Characterizing Emotional Prosody using Human-in-the-Loop Algorithms

POL VAN RIJN

Dissertation
zur Erlangung des akademischen Grades Doktor-Ingenieur (Dr. Ing.)

Characterizing Emotional Prosody using Human-in-the-Loop Algorithms

Pol van Rijn

23. Juni 2025

eingereicht an der
Fakultät für Angewandte Informatik Lehrstuhl für Menschzentrierte Künstliche Intelligenz
Universität Augsburg

Erstgutachterin: Prof. Dr. Elisabeth André
Zweitgutachter: Prof. Dr. Björn Schuller
Drittgutachter: Prof. Dr. Nori Jacoby

Tag der mündlichen Prüfung: 23. Juni 2025

Pol van Rijn Characterizing Emotional Prosody using Human-in-the-Loop Algorithms

Abstract

Speech conveys rich information beyond the spoken content including inferential cues about the speaker's intentions, personality, conversational goals, and emotions. This information is conveyed through prosody, characterized by variations in pitch, loudness, timing, and voice quality. Emotional prosody, in particular, is about how people speak when they are expressing emotions. The communication of emotions is crucial for successful communication in human-computer and human-robot interaction, which requires large datasets of emotional speech. In this thesis, we identify three core methodological problems in creating such corpora and propose solutions to them: obtaining a representative sample of all emotional prosodies (stimulus selection problem), identifying appropriate emotion annotation (taxonomy curation problem), and aligning emotional concepts across languages (lost-in-translation problem).

This thesis consists of three parts. In the first part, we develop Human-In-The-Loop (HITL) algorithms that provide solutions to the identified problems in emotional prosody. While corpora only indirectly capture the association between prosody (stimulus space) and emotions (semantic space), the actual association is stored in the minds of humans. HITL algorithms can sample this information directly from humans, by incorporating human decisions into computer algorithms. In particular, sampling algorithms from machine learning are used to iteratively characterize high-dimensional probability distributions. Here, we incorporate humans as part of the iterative procedure to obtain representative and diverse samples of stimuli over a distribution of latent concepts in human minds, such as the joint distribution of prosodic features and emotions.

Concretely, we propose three HITL algorithms: (i) Gibbs Sampling with People (GSP) to efficiently find instances of prosody that sound like a particular emotion using a voice model, (ii) Genetic Algorithm with People (GAP) to obtain a diverse set of emotional recordings through the process of mutation and selection, and (iii) Sequential Transmission Evaluation Pipeline (STEP) to distill a taxonomy of emotions from prosody. While the first two algorithms provide solutions to the stimulus selection problem, the last algorithm provides a solution to the taxonomy curation problem.

In the second part of the thesis, I establish an infrastructure to run massive online experiments across the globe. This infrastructure allows deploying the algorithms across languages, providing a solution to the lost-in-translation problem. We benchmark the created infrastructure by running a large-scale, cross-lingual experiment in a low-dimensional and well-studied domain.

We recognize that these three problems identified for emotional prosody are pervasive and exist for most machine learning datasets. For example, when constructing a corpus for object recognition, one has to select a representative sample of objects, decide on a taxonomy to label the objects, and for multilingual datasets decide how to align those taxonomies.

In the last part of the thesis, we demonstrate that these HITL algorithms, which have been developed to solve core scientific problems in emotional prosody, can be applied in adjacent domains. In particular, we show how GSP can be used for voice personalization for digital agents and avatars, and we demonstrate how the combination of GSP and STEP can be used to align impressions of robots across the auditory and visual modality.

The HITL algorithms developed in this thesis enable the creation of large-scale, high-quality datasets, by leveraging human decisions to more directly sample from the associations between the stimulus and the semantic space. In a broader context, these algorithms allow the creation of more representative corpora that can be used to train machine learning models that are more balanced and diverse and can be used to benchmark the performance of state-of-the-art models.

Zusammenfassung

Sprache vermittelt weit mehr als nur den gesprochenen Inhalt – sie enthält auch Informationen über die Absichten, die Persönlichkeit, die Ziele und die Emotionen eines Sprechers. Diese zusätzlichen Informationen werden über die Prosodie übermittelt, die sich durch Variationen in Tonhöhe, Lautstärke, Timing und Stimmqualität auszeichnet.

Emotionale Prosodie beschreibt insbesondere die Art und Weise, wie Emotionen in der Sprache ausgedrückt werden. Die erfolgreiche Kommunikation von Emotionen ist ein zentraler Bestandteil der Mensch-Computer- und Mensch-Roboter-Interaktion, setzt jedoch die Verfügbarkeit großer, qualitativ hochwertiger Datensätze mit emotionalen Sprachaufnahmen voraus.

Diese Arbeit identifiziert drei methodische Kernherausforderungen bei der Erstellung solcher Korpora und schlägt entsprechende Lösungen vor: (i) die Gewinnung einer repräsentativen Stichprobe aller emotionalen Prosodien (Stimulus-Selektion), die Identifikation geeigneter Emotionsannotationen (Taxonomie-Kuration) und (iii) die Identifikation und Abstimmung emotionaler Konzepte in verschiedenen Sprachen („Lost-in-Translation"-Problem).

Diese Dissertation gliedert sich in drei Teile. Im ersten Teil entwickle ich Human-In-The-Loop (HITL) Algorithmen, die Lösungen für die identifizierten Probleme bieten. Sprachkorpora erfassen nur indirekt die Assoziation zwischen Prosodie (Stimulusraum) und Emotionen (semantischer Raum), während diese Assoziation eigentlich im menschlichen Gehirn gespeichert ist. HITL-Algorithmen ermöglichen es, diese latenten Assoziationen zu extrahieren, indem menschliche Entscheidungen in den Algorithmus integriert werden. Hierzu nutze ich Sampling-Algorithmen aus dem Bereich des maschinellen Lernens, um iterativ hochdimensionale Wahrscheinlichkeitsverteilungen zu beschreiben. In solchen Verfahren, werden Menschen aktiv eingebunden, um repräsentative und vielfältige Stichproben von Stimuli über die gemeinsame Verteilung prosodischer Merkmale und Emotionen zu generieren. Konkret schlage ich drei HITL-Algorithmen vor: (i) Gibbs Sampling with People (GSP) – ein effizientes Verfahren zur Identifikation von Prosodien für bestimmte Emotionen mithilfe eines Sprachmodells. (ii) Genetic Algorithm with People (GAP) – ein evolutionärer Algorithmus zur Gewinnung vielfältiger emotionaler Sprachaufnahmen. Und (iii) Sequential Transmission Evaluation Pipeline (STEP) – ein Verfahren zur Ableitung einer Emotions-Taxonomie aus Sprachkorpora. Während GSP und GAP das Problem der Stimulus-Selektion adressieren, dient STEP der Lösung der Taxonomie-Kuration.

Im zweiten Teil der Arbeit entwickle ich eine Infrastruktur für groß angelegte Online-Studien. Diese Infrastruktur ermöglicht es, die entwickelten Algorithmen weltweit anzuwenden. Insbesondere ermöglicht die sprachübergreifende Anwendung von STEP die Untersuchung emotionaler Konzepte in verschiedenen Sprachen, womit das „Lost-in-Translation"-Problem adressiert wird. Die Infrastruktur wird in einem groß angelegten, sprachübergreifenden Experiment evaluiert.

Im letzten Teil der Arbeit wende ich die entwickelte Algorithmen auf angrenzende Forschungsgebiete an. So zeige ich, wie GSP zur Personalisierung von Stimmen von digitalen Agenten und Avataren genutzt werden kann. Zudem demonstriere ich, wie die Kombination aus GSP und STEP dazu beitragen kann, Eindrücke von Robotern aus verschiedenen Modalitäten (auditiv und visuell) aufeinander abzustimmen.

Die in dieser Arbeit entwickelten HITL-Algorithmen ermöglichen die Erstellung groß angelegter, qualitativ hochwertiger Datensätze, indem sie menschliche Entscheidungen gezielt zur effizienteren Erfassung der Assoziationen zwischen Stimulus- und semantischem Raum nutzen.

In einem breiteren Kontext tragen diese Methoden zur Entwicklung repräsentativerer Korpora bei, die für das Training ausgewogener und diverserer maschineller Lernmodelle verwendet werden können. Dadurch verbessern sie nicht nur die Benchmarking-Leistung moderner Modelle, sondern leisten auch einen wichtigen Beitrag zur besseren Erfassung und Nutzung emotionaler Prosodie in technischen Systemen.

Acknowledgements

First and foremost, I extend my deepest gratitude to my supervisors for their invaluable guidance and unwavering support throughout my PhD journey. I am profoundly thankful to **Elisabeth André** for her meticulous feedback on manuscripts, her dedication to fostering a stimulating research environment, and her broad scientific curiosity that inspires interdisciplinary collaboration. Your insights have been instrumental in shaping my work.

I also wish to express my heartfelt thanks to my other supervisor, **Nori Jacoby**, whose mentorship has profoundly influenced my growth as a researcher. Your enthusiastic supervision, deep involvement in my research, boundless energy, and passion for excellence in science have been truly transformative. I am incredibly fortunate to have had your guidance.

A special acknowledgment goes to **Björn Schuller**, for the generous commitment of reading and reviewing the thesis. I want to thank both supervisors for their generous support, the detailed feedback, and the joint work over the years.

This dissertation would not have been possible without the contributions and collaboration of my exceptional colleagues and collaborators. At the Chair of Human-Centered Artificial Intelligence, I am especially grateful to **Silvan Mertes**, **Dominik Schiller**, **Kathrin Janowski**, and **Katharina Weitz** for the enjoyable projects we undertook together and the papers we wrote together.

From the Computational Auditory Perception group, I would like to highlight my fruitful collaboration with **Harin Lee**. Working with you has been an absolute pleasure, and the synergy between our research topics has been remarkable. I am also extremely grateful to **Peter Harrison**. Thank you for your support, your openness to new ideas and features in PsyNet, and your thoughtful feedback on countless pull requests were immensely helpful. I also wish to thank **Frank Höger** for his technical support and collaborative efforts on PsyNet.

A heartfelt shout-out goes to **Raja Marjieh** and **Ilia Sucholutsky**, two of the most remarkable collaborators I have had the privilege to work with. Thank you for the exciting projects we tackled together over the years, the complementary expertise you brought to our work, and the countless hours of enjoyable research we did together.

I am also grateful to my other group members: **Manuel Anglada-Tort**, **Eline Van Geert**, **Ofer Tchernichovski**, and **Yoko Urano** for their engaging discussions and thoughtful contributions to my work. A special mention goes to student assistants and interns **Elif Celen** and **Dun-Ming "Brandon" Huang**, whose dedication and contributions to our shared projects have been truly commendable. It has been a privilege to work with all of you!

On a personal note, I wish to thank my family for their unconditional support. Especially to my wife, **Sophie**, for your encouragement, your patience, and for regularly taking our children, **Lieve** and **Jakob**, for enabling me to dedicate time to research and attend conferences. Finally, I owe immense gratitude to my parents, **Frans** and **Fros**, for their lifelong support and encouragement.

Last but not least, I extend my sincere thanks to the more than 11,000 participants who generously took part in my studies. Your contributions were vital to the success of this research.

Pol van Rijn

Contents

List of Publications

1. Reference: Peter M. C. Harrison, Raja Marjieh, Federico Adolfi, **Pol van Rijn**, Manuel Anglada-Tort, Ofer Tchernichovski, Pauline Larrouy-Maestri, and Nori Jacoby. 2020. 'Gibbs Sampling with People'. *Advances in Neural Information Processing systems*.
   Contributions: Designed and conducted emotion experiments, created figures, and contributed manuscript sections.
2. Reference: Dominik Schiller, Silvan Mertes, **Pol van Rijn**, and Elisabeth André. 2021. 'Analysis by Synthesis: Using an Expressive TTS Model as Feature Extractor for Paralinguistic Speech Classification'. *Interspeech*.
   Contributions: Conceived the idea together with the two other authors, created figures, and provided manuscript feedback.
3. Reference: **Pol van Rijn**, Silvan Mertes, Dominik Schiller, Peter M.C. Harrison, Pauline Larrouy-Maestri, Elisabeth André, and Nori Jacoby. 2021. 'Exploring Emotional Prototypes in a High Dimensional TTS Latent Space'. *Interspeech*.
   Contributions: Conceived the idea, designed and conducted all experiments, created figures, and wrote the manuscript.
4. Reference: **Pol van Rijn**, Silvan Mertes, Dominik Schiller, Piotr Dura, Hubert Siuzdak, Peter M. C. Harrison, Elisabeth André, and Nori Jacoby. 2022. 'VoiceMe: Personalized Voice Generation in TTS'. *Interspeech*.
   Contributions: Conceived the idea, designed and conducted all experiments, created figures, and wrote the manuscript.
5. Reference: Hubert Siuzdak, Piotr Dura, **Pol van Rijn**, and Nori Jacoby. 2022. 'WavThruVec: Latent speech representation as intermediate features for neural speech synthesis'. *Interspeech 2022*.
   Contributions: Validated the model, created figures, and provided manuscript feedback.
6. Reference: **Pol van Rijn**, Harin Lee, and Nori Jacoby. 2022. 'Bridging the Prosody GAP: Genetic Algorithm with People to Efficiently Sample Emotional Prosody'. *CogSci*.
   Contributions: Conceived the idea together with other authors, designed and conducted the experiments, created figures, and wrote the manuscript.
7. Reference: Raja Marjieh, **Pol van Rijn**, Ilia Sucholutsky, Theodore Sumers, Harin Lee, Thomas L. Griffiths, and Nori Jacoby. 2023. 'Words Are All You Need? Language as an Approximation for Human Similarity Judgments'. *ICLR*.
   Contributions: Implemented and conducted STEP experiments, created figures, and provided manuscript feedback.
8. Reference: Dominik Schiller, Silvan Mertes, **Pol van Rijn**, and Elisabeth André. 2022. 'Bridging the Gap: End-to-End Domain Adaptation for Emotional Vocalization Classification using Adversarial Learning'. *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*.
   Contributions: Performed partial analyses, created figures, and provided manuscript feedback.
9. Reference: Raja Marjieh, Ilia Sucholutsky, **Pol van Rijn**, Nori Jacoby, and Tom Griffiths. 2023. 'What Language Reveals about Perception: Distilling Psychophysical Knowledge from Large Language Models'. *Proceedings of the Annual Meeting of the Cognitive Science Society*.
   Contributions: Conducted partial analyses, created figures and interactive visualizations, and provided manuscript feedback.
10. Reference: **Pol van Rijn**, Yue Sun, Harin Lee, Raja Marjieh, Ilia Sucholutsky, Francesca Lanzarini, Elisabeth André, and Nori Jacoby. 2023. 'Around the World in 60 Words: A Generative Vocabulary Test for Online Research'. *CogSci*.
    Contributions: Conceived the idea, implemented the pipeline and experiments, created figures, and wrote the manuscript.
11. Reference: **Pol van Rijn** and Pauline Larrouy-Maestri. 2023. 'Modelling Individual and Cross-Cultural Variation in the Mapping of Emotions to Speech Prosody'. *Nature Human Behaviour*.
    Contributions: Conceived the idea, curated corpora, performed all analyses, created figures, and wrote the

manuscript.

12. Reference: Jakob Niedermann, Ilia Sucholutsky, Raja Marjieh, Elif Celen, Thomas L Griffiths, Nori Jacoby, and **Pol van Rijn**. 2024. 'Studying the Effect of Globalization on Color Perception Using Multilingual Online Recruitment and Large Language Models'. *CogSci*.
    Contributions: Supervised the first author, implemented experiments, developed analysis pipeline, created figures, and wrote the manuscript.
13. Reference: Raja Marjieh, **Pol van Rijn**, Ilia Sucholutsky, Harin Lee, Tom Griffiths, and Nori Jacoby. 2024. 'A Rational Analysis of the Speech-to-Song Illusion'. *Proceedings of the Annual Meeting of the Cognitive Science Society*.
    Contributions: Implemented the experiment and provided manuscript feedback.
14. Reference: Raja Marjieh, **Pol van Rijn**, Ilia Sucholutsky, Harin Lee, Nori Jacoby, and Thomas Griffiths. 2024. 'Characterizing the Large-Scale Structure of Grounded Semantic Network'. *Cognition*.
    Contributions: Implemented and collected data for STEP experiments, created figures, and provided manuscript feedback.
15. Reference: **Pol van Rijn**, Silvan Mertes, Kathrin Janowski, Katharina Weitz, Nori Jacoby, and Elisabeth André. 2024. 'Giving Robots a Voice: Human-in-the-Loop Voice Creation and Open-Ended Labeling'. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.
    Contributions: Designed and conducted all experiments, performed all analyses, created figures, and wrote the manuscript.
16. Reference: Harin Lee, Manuel Anglada-Tort, Oleg Sobchuk, **Pol van Rijn**, Marc Schönwiesner, Ofer Tchernichovski, Minsu Park, and Nori Jacoby. 2024. 'Global Music Discoveries Reveal Cultural Shifts during the War in Ukraine'. *10.31234/osf.io/7b98u*.
    Contributions: Created interactive figures and provided manuscript feedback.
17. Reference: Dun-Ming Huang, **Pol van Rijn**, Ilia Sucholutsky, Raja Marjieh, and Nori Jacoby. 2024. 'Characterizing Similarities and Divergences in Conversational Tones in Humans and LLMs by Sampling with People'. *ACL*.
    Contributions: Created figures and provided manuscript feedback.
18. Reference: Raja Marjieh, Ilia Sucholutsky, **Pol van Rijn**, Nori Jacoby, and Thomas L. Griffiths. 2024. 'Large Language Models Predict Human Sensory Judgments across Six Modalities'. *Scientific Reports*.
    Contributions: Created figures, developed interactive visualizations, and provided manuscript feedback.
19. Reference: Elif Celen, **Pol van Rijn**, Harin Lee, and Nori Jacoby. 2025. 'Are Expressions for Music Emotions the Same Across Cultures?'. *arXiv:2502.08744*.
    Contributions: Implemented experiments, developed international infrastructure for deployment, created figures, conducted all analyses, and co-wrote the manuscript.
20. Reference: Harin Lee, Eline Van Geert, Elif Celen, Raja Marjieh, **Pol van Rijn**, Minsu Park, and Nori Jacoby. 2025. 'Visual and Auditory Aesthetic Preferences Across Cultures'. *arXiv:2502.14439*.
    Contributions: Developed international infrastructure for deployment, provided technical support, and manuscript feedback.
21. Reference: Harin Lee, Elif Celen, Peter M. C. Harrison, Manuel Anglada-Tort, **Pol van Rijn**, Minsu Park, Marc Schönwiesner, Nori Jacoby. Submitted to the *International Society for Music Information Retrieval (ISMIR)* conference.
    Contributions: Finetuned the CLAP model, develeoped the data collection pipeline, and provided technical support.

List of Figures

List of Tables

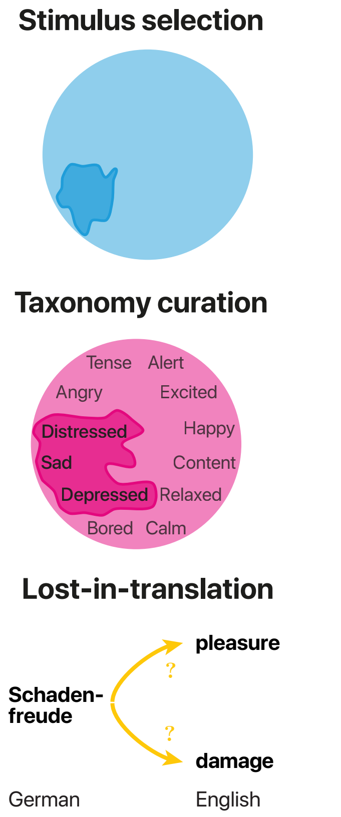
Chapter 1

Introduction

Imagine a scenario where a virtual assistant is helping a user who sounds distressed. If the assistant only processes the words being spoken but misses the emotional tone, it might respond inappropriately, offering routine solutions instead of empathizing or prioritizing urgent assistance. This mismatch can lead to frustration and uncanny or repulsive responses [21–26]. Understanding emotional prosody—the nuances of pitch, loudness, timing, and voice quality that convey emotions—enables machines to interpret not just *what* is said, but *how* it is said. This ability is essential for creating intuitive, responsive, and empathetic human-computer interactions, especially in contexts like health care [27], education [28–30], or preparing job interviews [31–36].

To make this communication successful, it requires at least two components: reliable detection of emotions and expressive synthesis of speech [37, 38]. In this thesis, I will mainly focus on the latter. Both methods rely on corpora, which are large collections of audio recordings. To learn the associations between emotions and prosody—for example, angry speech tends to be loud [39], and sad speech tends to be slow [40]—the corpus recordings need to be annotated with the intended emotion or the recognized emotion.

However, creating such corpora is challenging, because it has to overcome the following three problems (see Figure 1.1):


Stimulus selection

Taxonomy curation

Lost-in-translation

Figure 1.1: Core challenges in emotional prosody Biased sample from stimulus or semantic space or translation across languages.

- **Stimulus selection problem**: Imagine a corpus of emotional speech that does not contain any loud recordings, such as shouting, screaming, or yelling. This is unlikely to be a complete collection of all emotional stimuli (i.e., recordings), because loudness is a common feature of various intense emotions, such as "surprise", "fear", or "anger" [39]. Such a corpus will underrepresent all prosodies conveying emotions and thus models trained on it will have an impoverished representation.
- **Taxonomy curation problem**: The creation of emotional speech corpora often relies on a particular emotion taxonomy. For example, in acted emotional speech corpora, actors are prompted to say a sentence for a particular emotion [41–46], in spontaneous emotional speech corpora human annotators are asked to select a particular emotion label that is most adequate for a given segment or provide continuous ratings along selected dimensions [47–50], and in naturalistic data, human participants are asked to find sequences for a particular emotion [51–53]. This is problematic because the creation of the corpora is mediated through a particular, potentially incomplete taxonomy, leading again to a biased sample of the stimulus space (see above). Superimposing existing taxonomies on new data can be problematic, because (i) it is often unclear if the taxonomy fits the data (e.g., too broad or too narrow), (ii) it does not allow for the discovery of new concepts and (iii) taxonomies tend to be culture- and language-specific.
- **Lost-in-translation problem**: Quite often, emotion recognition models are not only applied on a corpus from a single language, but involve multiple languages, potentially from multiple corpora [54, 55]. Here, the question naturally arises, how to align emotional concepts across languages. Existing work relies on dictionary translations [56] or on

translations by multilingual authors [52, 57, 58], however it is unclear if the emotional concepts in one language are equivalent [59] or even exist in another language [60–63]. For example, the German word "Schadenfreude" does not have a direct translation in English, but describes the feeling of joy when someone else is damaged.

These problems illustrate how difficult it is to obtain a representative sample of emotional prosody. While the association between emotions and prosody is only implicitly captured in the corpus, the actual association is stored in the minds of humans. Human-In-The-Loop (HITL) algorithms can sample this information directly from humans, by incorporating human decisions into computer algorithms, such as Markov Chain Monte Carlo [64], Coordinate Descent Optimizer [1], and Diffusion [65]. A large body of literature has shown that such HITL algorithms efficiently and reliably sample from human representations and biases [64, 66–80].

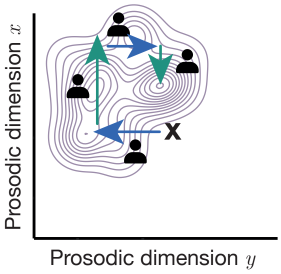


Figure 1.2: Gibbs Sampling with People Human participants sequentially control the parameters of a voice model to optimize for a particular emotion. Over iterations, participants explore dense regions of the prosodic space associated with the emotion.

However, these HITL algorithms involve decisions of a large number of participants. Here, we extend upon a line of work that has shown that crowdsourcing can be used to create large-scale datasets for affective computing [46, 56, 81–83] and develop three HITL algorithms to contribute to these three problems in the domain of emotional prosody.

1.1 Contributions

1.1.1 Human-In-The-Loop Algorithms

In the first HITL algorithm – Gibbs Sampling with People (GSP) – we work on the stimulus selection problem for emotional prosody. Concretely, we ask how can one identify emotions in the high-dimensional space of prosody? To do so, participants are provided with a tool to iteratively change the prosody of a voice model to make it sound like a particular emotion (see Figure 1.2). The voice model is trained on a large sample of varied prosodies and GSP allows us to efficiently identify which prosodic features are associated with a particular emotion thus providing a solution to the stimulus selection problem.

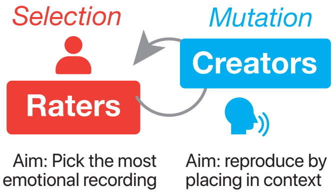

Selection
Mutation
Raters
Creators
Aim: Pick the most emotional recording
Aim: reproduce by placing in context

Figure 1.3: Genetic Algorithm with People Participants create emotional recordings by imitating previous recordings, separate raters select the most emotional recording which then propagates to the next generation of creators.

In the second HITL algorithm – Genetic Algorithm with People (GAP) – we also work on the stimulus selection problem. We have argued that most corpora of emotional prosody rely on a particular emotion taxonomy. To overcome this problem, we developed GAP, a HITL pipeline to obtain emotional recordings without pre-assuming particular emotions. GAP is a genetic algorithm involving either creators who provide emotional recordings by imitating recordings of previous creators or raters who must select the most emotional recording from a selection of recordings (see Figure 1.3). Crucial about the paradigm is that the creators are unaware that the experiment is about emotions and only imitate the previous recording. Over iterations, this alternating process between creation and selection leads to a set of emotional recordings without pre-assuming a particular emotion taxonomy.

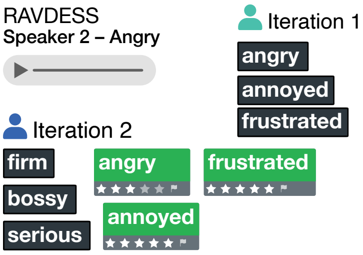


Figure 1.4: Sequential Transmission Evaluation Pipeline Participants listen to a speech recording, provide tags describing the emotion and rate tags provided by others. Over the course of iteration, this converges to a weighted bag-of-words representation of the emotional content of all recordings.

In the third HITL algorithm – Sequential Transmission Evaluation Pipeline (STEP) – we work on the taxonomy curation problem. To solve this, we developed a HITL paradigm, which involves participants providing an open-ended set of labels by describing the emotional content of speech recordings and by rating the labels of others (see Figure 1.4). This pipeline allows us to obtain a taxonomy of emotions perceived from a speech prosody and thus allows us to compile a more complete list of emotions. Furthermore, STEP allows for describing the relationship of emotional terms within the same language. For example, what is the relation between related concepts such as “afraid”, “anxiety”, “fear”, “frightened”, “scared”, “panic”, “terror”, and “worry”? Are they synonyms, or do they describe different aspects of the same emotion? STEP can provide answers to these questions by providing a weighted bag-of-words representation of the emotional content of the recordings.

All three paradigms are language-agnostic, meaning that they can be deployed in any language. If STEP is deployed in multiple languages, it can provide a solution to the lost-in-translation problem, because if the same stimuli are annotated in multiple languages, one can align the emotional concepts across languages.

1.1.2 Towards studying Cross-Lingual Differences

I have developed the infrastructure to run the HITL experiments across many languages. Concretely, we have co-developed a Python package called PsyNet that allows running large-scale online experiments across many languages on various recruitment platforms (e.g., Prolific, MTurk, Lucid). I have developed a language test to assess if the participants are fluent in the language they are participating in. Finally, we test the infrastructure by studying the low-dimensional, but well-studied domain of color naming [84–98], which has been shown to vary across languages [99].

1.1.3 Applications beyond Emotional Prosody

While these HITL algorithms have been developed to solve core methodological problems in emotional prosody, they can also be used to study a broader set of affective behaviors, such as personality traits or impressions. We show that GSP can also be used for personalization, by customizing voices for voice assistants and digital avatars. Concretely, participants optimize the voice to the appearance of a face, thus aligning the voice with their impression of the face.

In another project, we show that combining paradigms can be used to align impressions across modalities. Here, participants create a voice for an image of a robot using GSP. Both the impression of voice and the image of the robot are then annotated using STEP, which reveals the aligned perceptual space of both modalities. We show that the aligned space can then be used to predict fitting voices for new robots (e.g., find a voice for a fluffy, cute-looking robot).

1.2 Thesis Outline

The thesis is divided into three parts: HITL algorithms for emotional prosody, the development of the international deployment infrastructure, and the application of the HITL algorithms to other domains.

In Chapter 2, we provide general background information relevant to multiple chapters. Each chapter contains a background section, with information only relevant to that chapter.

In the first part of the thesis, we focus on emotional prosody. In Chapter 3, we conduct a large-scale meta-study, investigating the mapping between basic emotion labels and acoustic features [100] in emotional sentence recordings across the globe.[1] The work revealed fundamental weaknesses in studies on emotional prosody, which are summarized in the three problems mentioned above. In Chapter 4, we use GSP[2] to let people modify the prosody of a voice model[3] to make it sound like a particular emotion;[4] In Chapter 5, we develop GAP[5] to obtain emotional recordings without pre-assuming particular emotions; And in Chapter 6, we developed a HITL paradigm to annotate the semantic space of emotional prosody.[6,7,8]

In the second part of the thesis, I describe the infrastructure I have developed to make massive, cross-lingual, online experiments possible and use this infrastructure to study grounded semantics globally at scale. In Chapter 7, I describe my contributions as a core developer to Psynet, an open-source Python package, for implementing massive online experiments, automating the entire process from server provisioning to participant payment; In Chapter 8, I will present a language test, I developed[9] to quickly assess the vocabulary of a participant; And in Chapter 9, we benchmark this infrastructure to study the seminal problem of color naming across the globe.[10]

In the third and last part of this thesis, we will show that the HITL paradigms, developed to solve core methodological problems in emotional prosody, can also be used to solve more applied problems also in the voice domain. In Chapter 10, we show how GSP can be used for voice personalization[11] and in Chapter 11, how the combination of GSP and STEP can be used to align impressions across the auditory and visual modality.[12]

In the general discussion in Chapter 12, we will summarize the results, discuss limitations and sketch future directions.

[1] **Pol van Rijn** and Pauline Larrouy-Maestri. 2023. ‘Modelling Individual and Cross-Cultural Variation in the Mapping of Emotions to Speech Prosody’. *Nature Human Behaviour*.

[2] Peter M. C. Harrison, Raja Marjieh, Federico Adolfi, **Pol van Rijn**, Manuel Anglada-Tort, Ofer Tchernichovski, Pauline Larrouy-Maestri, and Nori Jacoby. 2020. ‘Gibbs Sampling with People’. *Advances in Neural Information Processing systems*.

[3] Dominik Schiller, Silvan Mertes, **Pol van Rijn**, and Elisabeth André. 2021. ‘Analysis by Synthesis: Using an Expressive TTS Model as Feature Extractor for Paralinguistic Speech Classification’. *Interspeech*.

[4] **Pol van Rijn**, Silvan Mertes, Dominik Schiller, Peter M.C. Harrison, Pauline Larrouy-Maestri, Elisabeth André, and Nori Jacoby. 2021. ‘Exploring Emotional Prototypes in a High Dimensional TTS Latent Space’. *Interspeech*.

[5] **Pol van Rijn**, Harin Lee, and Nori Jacoby. 2022. ‘Bridging the Prosody GAP: Genetic Algorithm with People to Efficiently Sample Emotional Prosody’. *CogSci*.

[6] Raja Marjieh, **Pol van Rijn**, Ilia Sucholutsky, Theodore Sumers, Harin Lee, Thomas L. Griffiths, and Nori Jacoby. 2023. ‘Words Are All You Need? Language as an Approximation for Human Similarity Judgments’. *ICLR*.

[7] Raja Marjieh, **Pol van Rijn**, Ilia Sucholutsky, Harin Lee, Nori Jacoby, and Thomas Griffiths. 2024. ‘Characterizing the Large-Scale Structure of Grounded Semantic Network’. *Cognition*.

[8] Raja Marjieh, Ilia Sucholutsky, **Pol van Rijn**, Nori Jacoby, and Thomas L. Griffiths. 2024. ‘Large Language Models Predict Human Sensory Judgments across Six Modalities’. *Scientific Reports*.

[9] **Pol van Rijn**, Yue Sun, Harin Lee, Raja Marjieh, Ilia Sucholutsky, Francesca Lanzarini, Elisabeth André, and Nori Jacoby. 2023. ‘Around the World in 60 Words: A Generative Vocabulary Test for Online Research’. *CogSci*.

[10] Jakob Niedermann, Ilia Sucholutsky, Raja Marjieh, Elif Celen, Thomas L Griffiths, Nori Jacoby, and **Pol van Rijn**. 2024. ‘Studying the Effect of Globalization on Color Perception Using Multilingual Online Recruitment and Large Language Models’. *CogSci*.

[11] **Pol van Rijn**, Silvan Mertes, Dominik Schiller, Piotr Dura, Hubert Siuzdak, Peter M. C. Harrison, Elisabeth André, and Nori Jacoby. 2022. ‘VoiceMe: Personalized Voice Generation in TTS’. *Interspeech*.

[12] **Pol van Rijn**, Silvan Mertes, Kathrin Janowski, Katharina Weitz, Nori Jacoby, and Elisabeth André. 2024. ‘Giving Robots a Voice: Human-in-the-Loop Voice Creation and Open-Ended Labeling’. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Chapter 2

Background

The study of emotional prosody draws on insights from various fields, including computer science (particularly affective computing and human-computer interaction), cognitive science (with a focus on language and culture), psychology (particularly speech perception), and linguistics (particularly phonology).

In my dissertation, I rely on insights from three key research areas: (i) Psychological perspectives on emotional prosody, examining emotion theories and prosody more broadly; (ii) Human-computer interaction research, specifically on Human-In-The-Loop (HITL) computation, which integrates human decision-making into algorithms to study mental representations such as emotions; (iii) Cognitive science research, investigating how language and culture influence perception.

2.1 Emotional prosody

2.1.1 Prosody

Prosody is an aspect of speech that conveys information beyond the literal meaning of the words [101, 102]. Before we formally define prosody and characterize its features, we start with a brief survey of speech production. A basic understanding of speech production is necessary to understand how prosody is produced, how it can be measured using acoustic features, how it can be manipulated to sound emotional [103], and how the emotional state of the speaker can influence speech production [39, 104].

2.1.1.1 Physionomy of Speaking

Human speech production can be divided into four subsystems: respiration, phonation, resonation, and articulation [105]. All subsystems are illustrated in Figure 2.1.

The first subsystem, the respiratory system, is responsible for in- and exhalation of air, which is needed to produce the necessary air pressure to speak [106]. During phonation, the second subsystem, the vocal folds rapidly open and close, creating vibrations many times per second [107]. The vocal folds are part of the larynx, which sits above the windpipe (trachea) and in front of the food pipe (esophagus) [108]. The number of the vocal fold vibrations is called fundamental frequency (f_0) and is usually measured in cycles per second (hertz, Hz) [109]. While pitch is the perceptual correlate of the periodicity of the acoustic signal [110], fundamental frequency is the main acoustic correlate of pitch. f_0 is primarily determined by the tension and length of the vocal folds and to a lesser extent by the subglottal air pressure below the folds [111]. So since children and females – on average – have shorter vocal folds, they have higher-pitched voices than men. Like string instruments, the rate of vibration increases when the folds are tight. The voice range of a speaker, which is the minimum and maximum f_0 a speaker can produce, is bounded by the

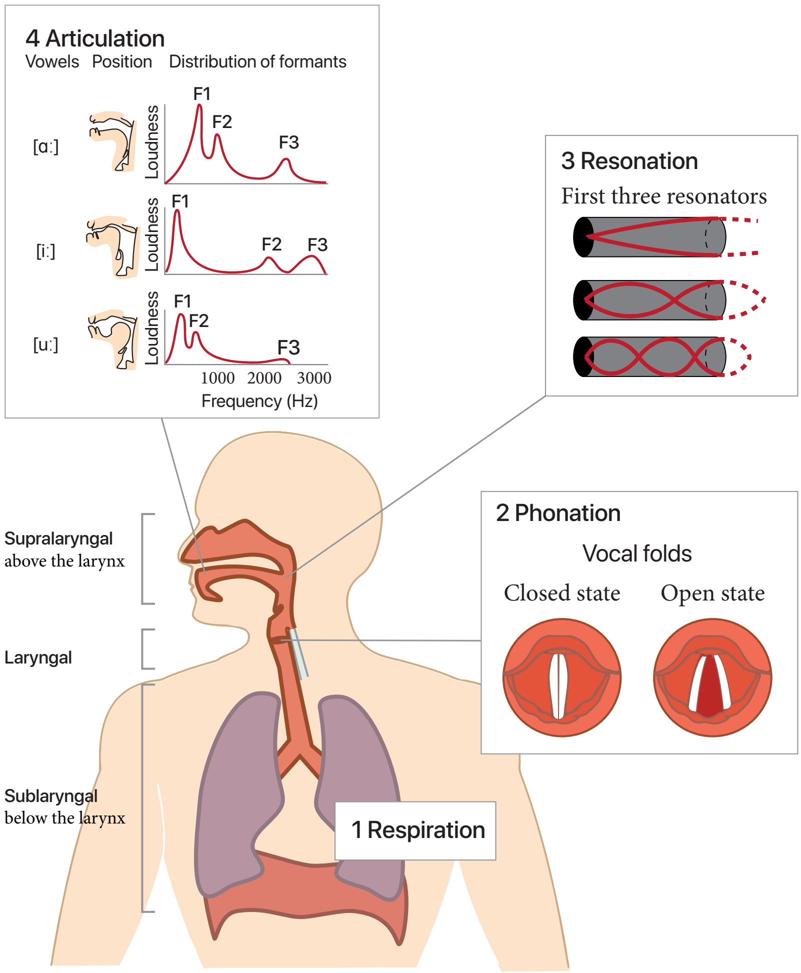


Figure 2.1: Four subsystems in speech production Respiration, phonation, resonation, and articulation. The distribution of formants is taken from [113] and the position of speech organs from [107].

physiology of the individual [106]. Still, within this range, the speaker has a fair amount of control over their f_0 [112] However, it is also implicitly influenced by the mental state of the speaker (e.g., psychological stress), hence making f_0 a particularly interesting feature for emotional prosody [39, 104].

During resonation, the fundamental frequency created by the vocal folds resonates in the trachea (see inset under "Resonation" in Figure 2.1), creating harmonic partials, which occur at integer multiples of the fundamental frequency. Since the windpipe is not a perfect tube, certain partials resonate more than others. These partials are called "formants" [114], leading to the unique sound of someone's voice (often referred to as "timbre").

During the articulation, the distribution of the formants is modified by the speech organs, like lips or parts of the tongue [107]. In Figure 2.1 (below "Articulation"), it is depicted how different positions of speech organs lead to the production of different vowels and hence to a different distribution of formants.

The result of this process can be conceptualized as a complex tone, which is a sum of all partials. However, unlike a real complex tone, which is fully periodic, human tones are semi-periodic due to slight irregularities in the vibrations of the vocal folds. Irregularities in the length of a single period are called "jitter" and variations in amplitude across periods are referred to as "shimmer". These irregularities are perceivable by humans and are common indicators for voice quality [115, 116].

Next to tones, the four subsystems can also produce non-periodic sound (noise). This means when one hears noise, one cannot hear pitch or measure a fundamental frequency [117]. We, therefore, call noise "voiceless" since the vocal folds do not vibrate. Noise is produced by the articulators, like by the lips in the consonant /p/. Noise can again be divided into two groups: fricative noise, spread out over a certain amount of time (i.e., continuant) like in /s/ or /f/, and burst noises that occur in a single burst as in /p/ or /t/ [118]. Tones and noises can also be mixed, which is the case for voiced consonants, like /v/ or /d/.

With this brief description of speech production, we now move forward to define prosody.

2.1.1.2 Definition of prosody

While there is no doubt that prosody plays essential contribution in the expression of emotion in speech [39, 40, 112, 119, 120] there is no consensus on the exact definition of prosody [121–123]. However, it is generally agreed upon that prosody:

- Refers to **suprasegmental aspects of speech**, which are overarching features that span multiple segments (e.g., words or syllables),
- consists of **pitch, duration, amplitude, and voice quality** to mark contrasts[1] or change meaning[2], and
- conveys **paralinguistic meanings** (e.g., communicate emotions or attitudes).

Prosody has been studied for centuries [124], initially for the recitation of religious texts and poetry [125], since the 20th century in phonology [126–128], and since the last twenty years in affective computing [129, 130]. Darwin was one of the first scholars to identify the importance of prosody in the expression of emotions [131]. While early studies relied on human hearing and manual annotation of prosodic features [132], the advent of computers and the development of automatic speech processing tools allowed for the automatic extraction of prosodic features from speech signals [133–135].

2.1.1.3 Acoustic features

Aspects of prosody can be measured using acoustic features, which are continuous signals extracted from the speech waveform. For example, one can estimate fundamental frequency (f_0) by correlating sliding windows of the signal with preceding windows (autocorrelation). This gives autocorrelation values for each lag, where the lag with the highest value corresponds to the period of the signal, which is the inverse of the fundamental frequency.

Since the signals are continuous, they can be summarized into summary statistics, like the mean or standard deviation, usually over the whole utterance. However, some features can not always be measured. For example, f_0, jitter (irregularities in the length of a single period), and shimmer (variations in amplitude across periods) can only be measured in voiced sounds. Such values are treated as missing values and are taken out when computing summary statistics.

[1] Emphasizing a different word can lead to a different focus. For example, "I talked to the guy with the ORANGE pants" and "I talked to the GUY with the orange pants" changes the focus from the color to the person.

[2] "Let's EAT GRANDPA" means something different than "Let's eat, GRANDPA" (literally eating your grandpa vs. calling your grandpa for dinner).

The features can be divided into four categories: time, frequency, amplitude, and spectral domain features [100] (see Table 2.1). Each of these features captures a different aspect of the speech signal and can be relevant for paralinguistic meanings like emotions.

Time domain features are mainly related to speech rate and rhythm, such as the number of loudness peaks per second, the number of voiced regions per second, and the duration of voiced or unvoiced regions. Speech rate can be an indicator of arousal or stress [136] and is thus emotionally meaningful.

Frequency domain features include f_0, jitter, and formant frequencies that can be obtained via a Fourier Transform of the signal. Fundamental frequency is relevant for the communication of emotion since higher arousal levels lead to higher muscle tension, which in turn leads to higher f_0 [104]. Also, jitter and shimmer are relevant for the communication of emotion, as they are perceptual correlates of voice roughness [115, 116], which can be an indicator of stress or anxiety [137].

Amplitude domain features include loudness, which is the perceived intensity of the sound, and the Harmonic-to-Noise Ratio (HNR), which is the ratio of the energy in the harmonics to the energy in the noise. Both loudness and HNR have been shown to be relevant for the communication of emotion [138, 139].

Spectral domain features can be computed on the frequencies or on the energy in different frequency bands and generally measure voice quality [140–142] and vocal effort [143] that are relevant for the communication of emotion [144]. Spectral energy features are often computed by measuring ratios (e.g., alpha ratio, Hammarberg index) or slopes (e.g., spectral slope) across the spectrum (i.e., the distribution of energy across frequencies). Spectral balance features are computed by measuring the energy in different frequency bands (e.g., spectral flux). Another commonly used spectral feature are Mel-Frequency Cepstral Coefficients (MFCC), which are computed by taking the Fourier Transform of the signal, applying a Mel filterbank, and taking the logarithm of the energy in each filter. This condensed representation of the signal based on the Mel scale (which aligns better with human perception) has been shown to be particularly useful for emotion recognition [145–147].

In linguistics, acoustic features are commonly extracted with the software Praat [133]. While it is a powerful tool for phonetic analysis, it has several limitations for emotion recognition:

- it does not provide a standardized feature set (and there are many hyperparameters to compute each feature), which makes it difficult to compare results across studies,
- it does not compute features in real-time, which is necessary for real-time emotion detection, and
- it does not have a convenient API and is thus not suitable for large-scale analysis of speech corpora.

To overcome these limitations, OpenSMILE [134, 135] was developed, which is a software that extracts standardized acoustic features from speech signals in real-time. For emotional prosody, the eGeMAPS standard feature set [100], has been extensively used and with 88 features it spans most prosodic dimensions [40, 119, 148]. In Table 2.1, we summarize the perceived correlates for each of the features. We can see that most of the perceived correlates in Table 2.1 are relevant for the communication of emotion.

Table 2.1: Description of the features included in the eGeMAPS feature set Also, see the summary in [152]

Type	Acoustic cue	Definition and measurement	Perceived correlate
Time	Rate of loudness peaks	Number of loudness peaks per second.	Velocity of speech
	Number voiced regions per second	Number of continuous voiced regions per second; similar to syllable rate.	Velocity of speech
	Duration of (un-) voiced regions	Duration of consecutive voiced or unvoiced regions; unvoiced regions approximate pauses.	Speech rhythm and fluency
Frequency	Fundamental frequency (f_0)	f_0 describes the rate of vibration of the vocal folds. It is described with summary statistics (e.g., arithmetic mean). The change of f_0 over time (referred to as pitch contour) is solely described with a slope. f_0 tends to be higher in aroused states [104].	Pitch and intonation contour
	Jitter	Jitter refers to small perturbations in f_0 in one cycle to another. It is caused by irregular fluctuations in the time it takes to open and close the vocal folds.	Pitch perturbations; "roughness" in the voice [116]
	First three formants (f_1 to f_3)	Caused by resonance in and speaker modulations of the vocal tract.	Voice quality [142]
Amplitude	Intensity	Sum of amplitudes across all frequency bands. It reflects the effort of the speaker to produce the utterance. Another amplitude measure used is equivalent sound level, which expresses the amplitude in decibels.	Loudness of speech
	Shimmer	Variations in amplitude from cycle to cycle, caused by irregular fluctuations in amplitude.	"Roughness" in the voice [115]
	Harmonic-to-Noise Ratio (HNR)	Proportion between harmonic (e.g., in vowels) and noise components (e.g., in unvoiced speech) in the voice.	"Breathy voice" [149]
Spectral	Alpha ratio	Ratio between the summed amplitude in the 50-1000 Hz and 1-5 kHz frequency bands.	Voice quality [140]
	Hammarberg index	Ratio of the strongest peak amplitude in the 0-2 kHz and the 2–5 kHz frequency bands.	Vocal effort [143]
	Spectral slope	Linear regression slope of the amplitudes of two frequency bands 0-500Hz, 500-1500Hz.	Voice quality [141]
	Energy proportion	Energy below and above 500 Hz, and 1000 Hz respectively.	Voice quality (related to spectral slope) [141]
	Harmonic difference	Difference H1 and H2 and H1 and A3, where the first f_0 harmonic is H1, and the second harmonic is H2; A3 is the highest harmonic in the third formant range.	Voice quality (also related to spectral slope) [141]
	Relative energy in f_{1-3}	Amplitude of the formants relative to f_0	Voice quality [142]
	Spectral flux	Speed at which energy distribution in different frequencies changes over time.	Rhythm and timbre [150]
	MFCCs (1–4)	Mel-Frequency Cepstral Coefficients (MFCC) using the Mel frequency scale which mimics human hearing.	Timbre [151]

More recently, deep learning methods have been used to automatically learn the relevant features from the data. Initially, unsupervised methods were used to directly learn the prosodic features from the audio alone, either the raw signal [153] or from a processed audio representation, such as a spectrogram [154, 155]. The acoustic features are then obtained by taking the output of the network at a particular layer. Later approaches, involved self-supervised learning, in which the model is trained to predict the next sample in the audio [156–161]. The latent audio representation can then be used to predict the emotion.

These advancements raise the question of whether you still need handcrafted features at all. While the handcrafted features performed on par with learned features for paralinguistic recognition tasks [162] or better [163], the use of handcrafted features is still justified, especially in cases with limited data (for example in low-resource languages), when the interpretability of the features matters (e.g., which acoustic features are used to communicate particular emotions across multiple cultures), or when the model is used in a real-time setting (e.g., in a call center).

2.1.1.4 Classifying Emotional prosody

These acoustic features can then be extracted from speech recordings with an annotation of the intended emotion or the recognized emotion. These annotations can either be discrete labels (e.g., basic emotions [59] such as "sad" or "happy") or dimensional (e.g., the circumplex model [164] with the dimensions valence and arousal). The extracted features were then put into a supervised learning method, such as a SVM, to predict the intended emotion or the recognized emotion from the features extracted from a new speech recording [129]. The classification performance of these datasets has been improved by the various challenges in the field of emotion recognition in the last two decades [165–173].

Most recently, foundation models have been proposed that are trained on text and data from other modalities [174–176] and can be used for emotion recognition [177]. For example, if the model is trained on sentence recordings and loud sentences tend to have captions containing the word "anger", the model would learn that loud sentences are associated with the word "anger". In contrast to previous methods, the supervision signal is no longer the emotion label in the corpus, but the association between the emotion term and the stimulus learned in the training data of the model. One can argue that foundation models solve the three core problems of creating emotional prosody corpora because emotion labels are no longer needed as a supervision signal, and the model learns the meaning of emotions from context. However, foundation models are trained on massive datasets scraped from the internet and this data might contain associations that are rather stereotypes than helpful for emotion recognition in real life (e.g., not all happy people smile or all angry people scream).

Now we have a basic understanding of how speech is produced, what prosody is, how prosody can be automatically extracted from speech signals, and how it can be used to classify emotions.

In the next section, we will discuss the psychological theories of emotion, which are relevant to the study of emotional prosody as well.

2.1.2 Emotion

For centuries people have been interested in the nature of emotions and how they are expressed [178]. In particular, the question to what extent emotions are universal or culturally constructed [59, 131]. The study of emotions has led to several influential theories that explain the relationship between physiological responses and emotional experiences.

2.1.2.1 How emotions emerge

Here is an overview of four major theories:

- **James-Lange theory** [179]: The theory proposes that emotions are the result of physiological changes in the body. Concretely, an external stimulus leads to a physiological response, and the perception of this response is what constitutes the emotional experience. For example, encountering a threatening stimulus (like a snake) causes physiological changes (like increased heart rate), and the awareness of these changes leads to the feeling (e.g., fear).

- **Cannon-Bard theory** [180]: In contrast to James-Lange theory, this theory proposes that emotions are the result of the *simultaneous* activation of the autonomic nervous system and the cognitive appraisal of the situation. So, according to this theory, the physiological arousal and the emotional experience are independent of each other. For example, encountering a snake leads to both physiological arousal and fear, but these two processes are independent of each other.
- **Schachter-Singer theory** [181]: This theory proposes that emotions are based on two factors: physiological arousal and cognitive labeling. It suggests that physiological arousal occurs first, and then the individual must identify the reason for this arousal to experience and label it as a specific emotion. According to this theory, physiological arousal is the same for different emotions, and the cognitive appraisal of the situation determines the type of emotion. For example, if you experience arousal (e.g., a racing heart) after encountering a stimulus, you interpret the context to determine whether you are feeling fear (e.g., encountering a snake), excitement (e.g., winning a lottery), or another emotion.
- **Lazarus' Cognitive-Mediational Theory** [182, 183]: This emphasizes the role of cognitive appraisal in the experience of emotion. It posits that our emotions are determined by our appraisal of a stimulus, which mediates between the stimulus and the emotional response. This appraisal can be immediate and often unconscious, determining whether we experience stress in response to a potential threat. For instance, hearing footsteps behind you in a dark alley might lead to an appraisal of danger, resulting in fear and physiological arousal.

These explanations differ to what extent emotions are hardwired (fearful stimulus leading to perception of fear) or constructed by the individual (appraisal of the situation leading to fear) and thus predict differences with respect to the universality of emotions. The explanations have influenced three major emotion theories: discrete [59], appraisal [184], and psychological constructivist theories of emotion [185]. The theories distinguish themselves by how emotions are conceptualized and internally structured [186].

2.1.2.2 Discrete Theories of Emotion

Discrete theories of emotions have largely been influenced by the work of Darwin [131], who proposed that emotions are universal and are a product of evolution to serve specific functions. For example, the idea that indicators for anger – such as bared teeth and narrowed eyes – are a functional preparation for an attack or that signs for fear – such as widened eyes and raised eyebrows – increase sensory input to detect threats and are thus a good preparation for flight.

Ekman's basic emotion theory [59] is a modern instantiation of this idea. He proposed that there are six basic emotions – happiness, sadness, anger, fear, disgust, and surprise – that are universally recognized across cultures. These emotions are thought to have a specific facial expression, which can be recognized by people from diverse cultural backgrounds [187].

So emotions are viewed as hardwired, biologically based responses, essentially "programs" that are activated by specific environmental triggers [186] and have developed over time to aid survival and social interaction. These automatic responses arise from particular stimulus events without prior learning or complex cognitive processing.

While the theories emphasize the universality of certain emotional expressions, there are also weaker versions of the theory that allow for some degree of variability in the expression of emotions. For example, the "in-group" effect [188] predicts that emotions are better understood by a member of the same community speaking the same "emotion dialect" [189, 190], hence allowing for some cultural variability in the expression of the same emotions [191]. Another example is "display rules", which are culturally specific norms that dictate how, when, and to whom individuals should express their emotions. For example, Matsumoto [192] found cultural differences in display rules across individualistic and collectivist cultures, differing in their social norms to express emotions. Further studies have explored how display rules vary not only across cultures but also within different social contexts, such as between work and non-work environments [193]. Also, particular emotions like "shame" seem to be more culturally specific [194].

So while discrete theories of emotion allow for some degree of variability in the expression of emotions, they emphasize the universality of certain emotional expressions and the biological basis of emotions.

2.1.2.3 Appraisal theories

Appraisal theories of emotion [184] take a different approach. They argue that emotions arise from an individual's cognitive evaluation or appraisal of an event or situation. These appraisals assess factors such as relevance to personal goals, congruence with desires, and perceived control over outcomes, thereby influencing the type and intensity of the emotion experienced. This perspective predicts that different individuals can experience varying emotional reactions to the same event based on their unique appraisals.

Lazarus' Cognitive-Mediational Theory (see Section 2.1.2.1) is a foundational framework within the appraisal tradition, emphasizing the adaptive function of emotions and their roots in personal meaning-making. Scherer further developed this theory [195] with the Component Process Model (CPM) of emotion [196], that emotions result from a continuous evaluation of events across multiple appraisal dimensions, leading to synchronized changes in experiential, physiological, and behavioral components.

Scherer [197] proposes a sequential appraisal process leading to the differentiation of emotional experiences, by implementing the following "appraisal checks":

- **Relevance**: Is the event relevant to the individual or their social group?
- **Implication**: Does the event imply a change in the individual's well-being? Does it have short- or long-term consequences?
- **Coping potential**: Does the individual have the resources to cope with the event?
- **Normative significance**: Does the event violate social norms and values? How does it influence the individual's self-concept?

By doing so, Scherer connects the appraisal process to the evolutionary function of emotions, where appraisal checks have evolutionary significance, such as preventing death, advancing reproductive goals, or avoiding social exclusion.

Another influential cognitive appraisal theory is the Ortony, Clore, and Collins (OCC) model of emotions [198], which provides a hierarchical structure for categorizing emotions based on the nature of what is being appraised. The theory posits that emotions arise from cognitive evaluations of three kinds of appraisals: events (e.g., joy, distress, hope, fear; for example when watching a movie), agents (e.g., admiration, reproach, gratitude, anger; for example towards your parents), and objects (e.g., love, hate, attraction, disgust; for example towards food). The OCC model is based on a hierarchical structure of cognitive appraisals, which are thought to lead to more than 20 distinct emotions based on different levels of cognitive appraisals. For example, "joy" and "distress" come from evaluating whether an event is desirable or undesirable, while "guilt" and "shame" arise from negatively appraising one's actions. The intensity of emotions is also influenced by additional factors such as personal relevance, expectations, and the likelihood of outcomes.

It has been shown that certain cognitive appraisals are rather universal (such as "valence" and "arousal"), others are more culturally specific (such as "responsibility" or "immorality") [199]. Compared to the discrete theories of emotion that predict a limited and universal set of basic emotions, appraisal theories predict more individual and cultural differences in the expression and recognition of emotions, as the emotional response depends on the appraisal of the situation and certain appraisals are culturally specific than others. Psychological constructivist theories of emotion, which are introduced in the next section, also predict more individual and cultural differences in the expression of emotions, but they assume that emotions are constructed from more basic psychological and physiological processes, rather than triggered by specific cognitive appraisals of events. This predicts a larger variation because there are more sources for psychological differences across individuals and cultures.

2.1.2.4 Psychological constructivist theories

Constructivist theories of emotion [185] extend the Schachter-Singer theory (see Section 2.1.2.1), arguing that emotions are constructed from more basic psychological and physiological processes [200]. These theories emphasize that emotions result from the interplay between an individual's perceptions, conceptual knowledge, and context. The brain constructs emotions by integrating sensory input, past experiences, and cognitive processes.

Concretely, like the Schachter-Singer theory, they assume a two-step process of emotions: In the first step, people receive an ongoing stream of bodily sensations that can give rise to emotions when interpreted in context. The idea is that emotions arise from basic feelings of arousal (activation–deactivation) and valence (pleasant–unpleasant) which is called "core affect". In the second step, core affective states are categorized and labeled. The ability to label these feelings is shaped by cultural, social, and individual differences, meaning that emotions are flexible and vary across cultures and individuals.

Psychological constructivist theories are influenced by Wilhelm Wundt's emotion theory [201] assuming three bipolar dimensions:

- **Pleasure–Displeasure** (Lust–Unlust): Ranges from positive to negative affective states.
- **Arousal–Relaxation** (Spannung–Lösung): Reflects the level of physiological activation or tension.

- **Excitement–Depression** (Erregung–Beruhigung): Captures dynamic changes in emotional intensity.

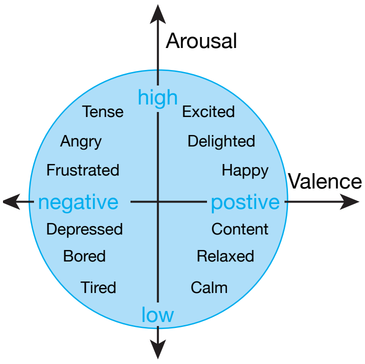


Figure 2.2: Circumplex model The circumplex model of emotion maps emotions onto a two-dimensional space of valence and arousal.

This led to the development of the PAD model of emotion [202, 203], which is a three-dimensional model of emotion that maps emotions onto the dimensions of pleasure (similar to valence), arousal, and dominance (degree of control–submission in a situation). The basic feelings referred to as core affect, are commonly described using the circumplex model of emotion [164] mapping emotions onto a two-dimensional space of valence and arousal (see Figure 2.2). While PAD allows to disambiguate certain emotions with similar valence and arousal values such as "anger" (high dominance) and "fear" (low dominance), the circumplex model can help to explain emotional similarity (e.g., anger and fear are both high-arousal negative emotions and are similar in this sense).

Psychological constructivist theories depart from the Schachter-Singer theory, by assuming that emotional responses are not only triggered by a particular stimulus (e.g., seeing a person cry can make you feel negative which is labeled as "sad") but instead can also be triggered by the individual's internal state (e.g., being hungry can make you feel negative which can be labeled as "sad"). This idea builds upon predictive coding [204], which posits that the brain is constantly making predictions about the world and updating these predictions based on incoming sensory information and interoception, referring to the brain's perception of internal bodily states [205]. In the context of emotions, this idea has been further developed [206, 207] to suggest that emotions are constructed by the brain based on the integration of sensory input and interoceptive predictions, leading to the theory of constructed emotions [205].

In contrast to discrete theories of emotion, psychological constructivist theories assume no hardwired links, but rather that emotions are constructed from more basic psychological and physiological processes. Also, appraisal theories, differ from psychological constructivist theories in two fundamental ways: (1) they assume that emotions are triggered by specific cognitive appraisals of events (and not constructed by the brain through the integration of sensory inputs, interoceptive signals, and learned concepts) and (2) they emphasize appraisals of external events, rather than re-evaluating internal states (interoception).

To summarize, the theories predict different levels of variability in the expression of emotions, but in part have in-theory explanations for the variability in the expression of emotions [188, 192, 199]. A limitation of these theories is that they all rely on theoretical constructs that may not align with empirical data from diverse populations or naturalistic settings. For example, discrete theories of emotion have been criticized for their reliance on a small set of emotions and facial expressions, which may not capture the full range of human emotional experiences [208], which may lead to limited generalizability. Also, appraisal theories have been criticized for their reliance on cognitive appraisals, which may not be universal across cultures [199]. Finally, psychological constructivist theories have been criticized for their reliance on the brain's construction of emotions, which may not be directly observable or measurable [206]. This is amplified by the fact that theories are often tested in small studies and struggle to incorporate large-scale data. Data-driven approaches, which we will discuss next, have the potential to overcome these limitations by uncovering latent patterns in emotional data without pre-existing theoretical biases and by incorporating large-scale data from diverse populations and naturalistic settings.

2.1.2.5 Cowen's Semantic Space Theory

The internet (e.g., via scraping and crowd-sourcing) and the development of low-cost sensors in mobile devices and have enabled a massive collection of data related to emotions [209]. Data-driven approaches have the potential to overcome the limitations of traditional theories of emotion by uncovering latent patterns in emotional data without pre-existing theoretical biases and by incorporating large-scale data from diverse populations and naturalistic settings.

Cowen and Keltner [210] are influential for the ideas developed in this thesis and have made a substantial contribution to the field by proposing the Semantic Space Theory, which is a data-driven approach to study emotions [210, 211]. Concretely, they asked participants to select an emotion label (from a pre-defined longer list of emotion labels based on previous literature) for a set of evocative stimuli. They then use mathematical methods to uncover the underlying structure of the emotional space, which they call the "semantic space". They study this space in various modalities including evocative videos [53] and responses to them [56], speech prosody [212], music [52], vocalizations [213, 214], facial expressions [208, 215], and art [216][3]. They focus on three aspects of the semantic space:

- **Dimensionality**: The number of distinct meanings of experiences or expressions within the space. In other words, how many different emotions can be communicated in a particular modality?
- **Conceptualization**: Whether the space is rather categorical or dimensional. Is the semantic space better described by a set of discrete categories or by a set of continuous dimensions?
- **Distribution**: How are the experiences in the space distributed? Do they form clusters or are there blends across states (gradients)?

To describe the **dimensionality** of the space, they developed a mathematical method which looks for orthogonal dimensions that have the highest covariance across two datasets (e.g., split-halves [53] or different cultures [52]). So, in other words, they look for dimensions that are consistent across different datasets. They then use statistical tests to determine which dimensions are significant. The number of significant dimensions tells us how many different emotions are communicated in a particular modality. They find the following number of distinct emotions for the following modalities: 27 for evocative videos [53], 12 for the responses to them [56], 12 in speech prosody [212], 13 for music [52], 24 for vocalizations [213], 28 for facial expressions [208], and 25 for art [216]. While the number of emotions varies substantially across modalities, it always covers a wide range of emotions that goes beyond the six basic emotions.

To characterize the **conceptualization** of the space, they compare dimensional ratings (such as valence and arousal) with categorical ratings. For facial expressions, they predicted category ratings from dimensional ratings and vice versa using linear regressions and found that categorical ratings explained but were not fully explained by general appraisal dimensions (appraisal judgments explained 56.1% of the variance in categorical judgments, while categorical judgments explained 91.4% of the variance in appraisal judgments) [208]. For music, they predict valence and arousal ratings from categorical ratings and show the correlation across two cultures (China and US) is higher between the predicted valence (r = .88 and .90) and arousal (r = .97 and .91) ratings than between the actual valence (r = .75) and arousal (r = .80) ratings across

[3] Click on the links to view the semantic spaces of the modalities: evocative videos, responses to them, music, non-verbal vocalizations, speech prosody, facial expressions, and art

the two cultures [52]. And for emotional prosody, they show that the category judgements correlate more (r = .80) than the appraisal judgements (r = .59) across two cultures (India and US) [212]. From these findings, they conclude that the space is better described by a set of discrete categories than a set of continuous dimensions, answering the question of conceptualization.

To address the **distribution** of the space, they apply dimension reduction [217] on the categorical ratings to project the stimuli into a 2D space. They show that expressions of emotions do not form distinct clusters, but instead are gradients of emotional states. This is criticized by Barrett, Khan, Dy, and Brooks, because the fixed position of the stimuli in the semantic space hides the variability in the perception of the stimuli, which is introduced by context, and individual or cultural differences [218]. In order to project the stimuli into a 2D space, Cowen aggregated the ratings of the participants, which means that individual differences in the perception of emotions are lost.

In a reply to this criticism, Cowen and Keltner do not provide an answer to the concern of aggregating out individual differences in the perception of emotions[4] but instead, they clarify that the method they used to estimate the dimensionality of the space was exploratory and not confirmatory as claimed by Barrett, Khan, Dy, and Brooks. This clarification is important to them because in a confirmatory analysis, there is a concern that the authors pre-assumptions influence the number of dimensions that are found. In the next paragraph, we will argue that Cowen's work actually suffers from authors pre-assumptions (participants pick from a pre-defined list of emotion labels, see "Taxonomy curation problem").

In addition, we identify the following three fundamental limitations of Cowen's work:

- **Stimulus selection problem**: The corpora he used in his work are either corpora of intended emotions [212, 213] – which pre-assume particular emotions (e.g., recordings for all six basic emotions) –, or he asked participants or search engines to propose stimuli for a particular emotion [52, 53]. This is problematic because the stimuli are not independent of the emotion labels, and he later uses the labels to create the semantic space. So, in essence, he shows that you can partially reconstruct the input emotions in the semantic space and it does not show how many emotions can be communicated in a particular modality.
- **Taxonomy curation problem**: In all his work, he relies on a pre-defined list of emotion labels (which are also not the same across studies!). He does not provide proof that the labels are the best labels to describe the data, or if gathering the labels in a data-driven way would lead to the same labels.
- **Lost-in-translation problem**: In his cross-cultural work [52, 56, 212, 214], he either relied on the translations of co-authors or on dictionary translations and treated the labels as synonyms across languages, but did not check systematically if the labels are actually used synonymously.

Other smaller limitations are that (i) while he did the semantic space analysis on many modalities, he did not compare the spaces across modalities (which is also difficult because the input labels differed across studies) and (ii) his "cross-cultural" work is only limited to a few cultures, which is not representative of the world's diversity, and (iii) he reported different statistics (e.g., for prosody he reports the correlation between categorical and appraisal judgements, but not for vocalizations) which makes it hard to compare across studies.

[4] For example, they could have estimated the variability in the position of the stimuli in the semantic space by using subsets of the data.

Recognizing these limitations, the work of Cowen demonstrates that: (i) the semantic space of emotions covers a larger array of emotions than the six basic emotions (dimensionality), (ii) the space is better described by a set of discrete categories than a set of continuous dimensions (conceptualization), and (iii) the expressions of emotions do not form distinct clusters, but are gradients of emotional states (distribution). For my thesis, this implies that I can focus on emotion categories as they can describe a rich emotional experience, that I can expect to find a large number of emotions in the data, and that single stimuli can evoke a wide range of emotions.

To summarize, in this section, we surveyed the literature of emotional prosody. First, we have explained how speech is produced, what prosody is, and how prosody can be automatically extracted from speech signals. For emotions, we have discussed three major theories and have shown a data-driven approach to study emotions using massive online experiments. Despite their potential, these approaches are fundamentally limited in a number of ways that we described above. To resolve these problems, I propose in my dissertation to extend the data-driven approach by using HITL sampling techniques. The advantages of these techniques are that they allow sampling in a balanced way from the mental spaces that are associated with emotional prosody and obtaining a representative sample of latent mental spaces, which is the key to the success of the proposed work. In the next section, we will describe the background and prior work describing HITL approach.

2.2 Human-in-the-loop algorithms

Corpora of emotional prosody enable to study associations between acoustic features and emotions. By learning these associations, it becomes possible to predict emotions from speech signals. However, in such corpora, the relationship between acoustic features and emotions is often indirect and noisy, influenced by factors such as background noise, linguistic prosody, and the speaker's accent. The true association between prosody and emotions, however, is stored in the minds of humans.

HITL algorithms leverage these mental representations by integrating human judgments into computational processes. Specifically, machine learning sampling algorithms are employed to iteratively characterize high-dimensional probability distributions. By involving humans in the iterative procedure, these algorithms obtain representative and diverse samples of stimuli, capturing latent concepts in the human mind, such as the joint distribution of prosodic features and emotions. These more direct samples from mental representations can be used to create better corpora to improve the generalization of machine learning models.

In this section, we will introduce existing HITL algorithms and demonstrate how they can be used to sample from latent mental representations.

2.2.1 Serial reproduction

Many of us enjoyed playing the game of telephone in our childhood (also known as Chinese whispers or "Stille Post" in German), where a secret message is whispered from one person to the other. The fun in the game often arises from that the message revealed by the last speaker is often completely different from the original. The reason why this occurs is that people respond not only to what they actually hear but what they think they hear; thus, their expectations affect their choices and accumulate over the course of the game. This game has been studied in the context of psychology to characterize human expectations, biases, and prior beliefs starting in the 1930s [71]. Later research in cognitive science inspired by this paradigm and used different variants to understand a wide range of cognitive phenomena, ranging from language [66] to perception [72, 74–76, 220]. Importantly, this is a simple, yet paradigmatical, example of HITL computation. The end result of the chain is influenced by all previous decisions, manifesting a simple computation.

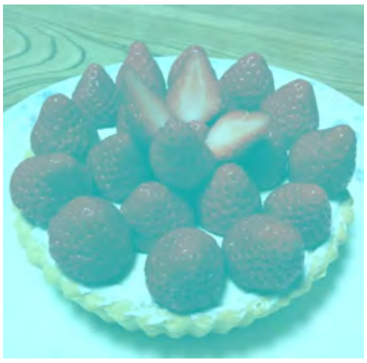


Figure 2.3: Visual illusion While this picture does not contain the color red, most people perceive the strawberries as red. The image by Akiyoshi Kitaoka is taken from [221] (CC BY-NC-SA).

Look at the strawberries in the image in Figure 2.3. What color are they? Most people perceive the strawberries as red, even though the image does not contain the color red. This visual illusion exemplifies how our perception is influenced by our expectations and biases. This observation is central to serial reproduction in which participants are shown a noisy stimulus – for example, by whispering, or by showing it briefly – and have to reproduce the stimulus from memory, amplifying biases of perception and memory, revealing shared priors that generate them.

Conceptually, one can think of serial reproduction as a Markov chain over stimuli (e.g., rhythms, stories, prosody) $x_0 \to x_1 \to \ldots \to x_t$ (see Figure 2.4). Each step involves a participant encoding and decoding the stimulus with a prior $\pi(x)$, capturing previous experiences with the stimulus, and a likelihood $p(x'|x)$, mapping stimulus x to the noisy percept x' (e.g., due to production constraints, perceptual noise, or memory limitations) [65]. So, for input stimulus x_i it is encoded as noisy percept x'_i and the reproduction decodes it into a new stimulus x_{i+1} [69]. This process repeats over iterations and will converge to the prior $\pi(x)$ if it is shared across participants [222]. Interestingly, this holds independent of the noise model $p(x'|x)$, which makes serial reproduction a powerful tool to study priors in a variety of domains [65].

Since then, serial reproduction has been used to study priors in a variety of domains, including spatial memory [72], rhythm [74, 75], and melodies [76, 77], which we will discuss in the next paragraphs.

2.2.1.1 Spatial memory

Langlois, Jacoby, Suchow, and Griffiths [72] shows that serial reproduction can be used to study spatial memory. Concretely, participants were shown a shape or a natural image with a point on top. The position of the point was initially randomly positioned (drawn from a uniform distribution) and participants had to reproduce the position of the point within the image (see Figure 2.5A). Throughout iterations, the initially randomly positioned points gradually move towards areas of high agreement, which tend to be landmarks in the image, such as the corners in the triangle and the roof of the lighthouse (see Figure 2.5B). The paradigm allowed to uncover spatial memory priors in unprecedented detail, and the authors showed that these perceptual biases

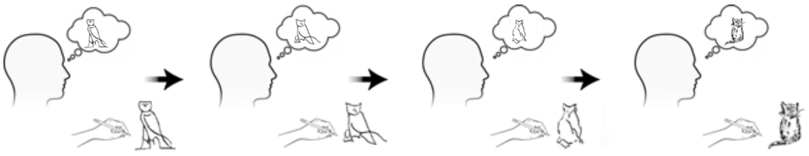


Figure 2.4: Serial reproduction Participants observe a stimulus (encode) and then reproduce it (decode) for the next participant. Drawings from the Bartlett's study [71], Figure reproduced from [65] with permission.

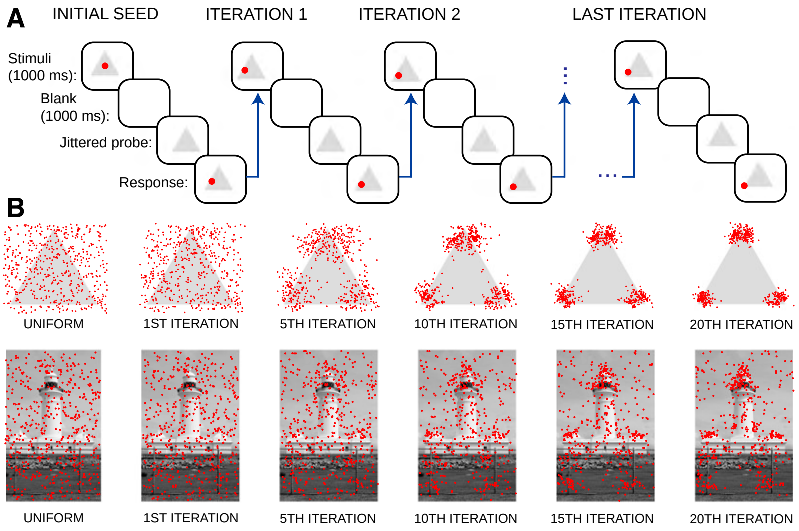


Figure 2.5: Using serial reproduction to study spacial memory **A** Participants are shown a shape or a natural image with a point on top. The image and the point disappear for 1 second and the image is now shown at a jittered position. The participant has to reproduce the position of the point within the image. **B** Over the course of iterations, the initially randomly positioned points gradually move towards landmarks in the image, such as the corners in the triangle (top) and the roof of the lighthouse (bottom). The figure is adapted from [72] with permission.

covary with variations in discrimination accuracy (participants tend to be more sensitive to differences in the areas of high agreement) [72].

2.2.1.2 Rhythm

Serial reproduction can also be used to study priors of rhythms. Jacoby and McDermott [74] have presented participants with four-tone rhythms that can be expressed with three intervals, which can be expressed as ratios. For example, 1:1:2 is a rhythm where the third interval is twice as long as the first two (see Figure 2.6A). In this example, the ratios are integers, but not all possible ratios are integers (e.g., 1.6:2.14:1). Each of the combinations of intervals can be visualized as a point in a 3D space, which can be visualized in a triangle (see Figure 2.6A). Participants listen to a rhythm and have to tap back the rhythm from memory. The authors show that initially uniformly sampled rhythms converge to integer ratios over the course of iterations (see Figure 2.6B). They also show that while both US and Tsimane' participants favor integer ratios, there are substantial differences in rhythm priors (see Figure 2.6C) [74].

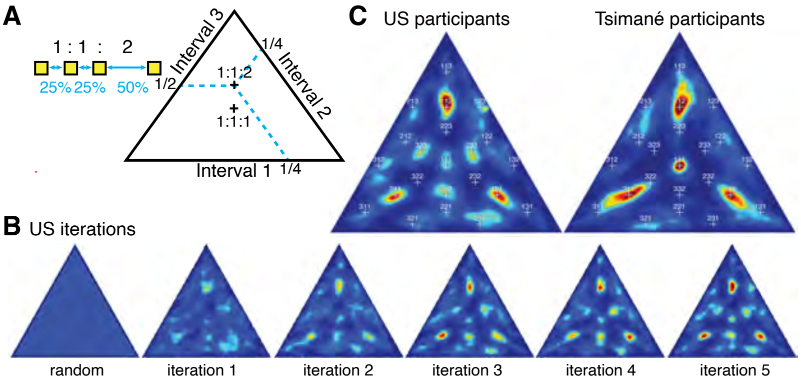


Figure 2.6: Using serial reproduction to study rhythm priors **A** Four-tone rhythms can be expressed with three intervals, which can be expressed as ratios. Each of the combinations of intervals can be visualized as a point in a 3D space, which can be visualized in a triangle. **B** Kernel density estimates of the reproduced rhythms over the course of iterations by US participants. Initially, the reproduced rhythms are uniformly distributed, but over the course of iterations, the reproduced rhythms converge to points in the triangle that correspond to integer ratios. **C** Cross-cultural comparison of reproduced rhythms by US and Tsimane' participants show that while both populations favor integer ratios, there are substantial differences in rhythm priors. The figure is modified from [75] with permission.

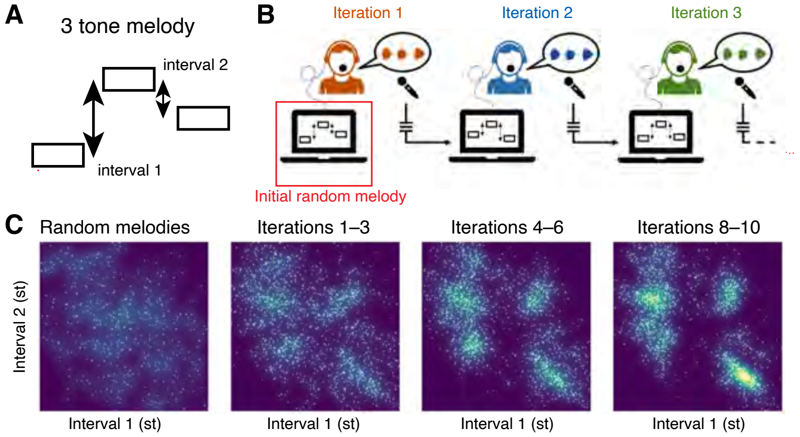


Figure 2.7: Using serial reproduction to study melody perception **A** Melodies are sequences of notes that can be expressed as intervals. In this example, a sequence of three pitches can be expressed as two intervals. **B** Participants listen to the previous melody and have to reproduce it by singing. The melody is automatically analyzed, synthesized, and played to the next participant. **C** Kernel density estimates of the reproduced melodies over the course of iterations by US participants. Each interval ranges from -12 to 12 semitones. Melodies converge to integer ratios over the course of iterations. The figure is modified from [76] with permission.

2.2.1.3 Melodies

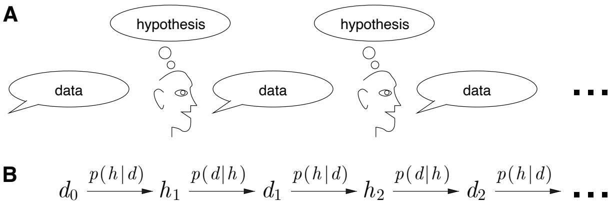
In another study, Anglada-Tort, Harrison, Lee, and Jacoby [76] have shown that the same procedure can also be used to study priors of melodies. In their study, participants listened to a melody and had to reproduce it by singing. A melody is a sequence of notes that can be expressed as intervals (see Figure 2.7A). The sung melody played to the next participant as the input melody (see Figure 2.7B). The authors show that melodies converge to integer ratios over the course of iterations (see Figure 2.7C) [76].

Previous work has also explored using serial reproduction for prosody. For example, Jacoby and McDermott [74] have shown that participants have similar rhythm priors for speech and music, and Kochanski [223] used it to study pitch imitation in speech. However, to the best of my knowledge, no study has used serial reproduction to study the priors of emotional prosody.

More recently, serial reproduction has also been recognized as a tool for explainable AI, by explosing the priors in Deep Neural Network (DNN)s [224].

In the next section, we will discuss an alternative formulation of serial reproduction, called iterated learning, which is used in the context of language evolution and learning.

2.2.2 Iterated learning



$$d_0 \xrightarrow{p(h|d)} h_1 \xrightarrow{p(d|h)} d_1 \xrightarrow{p(h|d)} h_2 \xrightarrow{p(d|h)} d_2 \xrightarrow{p(h|d)} \dots$$

Figure 2.8: Iterated learning **A** One can think of iterated learning as a process in which participants receive data from other participants, have to form a hypothesis about the data, and then emit data for the next participant. **B** This process can be formalized as a Markov chain. The schematic figure is taken from [68] with permission.

Serial reproduction and iterated learning are conceptually related, but differ in their focus. Where the former is about perception, the latter is about inference. Iterated learning is the process in which a population of agents learns from the output of the previous generation. Here, each participant receives data from other participants, has to form a hypothesis about the data, and then emits data for the next participant (see Figure 2.8A) [68]. Again this process is formalized as a Markov chain, in which a hypothesis is sampled from the posterior $p(h|d)$ and data is sampled from the likelihood $p(d|h)$ (Figure 2.8B), which will converge to its stationary distribution h if all learners have the same prior $p(h)$ [69, 70].

This conceptualization is particularly useful when one wants to study how structured patterns emerge or how culture shapes the transmission of information. One can consider each iteration in an experiment as a separate generation, where the data is passed from one generation to the next, where chain experiments are a way to mimic the process of cultural transmission.

Kirby, Cornish, and Smith use this paradigm to study language evolution [66]. In their experiment, participants are shown a novel artificial language and have to learn the language by observing the behavior of the previous participant. They found that over the course of iterations, the language becomes easier to learn and more structured, providing evidence that the process of cultural transmission can introduce gradual changes and regularities in the system [225].

In another study, Xu, Dowman, and Griffiths have studied how iterated learning shapes the emergence of color categories [67]. In their study, participants are shown examples of labeled colors and classify new colors based on those examples, and these classifications of the participants are used to generate new examples for the next participant. Participants are randomly assigned to a group in which a fixed number of color terms are used to describe the colors (the actual color terms were non-words). Initially, terms are randomly assigned to colors, but over the course of iterations, the categorization of colors becomes more similar to the color categorization in languages with the same number of color terms [226]. This indicates that the categorization of colors emerging from languages is not random, but follows certain universal regularities.

2.2.3 Markov Chain Monte Carlo with People

One limitation of serial reproduction is that it entangles perception and production, as the responses depend on the process of memorizing and internalizing the stimuli before reproducing. One way to overcome this limitation is to study mental representations without production constraints. This can be done with the paradigm Markov Chain Monte Carlo with People (MCMCP) [64]. The paradigm can be compared with the process of constructing a facial composite, where a forensic artist iteratively draws a face based on witness description until the drawn face matches the memory of the suspect's face. Instead of a forensic artist, faces can also be created using a generative model, but this requires a parametrization of the model (e.g., facial features of the suspect) and an efficient procedure to search all possible combinations.

Sanborn and Griffiths [64] realized that this process of finding a facial composite can be described in mathematical terms as a Markov Chain Monte Carlo (MCMC) sampler over the facial feature space. MCMC is a statistical method used to sample from probability distributions that are difficult to sample from directly. Metropolis-Hastings algorithm [227] is a popular MCMC algorithm [228], which starts at a random point in the space, proposes a new sample close to the current state using the proposal distribution, and then accepts or rejects the new sample based on the acceptance probability. The novelty proposed by Sanborn and Griffiths [64] was to use human decisions to decide which sample to take and hence included human decisions in a computer algorithm [64, 78].

In this section, we have shown how serial reproduction can be used to study biases in perception and memory, how iterated learning can be used to study the emergence of structured patterns in a population, and how Markov Chain Monte Carlo with People can be used to study mental representations. In the dissertation itself (Chapters 4–6), we will develop additional HITL paradigms that extend MCMCP and serial reproduction to study the priors of emotional prosody and for other similar applications.

2.2.4 Parallels to data augmentation

In this thesis, we use HITL algorithms to obtain representative samples that capture latent associations between emotions and prosody. Another approach to achieving this is data augmentation, a common machine learning technique that expands training datasets by applying transformations to existing data.

For instance, Liang, Chen, Zhao, Jin, Liu, and Lu proposed an adversarial learning framework to mitigate the influence of cultural differences in video-based emotion recognition [229]. Their approach frames emotion and culture recognition as adversarial tasks: the system learns to accurately identify emotions while simultaneously attempting to confuse a culture classifier. This dual-objective training encourages the model to extract emotion features that are invariant across cultural contexts.

A related challenge is that many emotion datasets are limited in size and restricted to discrete emotion categories. However, real-world emotional expressions often exist along a continuum, blending multiple emotions [53, 56, 208]. To address this, Mertes, Schiller, Lingenfelser, Kiderle, Kroner, Diab, and André introduced a generative approach using label interpolation within Generative Adversarial Networks (GANs) [230]. By interpolating between discrete emotion labels during training, their model learns to generate facial images that reflect nuanced emotional expressions, effectively bridging gaps between categorical emotions.

These studies demonstrate how data augmentation techniques can help refine the association between emotions and stimuli, denoising the latent associations between emotions and their expression.

2.3 Influence of language and culture

In this section, we will review the literature on how language and culture shape the expression of emotions, which is crucial for understanding the cross-cultural aspects of emotional prosody.

Recent studies have shown considerable variability in the expression of emotions [60, 231–233], touching upon the longstanding question in psychology and cognitive science whether emotions are shaped by universal constraints or by experience (e.g., influences of culture or language). This research question ties into broader inquiries in cognitive science, such as how language and culture shape mental representations and influence various cognitive processes. It also raises the issue of how to determine whether the demographic sample of participants is representative of the global population, ensuring the validity of general claims about human cognition.

2.3.1 Language shapes thought?

There is a long-standing debate in cognitive science about whether language shapes thought [236]. One particularly influential hypothesis is the Sapir-Whorf hypothesis, which posits that the structure of a language can shape the way people think and perceive the world. This idea has been supported by a number of studies showing that language can influence mental representations and cognitive tasks [90, 237–239].

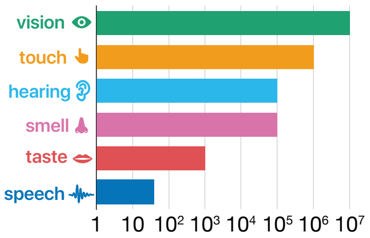


Figure 2.9: Bits of information per second Data from [234, 235]. Bits per second on a log10 scale.

Grounded semantics provides an interesting perspective on this debate, in which the same stimuli are presented and categorized by participants from different cultures and languages. The stimuli are mainly sensory experiences—such as viewing a color, smelling an odor, listening to sounds, or feeling a texture. To speak about these experiences, they need to be compressed, because speech only transmits at a rate of 39 bits per second [234], while humans receive more than 11 million bits of information per second through the sensory system [235] (see Figure 2.9).

Interestingly, the same sensory experiences are compressed differently across different languages for all five senses:

- **Vision**: Different languages have different color terms [99] or facial expressions for emotions [231].
- **Touch**: Different languages have different words for tactile experiences [240] and pain terms [241].
- **Hearing**: Different languages have different sound-meaning associations [242] or rhythm priors [220] and show differences in emotional non-verbal vocalizations [233] and prosody [11].
- **Smell**: Different languages have different odor terms [62]. For example, where Western languages tend to describe a scent by referring to the source (e.g., a banana) or the impression of the receiver (e.g., “stinky”), other languages (e.g., Jahai) have a distinct vocabulary for describing odors [243].
- **Taste**: Different languages have different taste terms [244], for example, where Westerners would describe the taste of Parmesan cheese as “salty”, Japanese people would describe it as “umami”.

Analyzing these differences can provide insights into the influence of language on mental representations and whether these differences – in part – can be explained by more general cognitive processes and biological constraints.

This question is tightly linked to a larger problem: obtaining a sufficiently diverse sample of cultures and languages to make general claims about human cognition. In the next section, we will discuss the limitations of unrepresentative population sampling in the social sciences.

2.3.2 Representative sample

2.3.2.1 WEIRD people

Henrich, Heine, and Norenzayan [245] raised awareness of the fact that the majority of participants in psychological studies are Western, Educated, Industrialized, Rich, and Democratic (WEIRD). This is particularly problematic because, based on a narrow sample of the world’s population, general claims are made about human cognition [246]. Where more than 90% of the participants in psychological studies are from Western countries (with the majority being from the US), the rest of the world accounts for less than 10% of the participants, but about 90% of the world's population [247].

Overgeneralizing from WEIRD participants to the rest of the world can lead to not recognizing differences in:

- **Sensory abilities and perception**: Recent work has demonstrated substantial cross-cultural abilities in underwater vision [248], olfaction [249], and spatial navigation [250].
- **Economic preference**: Researchers long assumed that preferences or motivations for reciprocity (i.e., mutual exchange of benefits, e.g., selling products for money), fairness, and altruism were universal. However, recent work has shown that these preferences can vary substantially across cultures [251].
- **Personality traits**: The "Big Five" [252] personality structure is often assumed to be universal, but recent work suggests this is not the case in some small-scale societies [253].
- **Morality**: Morality judgments were assumed to be made along five moral foundations [254], but recent work has shown that certain cultures have more moral foundations [255] or that the moral foundations are not as universal as initially hypothesized [256].
- **Emotions**: Recent work has shown substantial differences in emotion words [60, 61], facial expressions [231, 232], and non-verbal vocalizations [233] (e.g., moaning or screaming).

These findings highlight the need to include more diverse samples of cultures and languages in psychological studies to make more general claims about human cognition. This debate also has implications for other research fields. For example, in medicine, the majority of research is conducted on WEIRD bodies [257], which can lead to not recognizing differences in disease prevalence, treatment efficacy, or side effects across different populations. It also highlights problems in computer science where the majority of training data is from WEIRD people, which have become amplified with the rise of generative models [258].

Working with a WEIRD population also has other side effects, namely mainly involving participants who speak English. In the next section, we will discuss why this overreliance on English is problematic.

2.3.2.2 Overreliance on English

More than 70% of the participants in human experiments are from the USA [247] where for 80% of the participants English is their first and only language [259]. This is in sharp contrast to the world's population, where bilingualism and multilingualism (speaking at least one foreign language) are the norm [259].

Also, English is a non-representative language of the 7,000 languages spoken worldwide [260, 261]. This is problematic because the language someone speaks can influence mental representations and cognitive tasks [90, 237–239, 259].

For example, English is quite particular in that it:

- **writes from left to right**: But writing direction influences spatial cognition [262] and memory [263].

- **has a small vocabulary for certain sensory experiences**: For example, English has a very limited vocabulary for odors [62], with the word either referring to the object (e.g., "rose") or the impression of the smell (e.g., "stinky"). Other languages, such as Jahai, have a distinct vocabulary for describing odor perception [243].
- **is a right-branching language**: Whether a language is right- or left-branching (i.e., the order of the subject, verb, and object) influences working memory [264].
- **frequently uses a relative viewpoint**: English frequently uses a relative viewpoint from the speaker's perspective (e.g., left or right), which can influence spatial cognition [237, 265].
- **is optimized for direct information exchange**: In other languages, direct information exchange is considered rude and vagueness is preferred [266, 267].
- **prefers frequent expression of gratitude**: English speakers frequently express gratitude (e.g., "thank you", "please", "you are welcome"), but not all languages even have words for expressing gratitude [268].

These examples illustrate that English is not representative of all languages and that the particularities of a language can influence ostensibly non-linguistic tasks and capabilities, such as memory [263, 264], spatial cognition [262, 265], or odor perception [62, 237, 243].

2.3.2.3 Towards a more representative sample

As we discussed above, while the majority of research is conducted on WEIRD people, a fraction of research is conducted on remote societies. While this research allows one to obtain valuable insights into human cognition, the sample sizes are tiny, the expeditions come at a high cost, and there is very little room for automation.

In this thesis, we argue that recruiting participants from a diverse set of countries and languages is a more scalable and cost-effective way to study human cognition across the globe. As of 2024, 66.2 % of the world population has internet access, and with 5.35 billion smartphones on average, 69.4 % of the world population owns a smartphone [269]. While this population accessible through the internet is more exposed to other cultures[5], it is likely to still exhibit a strong linguistic and cultural diversity. Online cross-cultural experiments, unlike research expeditions, allow to collect massive amounts of data, are less expensive (only paying for the server infrastructure and for the local wage, which tends to be much lower than in the Global North), and have high potential for automation (e.g., recruitment, payment, translation).

To accomplish this goal, I have developed a fully automated pipeline for running massive online experiments and created an infrastructure to collect data from a large and diverse sample of languages spoken worldwide. In particular, I have implemented a recruitment service which, as of 2024 allows to recruit participants from 135 countries and 72 languages.

This is one step ahead towards a more informed sampling of languages and cultures to study cognition across cultures [270].

[5] Although remote societies are also becoming more connected through satellite phones, radios, and smartphones.

© Noah Buscher

Part I
Emotional Prosody

The first part of this thesis is about how emotions are communicated with the voice when speaking (emotional prosody). In Chapter 3, I conduct a large-scale meta-study of emotional prosody and show that there are considerable differences in the expression of emotions across countries, languages, and individuals, matching previous findings in the other modalities [60, 231–233]. The analysis, furthermore, reveals three core problems in corpora of emotional prosody (stimulus selection, taxonomy curation, and lost-in-translation problem) and we develop three Human-In-The-Loop (HITL) approaches to provide solutions to them. In Chapter 4 and 5, we present approaches to better sample from the stimulus space and, in Chapter 6, we develop an approach obtain an emotion taxonomy.

Chapter 3

Meta Analysis of Emotional Prosody

Based on
Pol van Rijn and Pauline Larrouy-Maestri. 2023. 'Modelling Individual and Cross-Cultural Variation in the Mapping of Emotions to Speech Prosody'. *Nature Human Behaviour*.

A key question in psychology is how prosody can communicate emotions in speech [119]. More specifically, the question is which acoustic features are used to convey particular emotions in speech and how these features potentially differ across cultures, languages, and individuals. This objective slightly differs from the research question in affective computing, where one of the goals is to detect emotions in real-life applications reliably. So in psychology, the emphasis is more on people's abilities to reliably detect the emotions from prosody alone (i.e., without context) [190, 191, 271, 272] and which acoustic features can be directly associated with specific emotions [39, 40, 42]. Studying the mapping between prosody and emotion typically relies on supervised data generated by recruiting actors to record sentences for specific emotions with neutral sentence meaning. The neutral content is important for psychological research because emotional cues in sentence content would override prosodic cues, leading to ceiling performance in human emotion recognition experiments [271, 273].

To study the mapping between emotions in acted corpora, existing studies investigate the mapping within a single speech corpus [40–46, 274–290] or conduct meta-studies on the reported acoustic features in the respective papers. Existing meta-studies traditionally:

- do not estimate differences in the mapping at more granular levels of analysis (e.g., individual speakers, languages, cultures),
- average over participants, leading to false confidence due to the removed variation,
- do not account for sampling imbalances (e.g., unequal number of recordings per culture, number of speakers, or number of sentences), and
- use the acoustic features as reported in the respective papers and do not recompute the features using identical preprocessing and feature sets [291].

At the time of writing the paper, there was no contemporary study investigating the mapping between prosody and emotions in acted corpora at a large scale. To overcome these shortcomings, I collected a large array of acted emotional recordings, including 432 individuals worldwide, speaking 2,963 unique sentences totaling 3,252 minutes of emotional speech, and preprocessed (Section 3.2.2) and extracted the features from the recordings (Section 3.2.2–1).

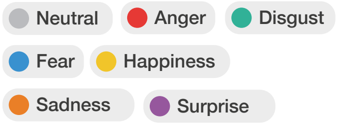


Figure 3.1: Basic emotions The six basic emotions and "neutral" are used as mapping targets.

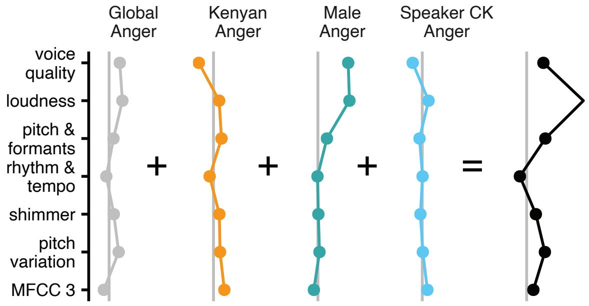


Figure 3.2: Multilevel mappings The model learns a multilevel mapping, consisting of a mapping that exists in all corpora as well as mapping deviations based on certain grouping variables, like culture or speaker. In this particular example, the mapping for "anger" for a male Kenyan English speaker is depicted.

Since the majority of studies rely on basic emotions (with some exceptions [42, 45, 212]), we only include speech corpora that cover basic emotions and "neutral" (Figure 3.1). In each corpus, a single speech recording is associated with one intended emotion.

To study the relationship between emotions and speech prosody, we employ Bayesian multilevel multinomial logistic regression models (Section 3.2.4). These models make a linear prediction for each emotion. The predicted emotion is the one with the highest value. Each predictor includes an intercept to account for potential imbalances in the base rate of emotion labels, as well as a series of coefficients for each of the seven acoustic features that describe the connection between speech prosody and emotions.

The overall procedure is depicted in Figure 3.2.

3.1 Background

3.1.1 Theories of Emotion

In the general background (Section 2.1.2), we described three theories of emotion: discrete [59], appraisal [184], and psychological constructivist theories of emotion [185]. We have argued that while all theories have in-theory explanations for a high variability in the expression of emotion, the discrete theory would predict a low variability, the appraisal theory a medium variability, and the psychological constructivist theory a high variability.

3.2 Methods

3.2.1 Speech corpora

To collect a large array of acted emotional recordings, we adopted a principled approach to scanning available speech corpora [191]. Using the approach, I identified 200 suitable corpora but obtained access to only 42. In total, 24 corpora passed our requirements[1] and were included in the analysis [40–46, 274–290]. The full list of corpora has been released here: `http://emotional.speechcorpora.com`.

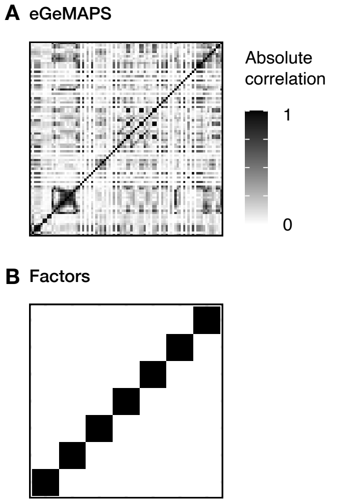


Figure 3.3: Correlatation across acoustic features **A** Correlations between the 88 features in the eGeMAPS set. **B** Correlations between seven factors with Varimax rotation.

[1] The requirements are full-sentence recordings with a single basic emotion annotation. See supplementary materials of the paper for more details [11].

3.2.2 Preprocessing

To process all corpora identically, I ran the following preprocessing steps. First, I made sure that there were no sounds other than speech that could disturb the acoustic feature extraction, like background music. For one corpus [288], I had to segment the speech from longer fragments into sentences. This was done with an adaptive algorithm changing a loudness threshold and a minimal silence duration in Praat [133] using Parselmouth [292]. If there were only video recordings of the spoken sentence, audio was extracted from the video signal. Finally, all recordings were converted to mono and downsampled to 16,000 Hz. For each file, I encoded the following information in the filename: corpus, intended emotion, sentence code, and speaker.

3.2.3 Acoustic features

We use the eGeMAPS standard feature set [100] since it has been extensively used to classify emotions. Since acoustic features are highly correlated (see Figure 3.3A) and including highly correlated features in a model can lead to multicollinearity, we reduce the dimensionality of the feature set by performing factor analysis.

We conducted a factor analysis since 74 of the 88 features are correlated at least .3 with at least one other feature, suggesting reasonable factorability, and the Kaiser-Meyer-Olkin measure of sampling adequacy is .87, and Bartlett's test of sphericity is significant ($\chi^2(3828) = 9429598, p < 0.01$). Since we apply principal component analysis with Varimax rotation (using the R package `psych` [293]), the final features are decorrelated (see Figure 3.3B).

To assess the minimal number of dimensions, we plot the eigenvalues per component (see Figure 3.4), showing a decay around seven features. We also computed the Widely Applicable Information Criterion (WAIC) for models with an increasing number of factors (see Figure 3.5). Again, we can show that the WAIC improvement stagnates for factor solutions with more than seven features.

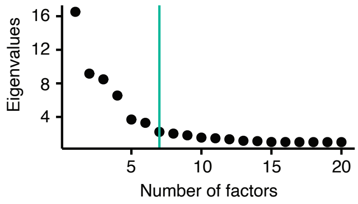


Figure 3.4: Screeplot Screeplot for factor solution on all 24 corpora.

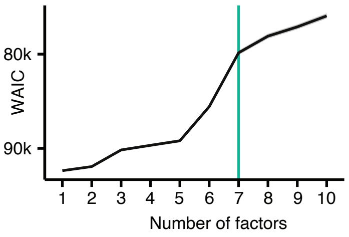


Figure 3.5: Watanabe-Akaike Information Criterion WAIC for models with an increasing number of factors. The dark area around the line is a 89% credible interval.

We, therefore, selected a seven-factor solution. The factors explain 12%, 11%, 10%, 10%, 6%, 4% and 4% of the variance (57% in total). Factor 1 "voice quality" mainly loads on Alpha ratio, Hammarberg index, and MFCC 1, 2, and 4 (see Figure 3.6 for the loading plot). Factor 2 "loudness" loads mainly on loudness and spectral flux. Factor 3 "pitch and formants" loads on fundamental frequency, on the formants (F_{1-3}), and mildly on HNR. Factor 4 "rhythm and tempo" mainly loads on durational features. Factor 5 "shimmer" loads on shimmer and mildly on HNR. Factor 6 'pitch variation" loads on pitch variation and jitter. Factor 7 "MFCC 3" loads on MFCC 3.

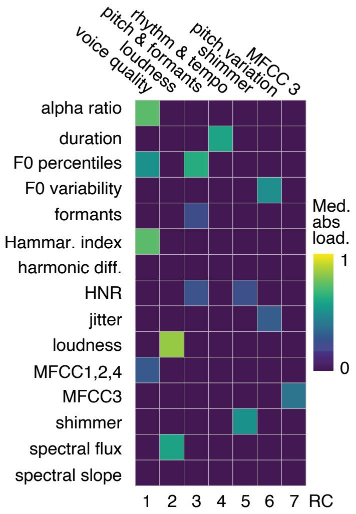


Figure 3.6: Factor loadings Factor analysis reveals seven acoustic dimensions that relate to perceptual qualities of speech prosody. To ease the visualization of the data, weak loadings (< .45) are not shown in the loading plot. The full loading plot can be found in [11].

To further assess the robustness of the factor analysis, we computed a seven-solution factor analysis for each of the four largest countries and largest languages, covering 87% and 89% of the data, respectively. We predict all data into the factor analysis of the respective country or language. For each country and each language pair, we compute the optimal alignment by maximizing the correlation between the dimensions. For each country and language pair, we compute an average correlation of each of the seven aligned factors. Some country and language pairs align better, but on average, we find a correlation of r = .67 and r = .65, indicating a fair overlap in factor solutions across the largest languages and countries.

3.2.4 Modelling

As described at the beginning of this chapter, we describe the mapping between speech prosody and emotion by using Bayesian multilevel multinomial logistic regression models (see Figure 2.1). The model learns a linear predictor of intercepts and coefficients for each of the seven factors. To obtain a prediction for a certain emotion for a particular speech sample, we multiply the acoustic factor values with the respective coefficients, sum the values up, and add the intercepts (see Figure 3.7).

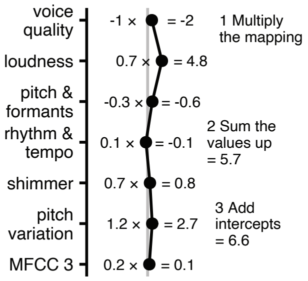


Figure 3.7: Obtain a prediction per emotion To obtain a prediction for a specific emotion, we take the mapping and multiply it with the respective acoustic factor values of some input stimulus, sum the values up, and add the intercepts.

This is repeated for all other emotions (except "neutral" which is the pivot category). The prediction is passed to a softmax and the model predicts the emotion with the highest value (see Figure 3.8). For an interactive version of this process, see `http://mapping-emotions.pol.works`.

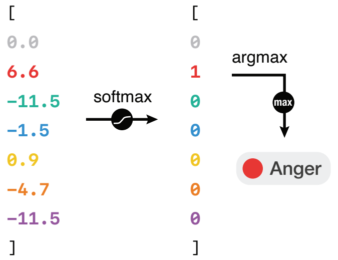


Figure 3.8: Model prediction Predictions for all six emotions. "Neutral" always obtains the prediction 0, as it is the pivot category. The seven values are converted into probabilities (softmax), and the emotion category with the highest probability is the category prediction for some input stimulus.

All multilevel models were fitted using the R package `brms` [294], which is a high-level interface to `Stan` [295]. The models use the categorical response distribution and logit link function. Where possible, standard normal priors are used (i.e., a normal distribution with a mean of 0 and a standard deviation of 1). The target distribution is explored using Hamiltonian Monte Carlo. The target acceptance rate is set to 99 % to avoid divergent transitions after warmup. To avoid exceeding the maximum tree depth, we set the hyperparameter to 12. For reproducibility, all models use the same seed. To speed up sampling, we use `cmdstan` as a backend. All models use eight chains, and we collected 4,000 posterior samples. The models reported in the paper were defined as follows:

- Corpus model: `emotion ∼ 1 + RC1 + RC2 + RC3 + RC4 + RC5 + RC6 + RC7 + (1 + RC1 + RC2 + RC3 + RC4 + RC5 + RC6 + RC7 | corpus)`,
- Null model: `emotion ∼ 1 + (1 | corpus)`,
- Base model: `emotion ∼ 1 + RC1 + RC2 + RC3 + RC4 + RC5 + RC6 + RC7 + (1 | corpus)`,
- Country model: `emotion ∼ 1 + RC1 + RC2 + RC3 + RC4 + RC5 + RC6 + RC7 + (1 + RC1 + RC2 + RC3 + RC4 + RC5 + RC6 + RC7 | country) + (1 | corpus)`,
- Language model: `emotion ∼ 1 + RC1 + RC2 + RC3 + RC4 + RC5 + RC6 + RC7 + (1 + RC1 + RC2 + RC3 + RC4 + RC5 + RC6 + RC7 | language) + (1 | corpus)`,
- Culture model: `emotion ∼ 1 + RC1 + RC2 + RC3 + RC4 + RC5 + RC6 + RC7 + (1 + RC1 + RC2 + RC3 + RC4 + RC5 + RC6 + RC7 | country:language) + (1 | corpus)`
- Big model: `emotion ∼ 1 + RC1 + RC2 + RC3 + RC4 + RC5 + RC6 + RC7 + (1 + RC1 + RC2 + RC3 + RC4 + RC5 + RC6 + RC7 | sex + country:language + speaker) + (1 | corpus)`

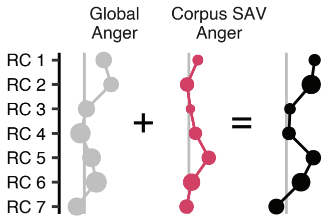


Figure 3.9: Example mapping for "anger" The model internally combines the population- and group-level effects. In this particular example, the estimates for "anger" in the corpus "SAV" for RC1–7 are depicted. The black line is the combined mapping, which is plotted in the following subplots. The larger the size of the dots, the smaller the credible interval.

3.3 Results

3.3.1 Low variability within and high variability across corpora

To assess the degree of variability in the mapping of emotional prosody, we fit a model that estimates a coefficient for each of the seven acoustic factors across the six emotions. On top of this "global mapping", a corpus-specific deviation from this coefficient is computed (see Figure 3.9). In doing so, we measure the variability of the mapping within a corpus and across corpora. The estimates are depicted in Figure 3.10.

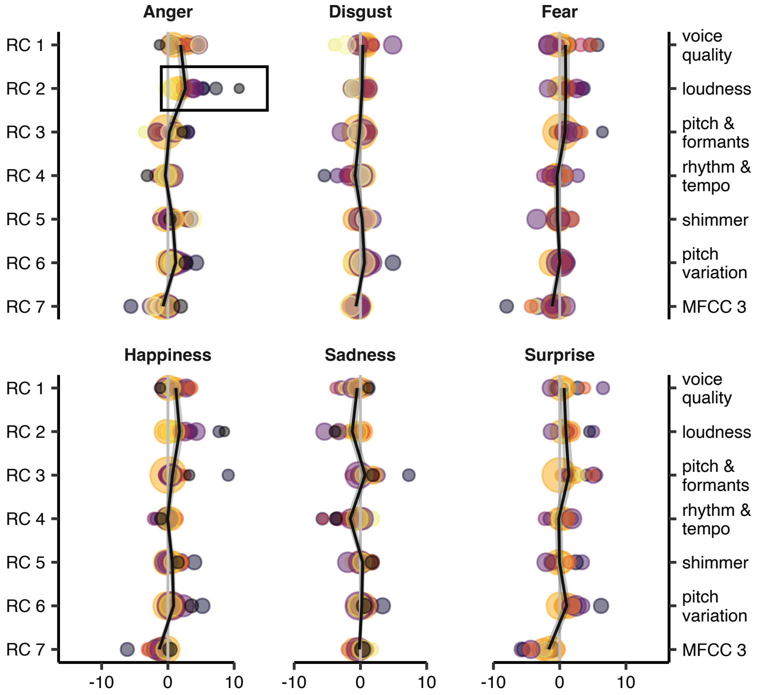


Figure 3.10: Model estimates Each colored dot represents a combined estimate for a specific corpus (average across corpora + corpus-specific estimate) of an acoustic factor (RC1–7) for all emotions. Large dots indicate small credible intervals (i.e., narrow distributions). The black line is the average coefficient, and the area around the line is a 89% credible interval. The vertical gray line indicates 0.

The estimated emotion coefficients across corpora mostly match empirical predictions from two reviews on emotion-specific acoustic profiles [39, 40]. For example, angry speech tends to be loud, and sad speech tends to be slow (see Figure 3.11).

However, the variability across corpora tends to be high. The variability within a corpus is characterized by the spread of the distribution of estimates. Wide distributions indicate more variability for the given estimate in a corpus (small dots indicate greater variability in Figure 3.10). On the other hand, variability across corpora can be described by the overlap in the estimated distributions across corpora. If there is a poor overlap of the distributions, then there is a great deal of variability across corpora.

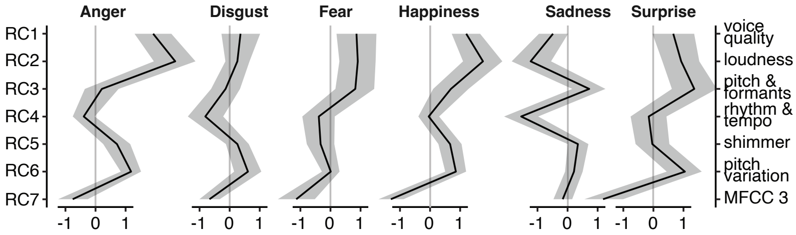


Figure 3.11: Global Mapping Zoomed in version of the global mapping presented in Figure 3.10.

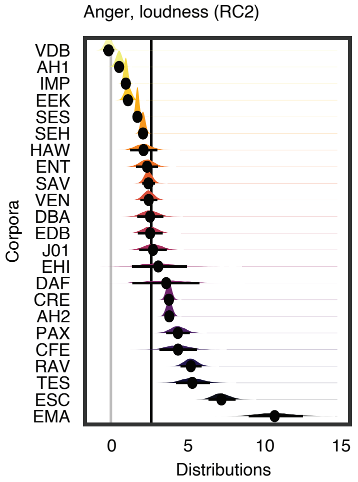


Figure 3.12: Zoomed-in mapping for "anger" Zoomed-in version of factor RC2 "loudness". The combined estimates per corpus (n = 4,000) rarely overlap. The black line below the distribution indicates an 89% credible interval. The vertical black line is the average coefficient (population-level effect), and the gray line is positioned at the origin.

In Figure 3.12, we zoom in on a single factor (RC2 for anger), and one can see that the estimates for the coefficients are rather tight (i.e., the distribution of estimates is narrow). This implies that the mapping of a certain acoustic factor to an emotion label is consistent within a corpus. However, across corpora, one can observe that the credible intervals of the distributions only partially overlap, which means that the estimates from one corpus to another often differ. If the mapping between acoustic features and emotion labels were identical across corpora, one would expect a greater degree of overlap. Given the observed variability in the estimates across corpora, the next step is to investigate the origin of the variability.

3.3.2 Identifying the origin of variability

Given that estimates across corpora are heterogeneous, we ran a series of models accounting for different moderators. Every model estimates a separate intercept for each corpus to account for possible imbalances in the base rate of emotions across corpora. Models are compared to each other using the Widely Applicable Information Criterion (WAIC), which provides an approximation of the out-of-sample deviance while penalizing overly complex models. Thus, the relative WAIC difference between contrasting models is important, where lower WAIC values indicate a better model fit.

As a lower boundary, we fit an intercept-only model estimating an intercept for each emotion and corpus. The "base" model additionally estimates a coefficient for each acoustic factor. As shown in Figure 3.13, the base model is much better than the intercept-only model.

Next, we fit a series of models inspired by the emotion dialect theory [189], based on the "in-group" effect. One way to model this membership is to add a group-level effect for languages and countries.

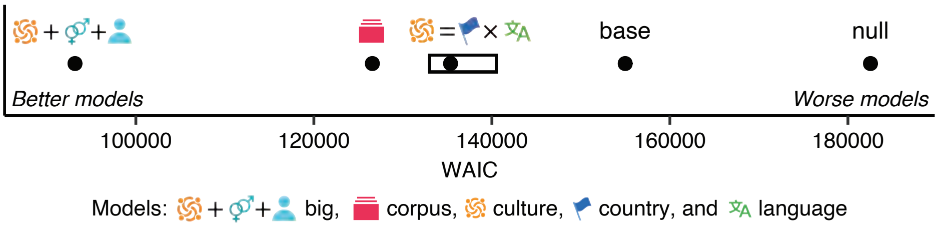


Figure 3.13: Model comparison The models from right to left: The null model containing only intercepts; the base model estimating the global mapping; the in-group model estimating the interaction between country and language; the corpus model from Figure 3.10; and the big model, which is the in-group model with additionally modeling speaker and sex differences. Error bars are standard errors of the WAIC.

As shown in Figure 3.14, the language model and the country model perform similarly well (the country model is slightly better). However, the combination of both categories "language" and "country" ("culture" model) performs even better, allowing for useful distinctions, such as between American and Canadian English.

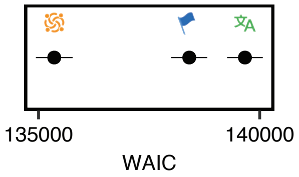


Figure 3.14: Model comparison for culture Zoomed-in version of the black box in Figure 3.13. It shows the WAIC of the in-group models modeling the group-level effect of countries, languages, or the interaction of both. Same legend as in Figure 3.13.

As shown in Figure 3.13, the culture model is outperformed by the corpus model from the reliability analysis (see lower non-overlapping WAIC value for the "corpus" model), as the grouping variable "corpus" contains the same grouping information as in "culture" — each corpus is usually assigned to one country and one language — and additionally consists of more specific information potentially relevant for the communication of emotion. For example, speakers are often recruited from the same area or institution (e.g., same city or university), targeting a more specific social group [188].

However, the grouping variable "corpus" is — in contrast to "language" or "country" — an artificial construct that is transcended by a series of more realistic constructs, such as cultural proximity and social belonging. We, therefore, extend the culture in-group model — and not the corpus model — by adding sex and individual speaker differences. As shown in Figure 3.13, this "big" model outperforms all other models.

The model comparison shows that models only assuming a global mapping are outperformed by models accounting for differences across cultures, sexes, and individual speakers. In the next section, we will investigate the relevance of these moderators for predicting the model.

3.3.3 Relevance of culture, sex, and individual differences

To examine how individual levels of the mapping contribute to the prediction of the model, we compute the contribution of each level of analysis to the prediction of the model. First, we obtain the model prediction on the data the model was fitted on, and we then measure how much each group-level contributes to the value for the predicted emotion (see Figure 3.15).

In all emotions (except "surprise") individual differences have the greatest impact on the model prediction. The second most important level of analysis is "culture" for most emotions, followed by the global mapping or sex differences.

Remarkably, only 20%–25% of the model prediction originates from the global mapping, as depicted by the pie charts in the right upper corner of each panel in Figure 3.15.

As depicted in Figure 3.15, the intercepts — marked by the darker color — play a subordinate role in the prediction of the emotion. In addition, the intercept of the corpus has in all emotions — except for "disgust" — the smallest contribution to the final prediction.

3.3.4 Variability is largest for speakers and cultures

In the previous analysis, we have shown that the contribution to the model prediction is largest for individual and cultural differences. In the current analysis, we investigate the degree of variability in the coefficients for different levels of analysis. We do this by extracting the posterior estimates for each

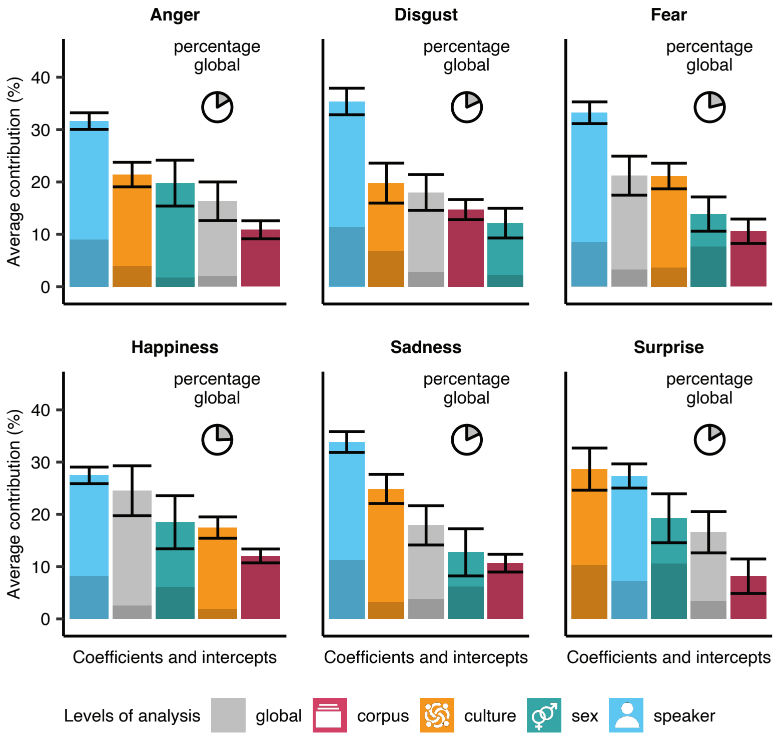


Figure 3.15: Contribution of different levels of analysis to the model prediction Each panel shows the mean contribution of different levels of analysis in all cases in which the emotion was predicted. The error bars are standard deviations across single posterior draws (n = 4,000). The color of the bars indicates the level of analysis. The darker section of the bars represents the contribution of the intercept. The lighter section represents the contribution of the acoustic coefficients. The pie chart in the upper right of each panel shows the contribution of global mapping to the full prediction.

acoustic factor, each emotion, and each group-level and computing the average standard deviation as a metric of the variability of the estimates.

As depicted in Figure 3.16, most variability can be found in the "speaker" and "culture" estimates. Overall, the first three acoustic factors (voice quality, loudness, and, pitch & formants) show the most variability (see inset in Figure 3.16). The remaining factors (except RC7, and MFCC 3) have decreased variability corresponding to their component number.

The variability results per emotion also show that the estimates for "speaker" and "culture" are the most variable. All estimates for the emotions are variable (see inset in Figure 3.17).

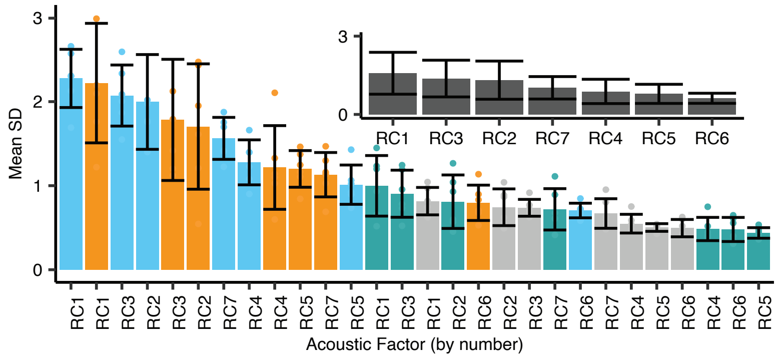


Figure 3.16: Variability of levels by acoustic factor Variability in the coefficients for different levels of analysis. For each group-level, emotion, and acoustic factor, a standard deviation was computed on all coefficients. In both panels, the average standard deviation is plotted by the acoustic factor. The error bars are in standard deviations. The subplots collapse over the different levels of analysis. The color legend is the same as in Figure 3.15.

3.3.5 Emotion expression across cultures, sexes, & individuals

In a follow-up analysis, we investigate the correlation between the mappings of different emotions across different levels of analysis showing how the expression of emotion differs across levels (see Figure 3.18). In this analysis, we correlate the mean mapping of an emotion to all other mappings of other emotions.

We start by correlating the global mapping across emotions. As depicted in the upper left panel of Figure 3.18, “sadness” is the only emotion with a distinct profile — as it has only a strong correlation with itself and low correlations with all other emotions. Interestingly, the profiles of the other emotions correlate more strongly with each other, especially the correlations among the profiles for “fear”, “happiness”, and “surprise” (note the visual resemblance across those emotions in Figure 3.11).

In three further analyses, we describe the relationship between emotions across sexes, cultures, and individuals. A first analysis showed that the mapping for a specific emotion correlates most strongly with the mapping for the same emotion of the other sex (right panel of Figure 3.11), For instance, female anger is, on average, closer to male anger compared to any other emotion. When compared to the global mapping, adding sex further increases the correlation among the profiles of “fear”, “happiness”, and “surprise”.

Adding “culture” or “speaker” to the global mapping leads to a substantial decrease in the overall correlations across emotions, indicating that the mapping for individual cultures and speakers is relatively distinct. The overall drop in correlation is greater for speakers than for cultures, confirming the pattern of results in the previous analyses (see Figure 3.18). Nonetheless, the diagonals are mildly preserved, indicating that the mapping for a given emotion is more similar across speakers and cultures than to another emotion.

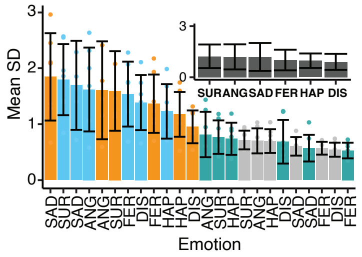


Figure 3.17: Variability of levels by emotions Identical plot as in Figure 3.16, but now the average standard deviation is plotted by emotion.

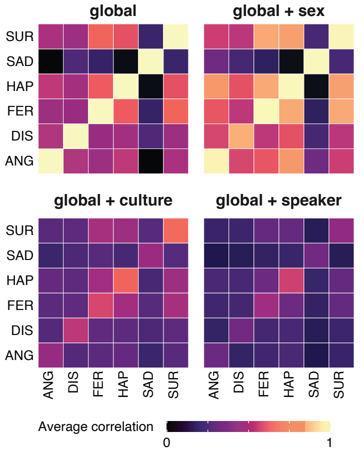


Figure 3.18: Correlation across mappings Left upper panel: mappings of all emotions correlated with each other. Diagonals are always 1. The remaining three panels: Correlation between the global mapping with sex, cultural, or speaker difference. The fill color is the average correlation (Pearson).

3.4 Discussion

3.4.1 Summary

Using a Bayesian modeling framework, we have shown that:

- there is considerable variability in the mapping of emotions to speech prosody (Section 3.3.1),

- differences across sexes, cultures, and individuals contribute to this variability (Section 3.3.2),
- individual differences followed by cultural differences are the most important for the prediction of the model (and the global mapping only contributes 20%–25%; Section 3.3.3),
- the variability across acoustic coefficients is largest for speakers and cultures (Section 3.3.4), and
- the expression of emotion differs more across speakers than across cultures and cultures than across sexes (Section 3.3.5).

The observed variability is most compatible with constructivist theories, predicting emotions are perceptually variable, and appraisal theories, predicting different appraisals will lead to different emotions. It is least compatible with affect program theories that predict low variability across sexes, cultures, and individuals because of the assumption of innate action programs. However, the theory also in-theory explanations for larger variability, such as:

- The notion of "refinement" [59] (the assumption that there are refined emotions within a basic emotion, e.g., "hot anger" as a subtype of "anger") might lead to more variability across cultures.
- Display rules [296] (some cultures might not allow the portrayal of certain emotions like "sexual pleasure") might lead to more variability across cultures.
- The auditory modality might not be sufficient to capture the full range of emotions, as information from other modalities is missing [297]. This might lead to more variability across individuals.

However, meta-analytic investigations like the current study do not allow to dismiss or confirm emotion theories, because they in part have in-theory explanations for the observed variability and the study has exposed several limitations present in emotional prosody corpora.

3.4.2 Limitations

The exposed limitations concern the stimulus space ($\mathbf{X}$), the semantic space ($\mathbf{Y}$), and conducting cross-lingual comparisons ($\mathbf{Z}$):

X1 **Overreliance on actors.** Most corpora rely on actors to portray emotions that are not representative of the general population. [2]

X2 **Unnatural prompting.** The present study is based on acted corpora, often involving an unnatural emotion prompting in which the actor has to record a sentence with a neutral meaning for a particular emotion. These recordings lack ecological validity and thus might not represent the expression of emotions in daily life [298–300] and might reproduce stereotypes [200, 232].

X3 **Production bias.** Not all individuals are physically capable of producing speech that conveys a particular emotion, especially if they have no training. But even for trained actors, producing particular emotions might be difficult (e.g., display rules).

X4 **Mismatch between felt and expressed emotions.** When prompting actors to record a sentence in a particular emotion, the actor might not feel the emotion they are portraying.

X5 **Lack of standardization.** The corpora are recorded in different settings, participants receive different instructions, and the recordings are made

[2] This is not specific to corpora of intended emotions but also applied to spontaneous recordings, e.g. actors engaging in an improvisation task.

with different equipment. This might lead to noise in the data and thus enhance the variability across corpora.

Y1 **Assumption of a single emotion.** Corpora of intended emotions assume that only a single emotion is communicated in a recording. However, in real life, emotions are often mixed [232].

Y2 **Assumption of existing emotion taxonomy.** By selecting a list of emotions to be recorded, the experimenter implicitly assumes the existence of particular emotions (and the non-existence or irrelevance of others).

Y3 **Unclear alignment of subtypes.** In meta-studies, one has to summarize over emotion labels. For example, Juslin & Laukka [40] count the emotions "afraid", "anxiety", "frightened", "scared", "panic", "terror", and "worry" all to the category "fear". However, it is disputable if they all refer to the same concept.

Z1 **Lost-in-translation.** The problem of unclear alignment of subtypes of emotions (e.g., "panic" and "fear") is further amplified once emotional concepts are translated. Since English is a lingua franca, emotion labels tend to be translated into the closest English equivalent, which is often not straightforward or even possible [60, 61]. This introduces a bias towards English-speaking cultures.

The subproblems of the stimulus space ($\mathbf{X}$) contribute to the stimulus selection problem, the semantic space ($\mathbf{Y}$) contributes to the taxonomy curation problem, and $\mathbf{Z}$ is the lost-in-translation problem.

3.4.3 Outlook

In the following three chapters, we will describe Human-In-The-Loop (HITL) approaches that address these limitations. In Table 3.1, we summarize the contributions of each paradigm. In each chapter, we will explain how the paradigm exactly solves the problem. The problem of overreliance on actors is solved for all paradigms because the HITL involves online participants who are a more general sample of the overall population.

Problem	GSP	GAP	STEP
X1 Overreliance on actors	✓	✓	✓
X2 Unnatural prompting	✗	✓	-
X3 Production bias	✓	✗	-
X4 Mismatch between felt and expressed emotions	✗	✓	-
X5 Lack of standardization	✓	✓	-
Y1 Assumption of a single emotion	✗	✓	✓
Y2 Assumption of existing emotion taxonomy	✗	✓	✓
Y3 Unclear alignment of subtypes	-	-	✓
Z1 Lost-in-translation	-	-	✓

Table 3.1: Core problems in studying the expression of emotion Checkmark indicates the paradigm addresses the problem, a cross indicates the paradigm does not address the problem, and a dash indicates the problem is irrelevant to the paradigm.

Chapter 4

Gibbs Sampling with People

Based on

Peter M. C. Harrison, Raja Marjieh, Federico Adolfi, **Pol van Rijn**, Manuel Anglada-Tort, Ofer Tchernichovski, Pauline Larrouy-Maestri, and Nori Jacoby. 2020. 'Gibbs Sampling with People'. *Advances in Neural Information Processing systems*.

Pol van Rijn, Silvan Mertes, Dominik Schiller, Peter M.C. Harrison, Pauline Larrouy-Maestri, Elisabeth André, and Nori Jacoby. 2021. 'Exploring Emotional Prototypes in a High Dimensional TTS Latent Space'. *Interspeech*.

Dominik Schiller, Silvan Mertes, **Pol van Rijn**, and Elisabeth André. 2021. 'Analysis by Synthesis: Using an Expressive TTS Model as Feature Extractor for Paralinguistic Speech Classification'. *Interspeech*.

A longstanding question in psychology is whether there exist acoustic profiles for particular emotions and whether they differ across cultures [39, 40, 42, 119]. To efficiently study this, one would like to associate a particular emotional label with a paradigmatic speech example. At the same time, one wants to avoid involving actors and limit production biases because it can significantly alter the mapping [299, 301], since asking people to depict a particular emotion does not necessarily make them feel the emotion. Gibbs Sampling with People (GSP) is a method to address these constraints, specifically, it can be used to sample a representative sample of a particular semantic category, and it can do that without relying on overt production. GSP is a HITL iterative procedure that optimizes stimulus features to subjectively match a given verbal description (such as "happy"), which makes it suitable for high-dimensional feature spaces like prosody.

Here, we show that this procedure can be used to sample a paradigmatical exemplar of a verbal descriptor from the space of emotional prosody in a principled way. The same method could also be applied to other domains as discussed below.

4.1 Background

Psychologists, cognitive scientists, and more recently computer scientists, have developed various tools to study human representations. Traditionally, these representations have been studied by presenting participants with stimuli and asking them to make judgments. In cognitive science, this is commonly done by presenting participants with pairs of stimuli and asking them how similar they are. From these similarity judgments, one can infer a representation of the stimuli in a high-dimensional space using Multidimensional Scaling (MDS) [302–305] (see Figure 4.1). Another example from psychology is reverse correlation [306], in which participants are shown the same base stimulus (e.g., an image of a face) with noise added to the stimulus. In each trial, the participant is asked to categorize the stimulus (e.g., the face looks either masculine or feminine) [307]. By averaging the noise patterns across selected and not selected trials, one can infer the mental prototype that participants use to categorize the stimuli (see Figure 4.2).

The problem with both approaches is that they require a large number of trials. For example, collecting similarity judgments for a large number of stimuli quickly becomes infeasible, since the number of judgments required is quadratic to the number of stimuli – for N stimuli, $(N*(N-1))/2$ judgments are required.

In both pairwise similarity judgments and reverse correlation, trials do not depend on previous trials. In the general background (Section 2.2), we showed that one can speed up the exploration of high-dimensional stimulus spaces by adding dependence across trials. In particular, we presented Markov Chain Monte Carlo with People (MCMCP), a HITL algorithm that integrates human judgments into the MCMC algorithm to explore high-dimensional stimulus spaces.

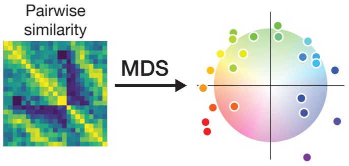


Figure 4.1: Multidimensional Scaling (MDS) A matrix of all pairwise similarities (left) is transformed into a low-dimensional space (right) such that the distances in the low-dimensional space approximate the similarities in the high-dimensional space. Data from [7].

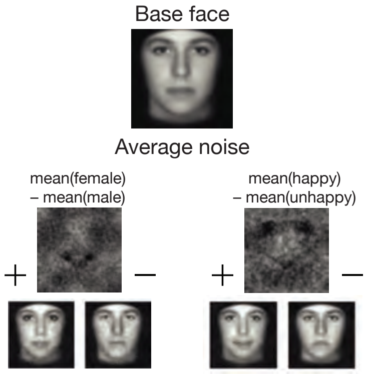


Figure 4.2: Reverse Correlation In each trial, pixel noise is added to the same base face (top). After collecting all the trials, the mean noise patch for the selected category (e.g., "happy") and the not selected category (e.g., "unhappy") is computed. The two noise patches are subtracted. By adding the resulting noise patch to the base face, one can infer the mental prototype that participants use to categorize the stimuli. Data from [307].

4.2 Paradigm

One problem with MCMCP is that it requires participants to make a binary choice, which is only one bit of information per trial. So to explore a high-dimensional stimulus space, one needs a lot of trials. The second problem is that the proposal distribution has to be defined, which is not always straightforward and requires some trial-and-error. While this is not a problem for computer algorithms (you can do a grid search over hyperparameters), it can become costly when running experiments with people. To overcome this problem, we developed Gibbs Sampling with People (GSP) [1].

4.2.1 Gibbs Sampling

A Gibbs sampler [308] is another instance of a Markov Chain Monte Carlo algorithm. The idea is that one can sample from a high-dimensional distribution by sampling from the conditional distribution of each dimension, given the other dimensions. So concretely, this means that one starts with a vector state $\mathbf{z}^{(0)} = (z_1^{(0)}, \ldots, z_n^{(0)})$ which is usually a random sample, where n is the number of dimensions. Then one iteratively updates each dimension (z_k) — where k is the index of the current dimension and v indicates how often all dimensions have already been visited — by sampling from:

$$p(z_k^{(v+1)} | z_1^{(v+1)}, \ldots, z_{k-1}^{(v+1)}, z_{k+1}^{(v)}, \ldots, z_n^{(v)}) \tag{4.1}$$

4.2.2 Gibbs Sampling with People

In Gibbs Sampling with People (GSP) the participant controls the value of the current dimension (z_k) by controlling a slider. The participant is instructed to move the slider to optimize for a particular target, for example, the color that is most similar to the concept of "sky". In GSP, each point (i) of the slider $(x = \{z_k^v\}_i)$ is associated with a utility value $(U(x))$. Where the utility function consists of a deterministic loss while keeping the other fixed dimensions (z_{-k}) fixed and a noise component associated with perceptual noise (n_k^v):

$$U(x) = \ell(x, z_{-k}) + n_k^v \tag{4.2}$$

To reduce the noise associated with a particular trial, one can assign the same chain to multiple participants (so the perceptual noise does not come from an individual participant) and potentially aggregate the responses of the participants to average out the noise, for example by taking the mean (Figure 4.3).

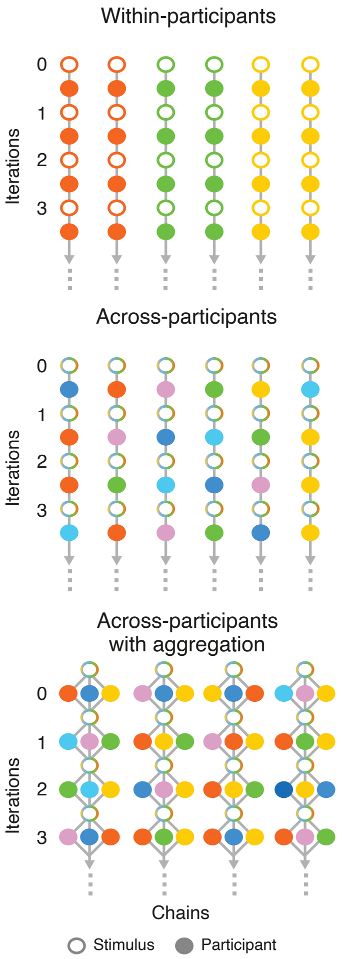


Figure 4.3: Reducing participant noise Chains can be assigned to single participants, across participants, or responses of participants can be aggregated (e.g. taking the mean or median response).

4.2.3 Comparison to MCMCP

To compare GSP to MCMCP, we conducted a study in which participants had to explore the three-dimensional space of color. Participants had to match the color to the following targets: “chocolate”, “cloud”, “eggshell”, “grass”, “lavender”, “lemon”, “strawberry”, and “sunset”. The color space was parametrized by hue, saturation, and lightness (Figure 4.4, see also [309]).

We compared three conditions (all across-participant chains): MCMCP, regular GSP, and GSP with aggregation (mean over 10 participants). Each chain was 30 iterations long, 5 chains were run per target, and each participant contributed up to 40 chains (see Figure 4.5 for the participant interface). 422 participants were recruited from Amazon Mechanical Turk recruitment platform (MTurk) and had to pass a color-blindness test and a color-vocabulary test before continuing with the online experiment (see paper for details [1]).

In Figure 4.6A, we plot the individual chains for the different conditions. We can see that the MCMCP chains are noisier than the GSP chains, and sometimes don’t even converge to the color prototype (e.g., the third “lemon” chain). Also, the aggregated GSP chains are much cleaner than the regular GSP chains, indicating that the aggregation of responses can reduce the noise associated with a single trial. One might argue that the poor performance of MCMCP is due to the selection of a poor proposal distribution, however, in a follow-up experiment in which we did a grid search over proposal distributions, we found little difference in performance between the different proposal distributions. Another concern might be that aggregated MCMCP would be equally good as aggregated GSP, however, in follow-up experiments, we found that aggregated MCMCP still does worse than aggregated GSP (see paper for details [1]).

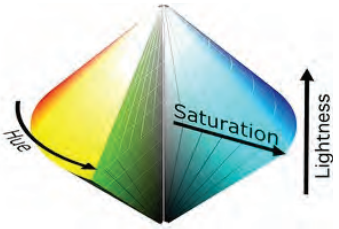


Figure 4.4: HSL Space Color space parametrized by hue, saturation, and lightness. Copyright Mike Horvath (CC BY-SA)

To validate our results, we conducted a validation experiment in which participants had to rate how well the samples matched the target color (see Figure 4.6B). We found that the aggregated GSP chains were rated significantly higher than the regular GSP chains and the MCMCP chains (see Figure 4.6C). From these findings, we conclude that GSP is a more efficient way to explore high-dimensional stimulus spaces than MCMCP (even when taking into account that a GSP trial takes longer than a MCMCP trial) and that aggregating responses can reduce the noise associated with a single trial.

4.2.4 Finding a good parametrization

While GSP only has a few assumptions concerning the parametrization of the stimulus space (e.g., continuous dimensions), in practice, we found that it can sometimes be challenging to find a good parametrization.

While there is no general prescription for constructing a good parametrization, the following properties are desirable:

- **Independence**: The dimensions are independent, namely, they operate on different features of the stimulus space. Circulating over highly correlated dimensions could result in unnecessarily long chains.

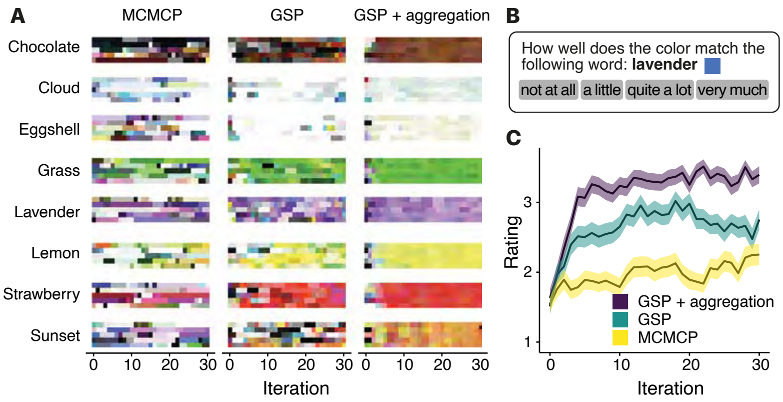


Figure 4.6: Color results **A** Individual chains for the different conditions. **B** Interface for the validation experiment. **C** Validation results.

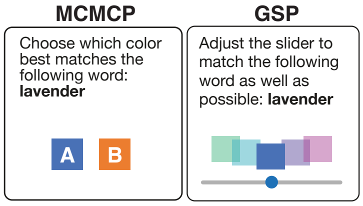

MCMCP
GSP
Choose which color best matches the following word: lavender
A
B
Adjust the slider to match the following word as well as possible: lavender

Figure 4.5: Color interface Interface for MCMCP (left) and GSP (right).

- **Perceptual alignment**: The dimensions are perceptually aligned, namely, the individual slider manipulations result in intuitive changes to the stimulus.
- **Smoothness**: Changes in the dimensions should result in smooth changes in the stimulus (i.e., a small change in the slider should not lead to an abrupt change in the stimulus).
- **Consistency**: The dimensions should be consistent (e.g., a slider of a face model where the first part of the slider changes the sex and the second part of the slider changes the age is not consistent). The consistency will also make the sliders more predictable, allowing for a more efficient exploration of the stimulus space.
- **Completeness**: All sliders allow for a complete exploration of the stimulus space. For example, only changing the red and blue color channels of a color space, but not the green channel, would not be complete.
- **Efficiency**: The stimulus space should not have inactive dimensions, namely, dimensions not relevant for the mental representation (e.g., when exploring the space of voices, linguistic prosody is not relevant).
- **Synthesis**: The dimensions are synthesizable in real-time (this can become a challenge when using generative deep-learning models).

While we applied GSP to four modalities in the paper [1], we will only describe the results of the emotional prosody experiment in the next section since we designed and conducted those experiments.

4.3 Emotional prosody

4.3.1 Methods

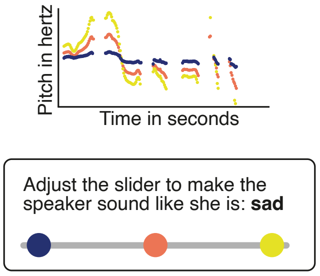

Pitch in hertz
Time in seconds
Adjust the slider to make the speaker sound like she is: sad

Figure 4.7: GSP interface for prosody Moving the slider changes only one prosody dimension. In this example, the range of the pitch contour changes.

In the emotional prosody experiment, participants use sliders to control one prosody dimension at a time (see Figure 4.7). We use three phonetically balanced sentences with a neutral meaning from the Harvard sentence text corpus [310], which were recorded by a female speaker [311]. Participants modify seven prosodic dimensions (all conducted in Praat [133] via the Python wrapper Parselmouth [292]) in the following order:

- **Pitch level**: From the extracted pitch contour (pitch floor: 100 Hz, ceiling: 500 Hz), participants shift the pitch contour up or down in the range of [-37, 37] Hz.

- **Pitch range**: Scaling the pitch contour in the range of [0.2, 1.8] centered at the mean pitch.
- **Pitch slope**: Changing the pitch contour in the range of [-37, 37] Hz, where the slope is defined as the difference between the first and last pitch value. We therefore added the linear function $f(t)$ to each pitch point:
$$f(t) = x * \frac{t - t_0}{t_1 - t_0} \tag{4.3}$$
- **Jitter**: The amount of frequency perturbation by applying Gaussian noise in the range of [0, 0.0001] to the glottal pulses. This mimics irregularities in the opening and closing of the vocal folds (see Table 2.1) and thus adds an impression of roughness to the voice [116].
- **Duration**: The speed of the fragment in the range of [0.8, 1.2] times the original duration.
- **Tremolo depth**: Tremolo is the periodic variation of the intensity of the sound. We choose to control tremolo because shimmer (the intensity perturbation) is difficult to control in existing recordings, and tremolo applies a similar effect. The tremolo depth is the amplitude of the intensity variation in the range of 0.01 dB (inaudible) to 10 dB (strong effect).
- **Tremolo frequency**: The frequency of the intensity variation is in the range of [0, 5] Hz.

Participants have to change the sliders such that the speaker either sounds: “angry”, “happy”, or “sadness”.[1] In contrast to the other experiments, the prosody experiment does not start at a random point in the stimulus space but starts with applying no effects (this does not violate any GSP assumptions) since a random initialization would lead to a too distorted voice in the beginning.

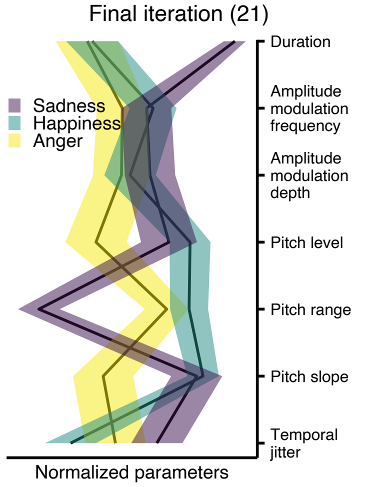


Figure 4.8: Acoustic profile for the final iteration 95% confidence intervals over participants.

4.3.2 Results

We recruited 110 participants from MTurk[2] who contributed 220 within-participant GSP chains with each 21 iterations (so two chains per participant). The mean feature values are plotted for the final iteration in Figure 4.8. One can see that the acoustic profiles of the three emotions are fairly distinct. For example, sad speech is associated with longer durations and a low pitch range, whereas happy and angry speech is associated with short durations and a high pitch range (note this is also consistent with the results from my meta-study, see Figure 3.11 and other previous research [42]).

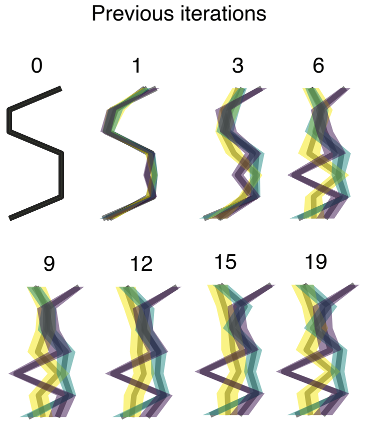


Figure 4.9: Acoustic profiles over iterations 95% confidence intervals over participants.

In Figure 4.9 we plot the mean feature values over the iterations. One can see that the acoustic profiles start to deviate, but start to stabilize once all dimensions have been visited once.

This is also consistent with the validation results (N = 161), where the ratings increase steadily for the first sweep of the parameter vector and then plateau with a reliable mean contrast of 0.9 points (see Figure 4.10). The contrast is the mean rating for the correct emotion minus the mean rating for the incorrect emotions. Thus, if the contrast is > 0, the ratings for the intended emotion are higher than the ratings for the not-intended emotions.

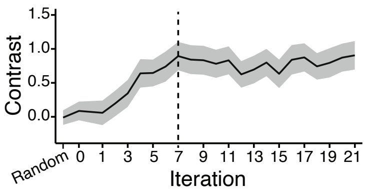


Figure 4.10: Contrast ratings The dashed line indicates one sweep of all dimensions. 95% confidence intervals over participants.

In another experiment (N = 210), we replicate our findings by running the experiment with across-participant chains. Figure 4.11 shows that the mean acoustic profile is not altered substantially if one takes the mean feature values over all within-participant chains or all across-participant chains.

[1] I selected those three emotions because they occupy distinct positions within the valence arousal space.

[2] All participants passed a headphone check [312].

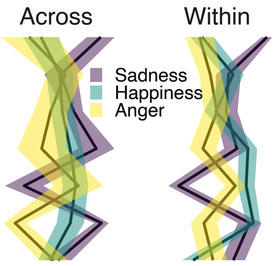


Figure 4.11: Across vs. within chains 95% confidence intervals over participants.

4.3.3 Summary

Compared to existing voice manipulation experiments [313, 314], GSP allows for efficient exploration of the stimulus space (on average visiting each dimension just once). Also, the produced stimuli are recognizable as the intended emotion by a separate group of participants in the validation experiment. These results show that GSP is an efficient algorithm to explore emotional prosody.

4.4 Expressive Text-To-Speech

While the previous study allowed for identification acoustic profiles associated with particular emotions, it has three major limitations:

- **Subjective selection**: By focusing on the seven acoustic features, we made strong assumptions about which acoustic manipulations are relevant for the expression of emotions (and one can see that while we assumed tremolo is relevant for the communication of emotions, it is not used by participants).
- **Low audio quality**: Changing acoustic features in existing recordings can lead to unnatural and distorted speech. This is because many acoustic features are correlated to each other (e.g., a pitch contour correlates with spectral properties of the sound), and modifying the acoustic features independently of each other can create artifacts.
- **Limited expressivity**: Traditional hand-crafted features such as pitch slope and pitch range struggle to capture the full expressivity of underlying pitch or intensity contours (e.g., it does not allow sampling from all possible pitch contours possible for the particular recordings).

4.4.1 Methods

To solve these problems, we use a deep learning model to generate the speech. The quality of Text-To-Speech (TTS) models has improved substantially over the last years, and they can now generate speech that is indistinguishable from human speech [315]. Recent developments have also allowed for controlling the sentence, speaker identity, and prosody in isolation from each other [316]. The rationale here is that if one has a TTS model that can learn a distinct representation of the prosody and the model is trained on expressive recordings, human participants can sample from all possible prosodies for particular sentences.

4.4.1.1 GST Tacotron

At the time of the study, GST Tacotron [316] was a popular TTS model that allows for control of the prosody of the generated speech. GST Tacotron extends the Tacotron base model [317] by adding the following components (see Figure 4.12A):

- **Reference encoder**: The input recordings are converted to a Mel spectrogram. The encoder now has to learn to map the Mel spectrogram to a fixed-size vector (called “reference embedding”).

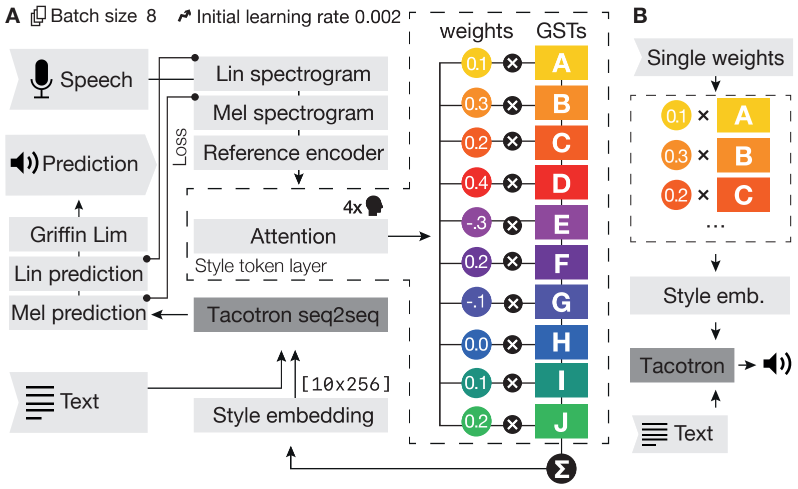


Figure 4.12: Tacotron architecture A Schematics of GST Tacotron. **B** Manipulation of style weights.

- **Style token layer**: The reference embedding is now fed into an attention module and then into the style token layer. The attention module learns a mapping between the reference embedding and a bank of "Global Style Tokens", hence applying weight to each of the learned 'styles'.
- **Style embedding**: The weighted average of the Global Style Tokens is called the style embedding, which is then fed into the Tacotron model together with the text embedding.

Instead of extracting the GST weights from the reference embedding, they can also be set via GSP (Figure 4.12B). The GST Tacotron has desirable properties for GSP: (i) the model can generate high-quality speech (at least for the time we wrote the paper), (ii) it learns self-emerging prosodic styles from the training data, and (iii) the weights allow to make smooth interpolations across styles (which would naturally happen in GSP).

We trained the model[3] on the Blizzard 2013 dataset, which consists of 9,741 segmented English utterances from expressive audiobook recordings by the professional speaker Catherine Byers [318]. After 380,000 epochs we stopped the training since we did not observe further improvement.

4.4.1.2 Experiment setup

We used the same targets as in the prosody experiment ("angry", "happy", and "sadness") and also used the Harvard sentences [310] as stimuli.

Each slider consisted of 32 equidistant points and the range of all dimensions is constrained to [-0.24, 0.38], corresponding to a 94% confidence interval of the attention weights given by the model in the training data (trade-off between expressivity and distortions created by using extreme slider values). When synthesizing audio, we used a fixed seed to make the audio samples deterministic. We used across-participant chains, because (i) the results of the prosody experiment showed that the mean feature values are not altered substantially, and (ii) due to the autoregressive nature of the TTS model, the generated speech takes quite some time to generate.

Since aggregation worked well for the domain of color, we also used aggregation for the weights (see Figure 4.13). We used median aggregation to avoid obtaining a sound none of the participants ever listened to (e.g., a mean value can lie between two slider selections). Every chain starts with all weights set

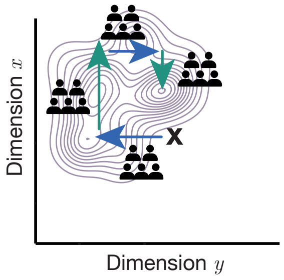
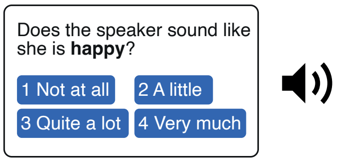


Figure 4.13: Aggregating over responses Each slider is presented to 5 participants, and the median response propagates to the next iteration. Over iterations, participants explore dense regions associated with a particular emotion in the feature space.

[3] https://github.com/syang1993/gst-tacotron

to 0. 130 US participants recruited from MTurk engaged in the experiment, finishing 39/45 chains after (20 iterations).

The results were validated by a separate group of participants ($N = 82$) who rated how well the samples matched the intended emotion on a four-point scale (see Figure 4.14). We also created 156 transfer stimuli by applying the median attention weights of the final GSP iteration to four novel sentences from the Harvard sentence corpus and adding 18 random samples. On average, every stimulus was rated 4.5 × for every emotion.


Does the speaker sound like she is **happy**?

1 Not at all
2 A little
3 Quite a lot
4 Very much

Figure 4.14: Example validation trial Audio plays automatically and the user is prompted to select one option.

4.4.2 Results

In Figure 4.15 we plot the first two principal components of the style embeddings of the chains at iterations 9–20.

One can see that the three emotions separate moderately well on these two components, indicating that the emotional sentences group together in the latent space regardless of the sentence.

This grouping manifests from early on (again after approximately visiting each dimension once), providing additional support for early convergence of the GSP process (see Figure 4.16).

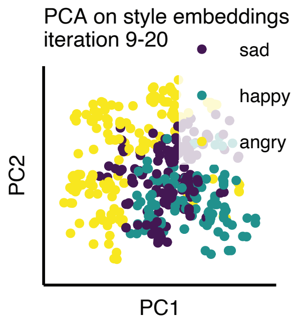


Figure 4.15: PCA on style embeddings Principal Component Analysis (PCA) on style embeddings of 39 chains at iterations 9–20.

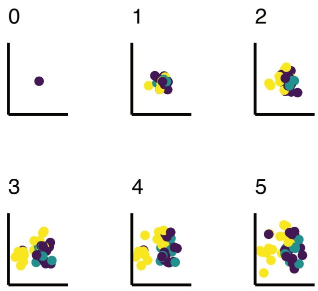


Figure 4.16: Style embeddings over iterations Development over iterations in Principal Component style embedding space at iterations 0–5.

In Figure 4.17, we plot the average ratings for the initial sample (iteration 0), binned iterations, the transferred prosody, and the random sample. One can see that the ratings for the intended emotion steadily increase throughout the iterations, whereas the ratings for non-intended emotions plateau or drop. Interestingly, there are imbalances in the rating of the initial and random samples, representing some perceived biases (e.g., iteration 0 sounds more happy than sad).

To summarize all ratings, we computed the contrast between the ratings (same as in the prosody experiment, Figure 4.10). The contrast shows that over the course of iterations, the intended emotion reliably achieves higher ratings than the non-intended emotions. Also, the transferred prosody receives a similarly high rating as the final binned iteration (see Figure 4.17 and 4.18). This provides additional evidence that the emotional prototypes are separated from the sentence and thus can be transferred to new sentences.

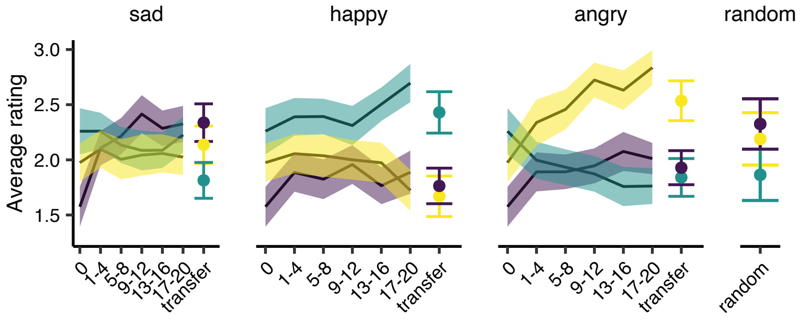


Figure 4.17: Average rating split by emotion Average ratings for the initial sample (iteration 0), binned iteration 1–4, 5–8, 9–12, 13–16, 17–20, the rating for the transfer and random sample (95% confidence intervals).

4.5 Discussion

4.5.1 Summary

In this chapter,

- We introduced Gibbs Sampling with People (GSP), a novel paradigm that allows for efficient exploration of high-dimensional stimulus spaces.
- we showed that GSP is more efficient than other methods such as reverse correlation and Markov Chain Monte Carlo with People (MCMCP)
- we applied GSP to the domain of emotional prosody and showed that the intended emotion of the produced stimuli is recognized.
- we improved the synthesis quality and limited expressivity of the prosody experiment by using a deep learning model that allows us to control the prosody of the generated speech.
- we showed that the emotional prototypes are separated from the sentence and the emotional prosody can be transferred to new sentences.

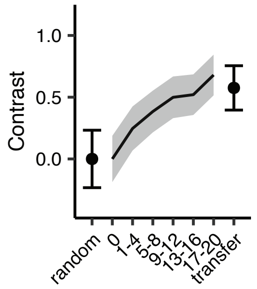


Figure 4.18: Contrasts between ratings 95% confidence intervals.

4.5.2 Contributions to Emotional Prosody

When GSP is applied to the domain of emotional prosody, it allows solving two identified problems (see Table 3.1):

- **Production bias**: GSP relies on a generative model controlled by participants via sliders, so participants do not have to produce the emotional prosody themselves.
- **Lack of standardization**: Related to the previous point, all participants receive the same slider interface and go through the same hardware checks (e.g., wired headphones), leading to a standardized procedure for all participants.

4.5.3 Limitations and Outlook

While the results are promising, several limitations need to be addressed in future research:

- **Generalization**: Both experiments only included three emotions, whereas the space of emotions is much larger.
- **Multi-speaker**: Both experiments only included one (female) speaker. Future research should investigate if prosody prototypes are speaker-specific and also if there are sex differences in the communication of emotions. In Chapter 10, we show how GSP can be used to create a specific voice. One possible direction would be to do some double GSP where participants first have to find a voice for a particular speaker and then have to find the prosody for a particular emotion.
- **Cross-lingual**: Currently, the experiment has only been conducted in English. It would be interesting to see if the results generalize to other languages, especially to tonal languages, like Mandarin or Thai.
- **Cultural context**: Finally, the experiments only included US participants recruited from MTurk, which is a very narrow population and unrepresentative of the world population. Follow-up research should include more heterogeneous populations and also train on non-Western

and non-English corpora to make valid claims about emotional prosody and to develop robust applications [246].

- **Stereotypes**: The prototypes might be stereotypical and might not fully represent how emotions are communicated in real life [200, 299]. Future research should investigate how the prototypes are perceived in different contexts and how they can be used to create more nuanced emotional expressions.

Chapter 5

Genetic Algorithm with People

Based on

Pol van Rijn, Harin Lee, and Nori Jacoby. 2022. 'Bridging the Prosody GAP: Genetic Algorithm with People to Efficiently Sample Emotional Prosody'. *CogSci*.

Existing corpora of emotional prosody rely on existing emotion taxonomies. For example, acted emotional speech corpora involve actors producing speech prosody for a particular emotion or spontaneous corpora require raters to annotate the emotion in speech segments, thus relying on pre-assumed taxonomies. Mediating the selection of stimuli through assumed emotion categories is problematic because it can lead to an unrepresentative sample of all emotional prosodies and training a model on such a corpus leads to an impaired representation of all possible prosodies. For example, for supervised models, this is problematic because the model can only predict emotions it was trained on, or for self-supervised models, the model might lack an acoustical representation typical for that emotion (e.g., screaming for "anger"). GAP tries to overcome this problem, by creating a high-quality corpora without any pre-specified taxonomy.

In this chapter, we will describe how corpora of emotional prosody are constructed, discuss the limitations of existing methods, and provide a Human-In-The-Loop (HITL) approach that can solve most of these limitations.

5.1 Background

5.1.1 Corpora of Emotional Prosody

Representative sampling from emotional recordings is particularly difficult, because emotional events are fairly rare in everyday speech, and emotional cues are usually transmitted using multiple simultaneous cues (e.g., body posture, facial expression, sentence meaning, prosody), which makes it hard to study how emotions are communicated through prosody (other cues also provide emotional information).

Corpora of emotional prosody – collections of emotional speech recordings – can be divided into two groups: corpora of intended and corpora of perceived emotion.

5.1.1.1 Corpora of intended emotion

In intended corpora, participants – usually actors – are asked to say the same sentence for multiple emotions. The sentences are usually constructed by the experimenter to capture a wide range of possible emotions (e.g., "Let me tell you something" [42, 46, 319]) or are jabberwocky sentences with no meaning (e.g. “I nestred the flugs”, [287]). This approach has the following advantages: It allows to study of emotional cues in isolation and provides high experimental control – both over the emotion elicitation procedure, the recorded sentences, and the (usually high-quality) recording.

However, the approach also has the following limitations:

- **Limited size**: Corpora of intended emotion usually consist of a small number of recordings involving different recordings of the same sentence by a limited number of speakers (potentially overemphasizing idiosyncrasies of individual speakers).
- **Overreliance on actors**: Saying the same sentence for different emotions is not easy for most people and thus requires training. Therefore, most corpora of intended emotion rely on professional actors. This makes the speakers of the corpus a very particular subsample of the full population (certain demographic groups are overrepresented) and makes acted corpora expensive to create (actors are usually paid more than laymen).
- **Unnatural elicitation procedure**: Each corpus of intended emotion uses its own emotion elicitation procedure. For example, some corpora only give the emotion word [44], whereas other corpora try to induce emotions by using scenarios [46]. This raises the concern if the portrayed emotion is also the felt emotion of the speaker – which largely depends on the emotion induction method (e.g., just reading the label “sad” probably does not make the speaker feel sad). Another concern is that most corpora of intended emotion try to induce a single emotion, which is not representative of real emotions in daily life since the expression of emotion is embedded in a larger context that is missing here. Taken together, this raises concerns about the authenticity of the expressed emotion.
- **Sentence bias**: The sentences read by the actors are usually chosen by the experimenter, but it often remains untransparent why this sentence was selected. For example, it is unclear if a neutral sentence can carry all emotions (e.g., “Let me tell you something” is more likely to carry an angry prosody than a disgusted prosody). Sentences also differ in the amount of emphasis they allow on words. One concern is that speakers in corpora of intended emotions overemphasize certain prosodic cues [119] and thus that the recordings consist of exaggerated or stereotypical prosodic patterns that do not reflect natural variations in emotional expression. Related to this concern is that if the same speaker reads the same sentence many times for the same emotion, the recordings start sounding less authentic due to fatigue or boredom.
- **Pre-assumed semantic space**: To create a corpus of intended emotion, the creators of the corpus have to make a selection of emotions that are used to elicit emotions in the speakers. This implies that the selection of stimuli is mediated through the selected categories of emotion. It is problematic to select such a corpus if one wants to make a statement about the semantic space of emotions [212] since the space of possible emotions was already constrained by selecting a subset of possible emotions.

5.1.1.2 Corpora of perceived emotion

Corpora of perceived emotion use (human) emotional annotations to identify and label emotional segments in speech recordings. The speech recordings can come from data scraped from naturalistic contexts [288, 289, 320] or by recording participants in game-like activities such as acting improvisation [321] or user interactions with robots [322]. The corpora vary in the amount of experimental control: some corpora use scraped recordings from the internet or from movies [320], whereas other corpora involve interactions between actors [49]. Compared to corpora of intended emotion, corpora of perceived emotion: are more scalable, and can be less expensive to create (do not necessitate actors, although in practice they often do [49, 50, 321]) and are more natural (the elicitation of emotions is embedded in a larger context and the spoken content is determined by the speaker and not by the experimenter).

However, perceived emotion corpora have the following limitations:

- **Consent**: The recordings are often scraped from the internet or from movies, which raises concerns about the consent of the speakers [323].
- **Inefficient data collection**: Since emotional events are rare in everyday speech, raters have to listen to a large number of speech recordings to find emotional segments.
- **Assumption that intended emotions can be recognized**: The labels given in perceived emotion corpora are based on the emotion label given by the raters. This builds upon the controversial assumption that emotions can be recognized at an above-chance level from the voice. However, low agreement across raters indicates that annotators quite often do not agree [324]. Furthermore, there is no consensus on cutoff values to call a fragment an instance of a particular emotion — such as the minimal number of ratings per stimulus or the minimal rating. One criticism expressed for intended emotion corpora – that the procedure leads to stereotypical depictions of emotions – also applies to corpora of perceived emotion: Recordings that are rated highly on a particular emotion, are not necessarily valid depictions of that emotion, but instead might also lead to the selection of prototypical depictions.
- **Entangled emotional cues**: The emotional information is encoded simultaneously through multiple channels [271, 273] but some channels are entangled with one another and cannot easily be separated (e.g., if you say angry things, you probably also say it with an angry prosody). This does not allow us to dissect whether the judgment provided by the annotators arises from a single channel (e.g., prosody) or a combination of channels.
- **Pre-assumed semantic space**: Corpora of perceived emotion also have to make a selection of emotions that are used to identify emotional segments in the speech recordings.

5.1.1.3 Synthetic corpora

A third category of corpora consists of synthetic corpora, generated using generative speech models. These approaches leverage expressive Text-To-Speech (TTS) models, which can manipulate prosody and other paralinguistic features [e.g., 316, 325, 326] (see Section 4.4), or Emotional Speech Synthesis (ESS) models, which incorporate an additional emotional supervision signal [for an overview: 327].

Expressive TTS models are typically trained on expressive speech datasets, such as audiobook recordings, that capture a broad range of affective states. However, since these datasets lack explicit emotion labels, HITL approaches can be used to identify emotionally expressive speech within the generative model. For instance, Mertes, Don, Grothe, Kuch, Schlagowski, and André proposed a HITL method for synthesizing voices that reflect specific personality traits [328]. This approach is an instantiation of MCMCP[64] (see Section2.2.3), where the generative model proposes a new sample based on the current state (akin to a mutation step), and a human rater evaluates whether it is an improvement (analogous to selection in an evolutionary algorithm). In Chapter 4, we demonstrated that GSP is a more efficient search method for high-dimensional spaces than MCMCP[1]. Using this procedure, we identified emotional speech samples within an expressive TTS model [3]. However, like other corpus-based approaches, this method remains constrained by predefined emotion spaces—optimizing for a fixed set of emotions does not necessarily capture the full diversity of emotional prosodies.

ESS models, by contrast, rely on corpora of intended (see Section 5.1.1.1) or perceived emotions (see Section 5.1.1.2) for training and thus inherent biases and limitations.

To address these challenges, we propose a novel method for constructing emotional prosody corpora that mitigates many of these issues.

5.2 Paradigm

5.2.1 Create and Rate

To overcome these problems, we developed a new paradigm called "Create and Rate" that is inspired by MCMCP. In MCMCP, a Markov chain is constructed by proposing new states close to the current state and by asking participants to accept or reject the move. In "Create and Rate", the proposals are not generated by an algorithm, but come from the creations of participants. In contrast to MCMCP, the proposal is human-generated and the acceptance criterion is not a binary decision. This has three advantages: (i) rating multiple creations instead of just two means more information per rating, (ii) the human creations are not random moves around the current state but represent human biases, and (iii) multiple raters implement a form of human aggregation.

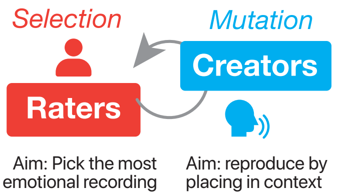

Selection
Mutation
Creators
Raters
Aim: Pick the most emotional recording
Aim: reproduce by placing in context

Figure 5.1: Create and Rate Illustration of Genetic Algorithm with People for Prosody. Creators generate mutant recordings of the previous generation by recording themselves saying the sentence as if they were in the same situation. Raters select the most emotional recording and hence the Darwinian selection is applied.

In the paradigm, "raters" are either rating (e.g., on a Likert scale) or selecting (forced choice) creations. The creations are generated by participants who view the selected previous generation's creation and then create a new one. The paradigm supports all kinds of creations, for example, text input (e.g., asking participants to describe all details in an image), voice recordings, or even a generative model (e.g., you can do GSP with human aggregation, where the same slider is presented to multiple creators and the raters pick the best slider). To select a creation for the next generation, the paradigm supports different aggregation methods (e.g., majority vote or highest average rating) and can in principle take multiple selection criteria into account (e.g., select descriptions of images that are both highly detailed and concise).

One can think of this paradigm as a genetic algorithm with people, in which the creations are mutations of the previous creation and the ratings implement selection (Figure 5.1).

[1] Because MCMCP yields only one bit of information per trial, whereas GSP provides information gain proportional to the slider's granularity.

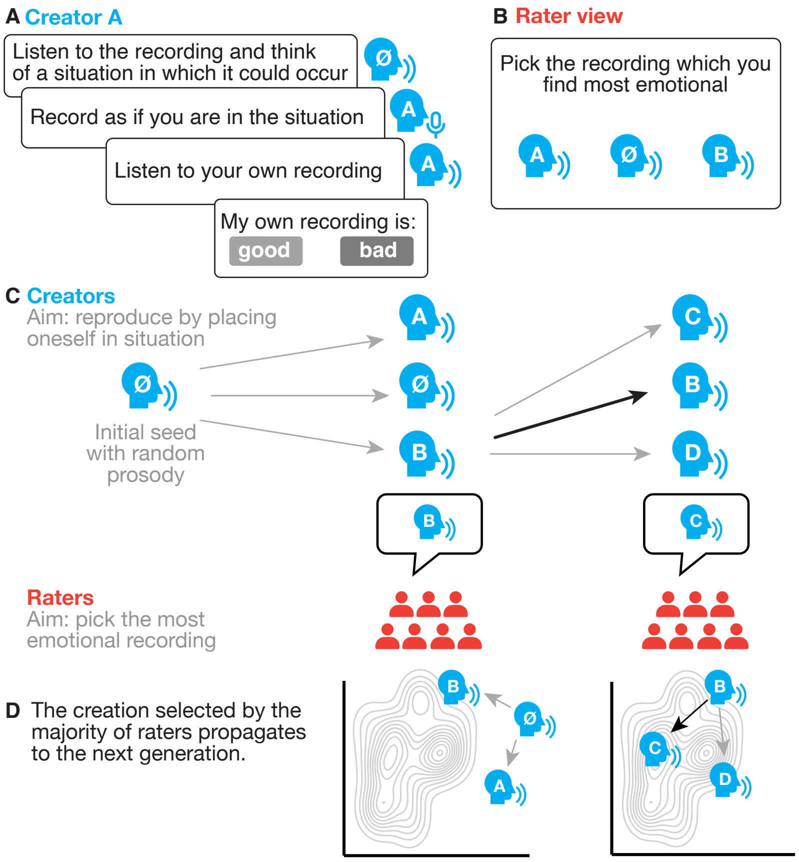

A Creator A
Listen to the recording and think of a situation in which it could occur
Ø
Record as if you are in the situation
A
Listen to your own recording
A
My own recording is:
good
bad
B Rater view
Pick the recording which you find most emotional
A
Ø
B
C Creators
Aim: reproduce by placing oneself in situation
Ø
Initial seed with random prosody
A
Ø
B
C
B
D
B
C
Raters
Aim: pick the most emotional recording
D The creation selected by the majority of raters propagates to the next generation.
B
Ø
A
B
C
D

Figure 5.2: Genetic Algorithm with People **A** Each creator listens to the creation from the previous generation (here the very first recording ø) and has to think of a situation in which the recording could occur. Next, they record themselves as if they were in the same situation as the previous speaker, followed by a playback of their recording to confirm that the recording is correct. **B** Raters listen to the creations and the recording presented to the creator (random order) and select the most emotional recording. **C** Schematics of the beginning of a chain. The initial recording is presented to the first two creators A and B. The seven raters select one recording. The majority vote (B) propagates to the next iteration. The process repeats in the next generation. **D** Over the course of generations, participants explore denser regions of the emotional prosody space.

5.2.2 Genetic Algorithm with People

We developed Genetic Algorithm with People (GAP), a variant of the Create and Rate, to especially solve limitations in corpora of emotional prosody. In this paradigm, two creators listen to the recording of the previous creator (see Figure 5.2A). When listening to the recording, they are asked to imagine themselves in the same situation as the speaker and then record themselves saying the sentence. They then listen to their own recording to confirm that the recording is correct (e.g., it captures the intended situation, the sentence is correctly spoken). The raters then listen to the recordings of the creators and select the most emotional recording (see Figure 5.2B). To avoid breaking progress in the chain (e.g., both creators have difficulty recording the sentence), the recording of the previous generation is always added to the set of proposals. The recording with the majority vote is then selected for the next generation.[2] The process is repeated for 10 generations. Each participant had a fixed role because raters and creators have different tasks that could potentially influence one another. This procedure differs from other corpora using emotional mimicry [331] in that GAP is an iterative procedure amplifying the emotional content over generations and that the creators are not aware they should optimize for emotionality, avoiding the production of stereotypes. The schematics of the experiment are shown in Figure 5.2C. Over the generations, the speech prosody becomes more and more emotional (see Figure 5.2D).

[2] Previous literature has shown that introducing a majority voting approach reduces participant error [329] and improves the quality of productions [330].

5.3 Methods

5.3.1 Sentences

Prior to conducting the experiment, we compiled a list of 10 semantically neutral sentences. Eight of these sentences are included in the two emotional corpora, which are later used in the validation experiment [42, 46]. The remaining two sentences come from the phonetically balanced and semantically neutral Harvard sentence corpus [310].

5.3.2 Initial seed generation

For each sentence, we generated five speech recordings each with a different speaker using the expressive TTS model Flowtron [325] trained on LibriTTS (2,456 speakers). We chose this TTS model, because (i) at the time of the experiment, it was one of the best available expressive TTS models (generating high-quality synthetic speech, while having varied prosody), and (ii) it was trained on a large number of speakers allowing to start each chain with a different speaker. We manually inspected if recordings contained glitches and replaced them with new ones when necessary.

5.3.3 Validation

To validate the robustness of our paradigm, we recruited an independent group of participants to provide annotations for the recordings, as well as the recordings obtained from two existing emotional prosody corpora [42, 46]. In addition to the stimuli created with GAP, we presented 20 recordings each from the 6 categories of emotion in CREMA-D [46], and 10 recordings each from the 11 categories of emotions in the US subset of VENEC [42].

For each recording, participants answered the following questions: perceived strength of emotion ("how emotional was the speech?", 4-point scale), valence ("how negative or positive was the speech?", slider value ranging from -50 to +50) and arousal ("how low or high in energy was the speech?", slider value ranging 0 to +100), and were asked to type a single word related to the mood that best describes the state of the speaker in the recording.

5.3.4 Pre-Screening

All participants had to pass a lexical decision task (LexTALE) [332] to ensure they were fluent speakers of English and had to pass a check to make sure they were wearing headphones [312]. Creators had to pass two additional tasks. The first task consisted of distinguishing good from bad recordings. Bad recordings were defined as recordings that were silent, contained too much noise, repeated sentences, or sentences cut out too early. If participants made more than one mistake in this test, they could not participate in the experiment. In the second task, participants were asked to reproduce the heard sentences and were excluded if there was a textual mismatch (identified using Google's speech-to-text API transcription).

In the validation experiment, we excluded participants posthoc who repeatedly gave the same answers for text labels (at least 4 unique words for 20 trials) and ones who had low response consistencies ($r < .40$) between the main experiment trials and the repeated trials at the end.

5.3.5 Participants

Participants were recruited from MTurk, had to reside in the US, had a HIT approval rate of over 99% on the platform, and had completed at least 2,000 tasks. 126 participants completed the GAP study and 131 participants completed the validation. All participants received monetary compensation at a rate of $9 an hour for their participation.

5.4 Results

5.4.1 Created stimuli

All stimuli generated in the experiment chains can be explored through an online, interactive visualization: https://polvanrijn.github.io/prosody-GAP/. The stimuli were validated by an independent group of participants.

5.4.2 Recordings become more emotional over iterations

In Figure 4.14, we show the average emotion strength for neutral stimuli, the seed stimuli of GAP, the selected creations by the raters, and the stimuli from CREMA-D and VENEC. The TTS-generated initial seed stimuli are slightly less emotional (mean = 1.79, sd = 0.80) than the neutral stimuli from CREMA-D and VENEC (mean = 2.09, sd = 0.86), but the creations become more emotional over the generations. They reach a plateau around the 6th generation, where the last generation (mean = 2.79, sd = 0.86) was slightly higher than CREMA-D (mean = 2.70, sd = 0.97) but lower than VENEC (mean = 3.11, sd = 0.88).

In line with this, we show that the average arousal and absolute valence of the recordings increase over the generations (see Figure 5.4) also indicating the recordings become more emotional over the generations.

The Kernel Density Estimation (KDE) showed dense concentration around the center of the valence and arousal 2-dimensional space (see Figure 5.5). The seed of GAP and the neutral sets of the other two corpora also showed comparable levels of arousal and absolute valence.

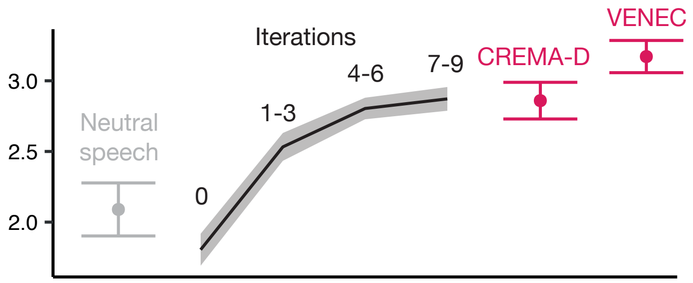


Figure 5.3: Emotion strength rating Average rating on the strength of emotions for the neutral stimuli (gray) and emotional stimuli from the VENEC (blue) and CREMA-D (green) corpus. The initial generation of GAP is bin 0 and the following generations are binned into 1–3, 4–6, and 7–9. The area and the error bars represent 95 % confidence intervals.

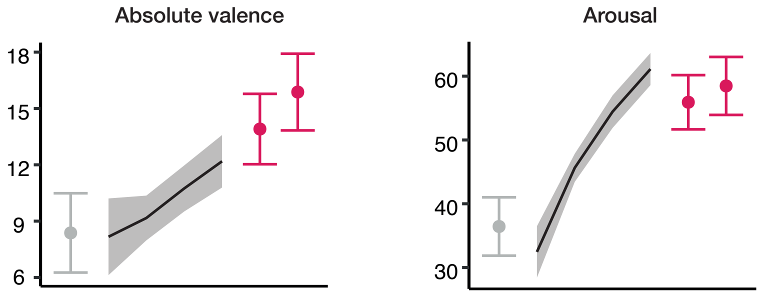


Figure 5.4: Valence and arousal rating
Average arousal and absolute valance.

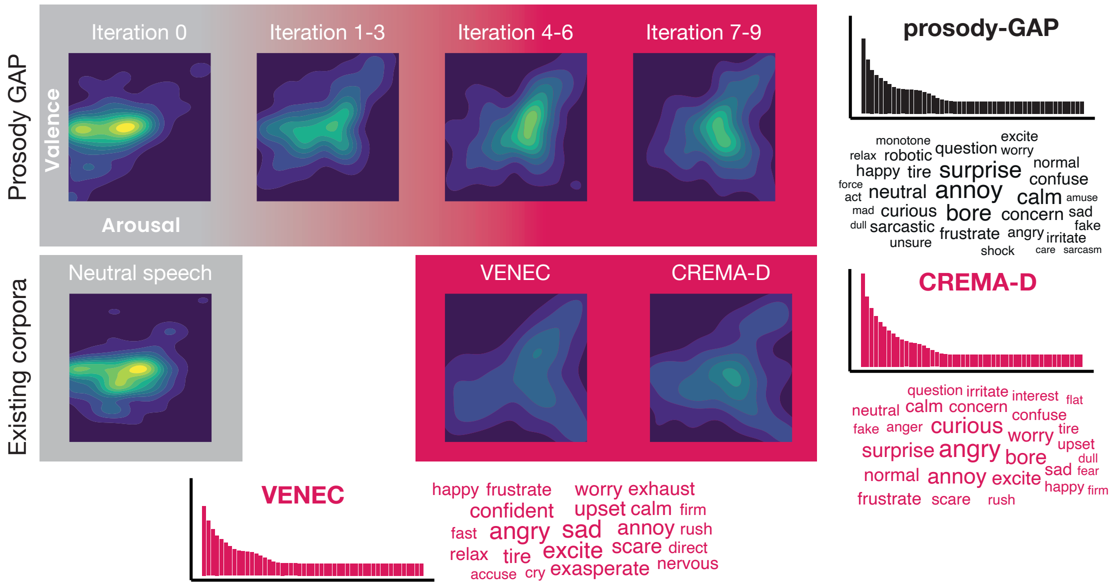


Figure 5.5: Kernel density estimates, word frequency distribution, and word clouds Arousal and valence KDE for single iterations and reference corpora. Word clouds of free-text responses and bootstrapped word frequency distributions.

Moreover, the coverage of the valence-arousal space dispersed over the generation, and by the last generation, one can observe that the covered regions were similar to CREMA-D and VENEC (see Figure 5.5).

5.4.3 Samples span a wide array of emotions

To further investigate the number of emotions captured by the recordings, we analyzed the word labels provided by the annotators in the validation experiment (see word clouds and word frequency distributions in Figure 5.5). We quantified the variability and the term-frequency distributions and made comparisons across the three datasets.

For GAP, we took the stimulus of the last generation where the ratings converged in each of the 50 independent chains (i.e., recording that the rater group judged as most emotional). Since the size of the stimuli was unbalanced across the three sets (GAP = 50, CREMA-D = 100, VENEC = 100; excluding neutral stimuli), we sampled 50 stimuli from CREMA-D and VENEC at random to match the stimuli size. We then computed 1,000 bootstraps without replacement in each set by randomly drawing 100 word label samples from all responses (where each stimulus can have multiple annotations made by independent annotators). All word labels were lemmatized using the `textstem` R package.

To measure variability, we counted the number of unique word labels in each of the bootstrapped samples. The results showed that both GAP (mean = 72.1, sd = 3.30) and VENEC (mean = 73.9, sd = 3.85) obtain comparable variability, and higher value on average than CREMA-D (mean = 64.0, sd = 4.16). High variability indicates that more diverse semantic labels are present, covering a wider range of semantic vocabulary associated with emotions, whereas low variability suggests that the frequency of words is concentrated in a smaller subset of words. Considering VENEC consists of more emotion categories than CREMA-D (11 and 6, respectively), higher variability for VENEC was expected. The fact that GAP achieves comparable variability to VENEC is indicative of the large breadth of emotion space GAP is able to capture.

5.5 Discussion

5.5.1 Summary

In this chapter,

- We proposed a new approach, Genetic Algorithm with People (GAP), for efficient sampling of the high-dimensional emotion prosody space by introducing genetic algorithms with human raters.
- We showed that with GAP the speech became more emotional over the generations (Figure 5.3) and also showed that the mean arousal and absolute valence increased over the generations (Figure 5.4).
- We quantified, using a word-frequency analysis, that the variability of the word labels obtained using GAP was comparable to VENEC and better than CREMA-D (Figure 5.5) indicating that GAP captures a wide array of emotions.

Overall, these results demonstrate the robustness of GAP and suggest that (i) emotional speech can be obtained in a less biased way without the prior assumptions of emotion categories, and (ii) the obtained recordings from online crowd-sourced samples can achieve comparable results to carefully curated corpora generated in professional settings. Furthermore, the convergence of emotional levels around the 6th generation demonstrates the efficiency of our method and shows highly promising potential for its scalability.

5.5.2 Evaluation and Limitations

In the Background section of this chapter (Section 5.1.1), we have outlined the limitations of existing corpora of emotional prosody. We will now explain how GAP at least in part addresses these limitations (see Table 5.1).

Problem	intended	perceived	GAP
Limited size	✗	✓	✓
Overreliance on actors	✗	✓	✓
Unnatural elicitation procedure	✗	✓	✓
Sentence bias	✗	✓	∼
Inefficient data collection	✓	✗	✓
Assumption that intended emotions can be recognized	✓	✗	∼
Entangled emotional cues	✓	✗	✓
Pre-assumed semantic space	✗	✗	✓

Table 5.1: Solving problems of corpora of emotional prosody Checkmark indicates the paradigm addresses the problem, a cross indicates the paradigm does not address the problem, and a tilde indicates the problem is only solved partially.

- **Limited size**: GAP is scalable (both in the number of stimuli as in the number of cultures sampled from) and can generate a large number of emotional recordings in a relatively short amount of time.
- **Overreliance on actors**: In contrast to corpora of intended emotion, speakers do not have to say a sentence in a particular emotion but instead have to imagine themselves in the same situation as the previous speaker. This is a much easier and more natural task than in intended emotion and does not need to involve trained actors. While we acknowledge that the online crowd workers are not a perfect representation of the general population either, the diversity of the participants is likely to be higher than in corpora of intended emotion.
- **Unnatural elicitation procedure**: Listening to a recording and then imagining oneself in the same situation is a more natural task than having to say a sentence in a particular emotion. For this task, it's also more likely that the portrayed emotion is the felt emotion of the speaker since you have to imagine yourself in the same situation. It also overcomes the problem of inducing a single emotion, since the emotion is embedded in a larger context (imagined by the participant).
- **Sentence bias**: While in GAP the sentences are also chosen by the experimenter (and the sentence bias is thus not completely removed), the creators read the same sentence much less often than in corpora of intended emotion, which reduces the risk of overemphasizing certain prosodic cues and potential fatigue or boredom.
- **Inefficient data collection**: Instead of searching for emotional segments in a large corpus as in perceived emotion corpora, in GAP, the prosody becomes more emotional over the generations (due to the evolutionary pressure asserted by the raters), which makes it an efficient way to sample from the emotional prosody space. Moreover, because the recordings can be collected online, it can considerably reduce the expenses and resources that are often necessary for a more traditional corpora curation (inviting participants to the lab, booking recording studios, manual annotation, etc).
- **Assumption that intended emotions can be recognized**: This problem is only partially solved by GAP. While we do not ask raters to recognize particular emotions, we do ask them to select the most emotional recording (implying they can detect emotionality from speech). Also, we acknowledge that the term "emotion" might be associated more with certain emotion terms. For example, people might think more of words such as "anger" or "happy" and less of words such as "surprise" or "disgust" when they hear the term "emotion". And these associations might differ between cultures.
- **Entangled emotional cues**: Since the same sentence is propagated throughout the generations, we can control for the influence of sentence meaning on emotion perception.
- **Pre-assumed semantic space**: In the paradigm, participants are either creators or raters. Creators do not know we are optimizing for emotionality and we explicitly avoid the use of words such as "emotions" or "feelings" in the experiment text. This allows us to minimize the potential biases in prompting participants to produce stereotypical emotions. Raters on the other hand are only asked to select the most emotional recording, and we thus do not make assumptions about the dimensionality of the emotional space.

5.5.3 Contributions to Emotional Prosody

From the previous summary, it becomes apparent that GAP solves the following identified problems for corpora of emotional prosody (see Table 3.1):

- **Unnatural prompting**: see "Unnatural elicitation procedure" (see the numeration in the previous section).
- **Mismatch between felt and expressed emotions**: since creators have to imagine themselves in the same situation as the previous speaker, it is more likely that the portrayed emotion is the felt emotion of the speaker than just a shallow imitation.
- **Assumption of a single emotion**: creators are not asked to produce a single emotion, but to imagine themselves in the same situation as the previous speaker.
- **Assumption of existing emotion taxonomy**: see "Pre-assumed semantic space" (see the numeration in the previous section).
- **Lack of standardization**: creators and raters go through various automated quality checks (e.g., microphone noise, headphone check, ASR transcript matching) to ascertain high-quality recordings.

5.5.4 Future work

While the current study only included US participants, the method can be easily extended to other languages and cultures (especially because it's scalable online and it only relies on a few language-specific tools that are now available for many languages [333, 334]). This makes it a particularly valuable tool for conducting research on emotional prosody cross-culturally and for low-resource languages.

Some recordings had – in spite of extensive screening tasks – poor audio quality. In future work, we plan to implement better audio control to screen for participants with bad microphone quality to improve the overall recording quality of the corpora.

Furthermore, we want to construct a more principled way for selecting sentences for initial generations that do not have to rely on an existing corpus and that can be more easily extended to other languages.

Chapter 6

Sequential Transmission Evaluation Pipeline

Based on
Raja Marjieh, **Pol van Rijn**, Ilia Sucholutsky, Theodore Sumers, Harin Lee, Thomas L. Griffiths, and Nori Jacoby. 2023. 'Words Are All You Need? Language as an Approximation for Human Similarity Judgments'. *ICLR*.
Raja Marjieh, **Pol van Rijn**, Ilia Sucholutsky, Harin Lee, Nori Jacoby, and Thomas Griffiths. 2024. 'Characterizing the Large-Scale Structure of Grounded Semantic Network'. *Cognition*.
Raja Marjieh, Ilia Sucholutsky, **Pol van Rijn**, Nori Jacoby, and Thomas L. Griffiths. 2024. 'Large Language Models Predict Human Sensory Judgments across Six Modalities'. *Scientific Reports*.

Identifying emotional concepts [335] and aligning them across languages is difficult [212]. In emotional prosody, participants are usually presented with a predefined list of emotions [42, 120, 191, 233, 271, 272, 287]. This involves adopting a predefined taxonomy, which is often based on previous research or is curated by the researcher.[1] However, this approach has several limitations. First, the manual creation process introduces researcher bias. Distinctions relevant to researchers might not be perceptually relevant. For example, researchers often distinguish between hot (short and impulsive) and cold anger (subtle and controlled) [335, 341], but it's unclear whether these distinctions are perceptually meaningful. Second, if taxonomies are sourced from previous research, it is unclear if they cover the full range of concepts in the dataset. Third, by using existing taxonomies, the taxonomy cannot be extended, thus the discovery of new concepts is not possible. Finally, often taxonomies are language- or culture-specific [246] and may not generalize across cultures or languages. This becomes a problem if taxonomies from one language are superimposed on another, or if the goal is to compare taxonomies across languages (e.g., a dictionary translation [40, 52] is not likely to yield an optimal alignment).

To solve these problems, We propose a novel adaptive tagging pipeline, called Sequential Transmission Evaluation Pipeline (STEP), that involves the creation of new tags and rating of existing tags. We describe previous work on HITL annotation pipelines and show how STEP extends these ideas. We will also show how STEP can be used to predict similarity judgments and how it can be used to show that grounded semantic networks show small-world and scale-free regularities.

6.1 Background

6.1.1 Games with a purpose

Games With A Purpose (GWAP) are a class of games that are designed to solve a specific problem while entertaining the players [342]. The most famous

[1] Notably, this limitation also applies to many machine learning datasets including ImageNet [336], COCO [337], Kinetics [338], AudioSet [339], Places [340] and others.

example of a GWAP is the ESP Game [343], where two players are shown the same image and have to guess the same word. If they guess the same word, they get points. The paradigm has also been applied to other modalities (e.g., music [344]) and other tasks [345, 346]. While GWAPs have been developed to create large-scale annotated datasets for machine learning, they have also been used to study human behavior [342, 343, 347].

6.1.2 Cascade

One limitation of the GWAP design is that participants are incentivized to come up with likely tags for the stimulus (which tend to be the most salient aspects of the stimulus) and might not capture all details. Cascade [348] solves this problem with a free creation step, followed by two filtering steps (selection and categorization enforcing the consensus of multiple participants). The paradigm consists of the following steps:

- **Creation**: Participants provide one tag for a given stimulus.
- **Selection**: Participants see all tags given for the stimulus and select the best one.
- **Categorization**: Participants see one stimulus and all tags assigned to it and have to categorize each tag (present or not present).

The three-step process is repeated if none of the tags is selected in the categorization step (this implements the consensus with the creators). The main problem with Cascade is the cost of the process, costing the same or more than hiring experts. Follow-up work suggested the process could be made more efficient by using an early stopping criterion, that stops the most costly step (categorization) earlier [349].

6.1.3 Visual Genome

Visual Genome [329] is a dataset developed to improve the performance of computer vision models. It contains structured annotations for 108,077 images from the MS COCO dataset [337] and YFCC [350]. It involved a sophisticated human annotation pipeline consisting of the following steps:

- **Region Descriptions**: Participants describe regions by drawing bounding boxes and providing a description (e.g., "yellow fire hydrant", "woman in shorts is standing behind the man", "man jumping over fire hydrant"). From the region descriptions, the following information is distilled: objects are extracted and merged across region descriptions (e.g., "yellow fire hydrant" is mapped to the hydrant in "man jumping over fire hydrant"), attributes are extracted from the objects (e.g., "yellow" is an attribute of the hydrant), and relationships between objects (e.g., "jumping over" in "man jumping over fire hydrant").
- **Objects**: For each of the extracted objects, a participant is asked to draw a bounding box. The bounding boxes are as tight as possible and should be within the bounding box of the region description.
- **Attributes and relationships**: For each of the extracted objects, participants are asked to create describe relationships as a triplet `(relationship, object, attribute|object)` (e.g., `(behind, man, woman)` or `(in, woman, shorts)`).

- **Question and Answers**: Participants now provide questions and answers (six Ws: who, what, where, why, when, how) for the image as a whole and for particular regions.
- **Verification**: Objects are validated by using rapid judgments [351] (i.e., participants are shown objects in a flashed manner and are asked to verify if the object is present in the image). All other annotations are validated using majority voting [352] asking three participants if the annotation is present (if two out of three participants agree, the annotation is accepted).

The detailed annotation process of Visual Genome allows one to create a validated scene graph for each image and a global taxonomy of objects, attributes, and relationships across all images.

6.1.3.1 Summary

GWAPs use the consensus of multiple participants to validate that newly given tags are likely to be adequate (they are used by participants). The gamification of the task (users receive points upon using the same word) allows to collect the data without financial compensation, however, the exploration of labels can be slow and inefficient since participants are incentivized to use words that are likely to be also used by other participants (and might be trivial, non-informative descriptions) potentially leading to an incomplete set of labels.

Cascade uses crowd workers to create taxonomies by letting participants write new tags, select the best tag from a list of possible tags for a stimulus, and mark which categories are present for a stimulus. While the process yields high-quality taxonomies only including laymen, it costs as much or more than taxonomies made by experts. One problem with this paradigm is that all tasks are done in isolation (creation, selection, and categorization) which makes the paradigm less efficient.

Visual Genome improves over this by letting participants write full captions for bounding boxes drawn in images and by extracting objects, attributes, and relationships by using NLP tools. They also use an efficient procedure for validating objects by using a flashed presentation. However, the approach is extensive (and thus laborious and costly) and it's tailored to the image modality (e.g., drawing bounding boxes would be more difficult for auditory experiences, especially if sounds overlap).

None of the three paradigms used the full potential of iterative annotation. In the next section, we will show how adding the dependence across trials can improve the efficiency of the annotation process.

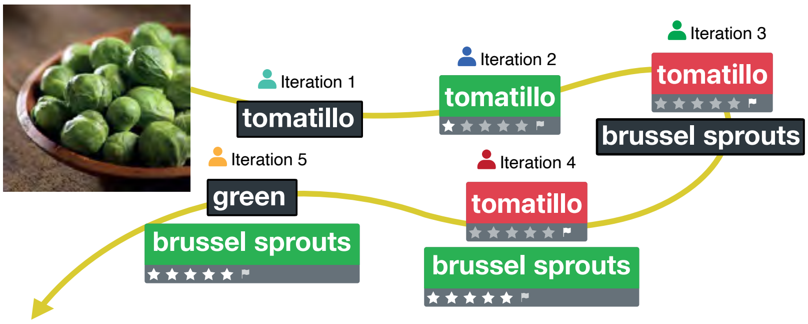


Figure 6.1: STEP Schematics An iterative HITL approach, in which participants can provide new tags, rate, and potentially flag tags of other participants.

6.2 Paradigm

Sequential Transmission Evaluation Pipeline (STEP) integrates ideas from iterated learning and serial reproduction [66, 70], by asking participants to iteratively provide new tags, and to rate or flag existing tags (see Figure 6.1).

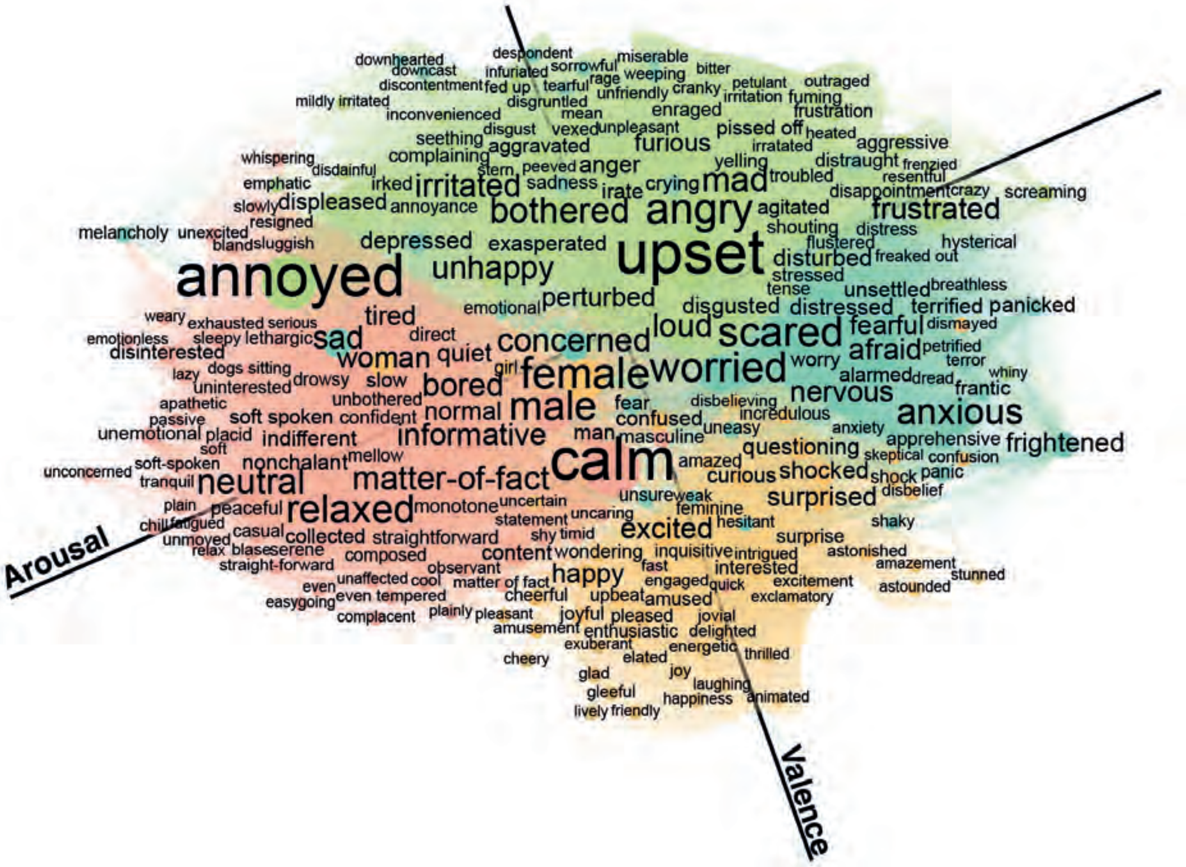


Figure 6.2: Co-occurrence graph of emotional prosody The color of the nodes is the modularity class. Modularity: 0.42. The size of the nodes is proportional to the degree of the node.

In contrast, to Visual Genome, STEP is modality-agnostic and can be applied to any stimulus type. In each trial, participants are presented with a stimulus (e.g., an image, audio clip, or video excerpt) and are asked to provide at least one tag. If the stimulus was presented before, participants are first asked to rate the relevance of the tags provided by previous participants (on a 5-interval Likert star-scale ranging from “not very relevant” to “very relevant”) or to flag a tag if they find it inappropriate. If a tag is flagged by two or more participants, it is removed from the list. However, participants are also given the opportunity to re-add the removed tag if they feel it is relevant. For example, in Figure 6.1, the image is first mistakenly labeled as “tomatillo”, which is then flagged twice and removed. Also, the correct label “brussel sprouts” is added by a participant in the third iteration and voted as highly relevant by the following participants. To increase overlap in the used labels across stimuli and participants, each response is stored in a global pool of tags. When participants start typing a new tag, we provide auto-completion suggestions based on the global pool of tags. Participants are incentivized to provide tags as single words, and unless it is a lexicalized concept (e.g., “brussel sprouts”, which has a different meaning from “brussel” + “sprouts”). Participants cannot provide the same tag twice for the same stimulus or provide tags that are already assigned to the stimulus.

We collected at least 10 iterations and the chain could have up to 20 iterations. Chains could end early if they converge if the last iteration has at least 2 tags that were rated at least 3 times and had a mean rating of 3 stars. After this process is finished, one yields a weighted bag of words describing each stimulus, i.e., a matrix T where T_{ij} is the mean rating of tag i for stimulus j.

One can also compute a co-occurrence matrix C where C_{ij} is the number of times tag i and tag j co-occur in the same stimulus. This matrix can be expressed as a graph where nodes are tags and edges are co-occurrences. In Figure 6.2, we show the graph of tags for tags provided for emotional prosody recordings taken from the RAVDESS corpus [44].

To detect semantic clusters in the graph, we used the Louvain community detection algorithm [353]. The resulting modularity score describes how well the network is compartmentalized into sub-networks. A modularity score of 1 indicates perfect compartmentalization, while a score of 0 indicates no compartmentalization. Each of the nodes is colored by the modularity class it belongs to. In Figure 6.2, one can see that similar tags are grouped together (e.g., “happiness” and “joy” or “anxious” and “frightened”), approximately following the circumplex model of emotion [164]. The position of the nodes is determined by the edges and their weights to other nodes (co-occurrence). While most nodes tend to be connected to the neighboring nodes in the graph (e.g., “male” and “man”), some nodes are not connected via an edge, e.g., “male” and “bored” but are close in the graph.

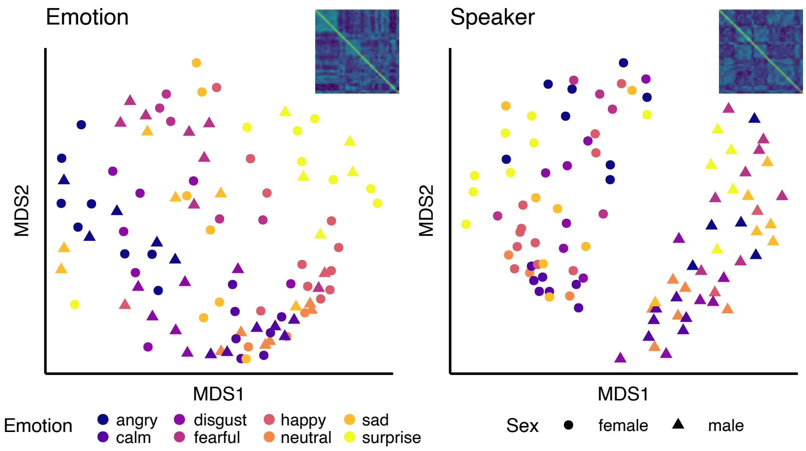


Figure 6.3: STEP respects MDS on the predicted pairwise similarity matrix based on tags. Participants listen to emotional speech recordings from the RAVDESS corpus and are either asked to provide tags describing the emotion (left) or the speaker (right).

Instead of asking participants to simply describe the stimulus, one can also ask participants to focus on a particular respect of the stimulus. In Figure 6.3, participants listened to the same emotional prosody recordings from the RAVDESS corpus [44] but were either asked to either describe the emotion or the speaker. While the MDS plot on the left shows that the tags are grouped by emotion, the plot on the right shows that the tags are grouped by the sex of the speaker, but the emotions are not as well separated.

6.3 Predicting pairwise similarity judgments

Pairwise similarity judgments are a common way to measure the similarity between stimuli, however, they are costly to collect since they require participants to compare each pair of stimuli. Here, we tried to predict pairwise similarity judgments from (i) SOTA machine learning models trained only on stimulus data and not on any text, (ii) human captions describing the stimuli, and (ii) tags collected with STEP. For the tag and caption data, we obtained embeddings either by using word or sentence embedding models (such as ConceptNet [354] or BERT [355]) or by using embedding-free methods (such as co-occurrence or cosine similarity on rouge score [356]). To then obtain the predicted similarity judgments, we computed the cosine similarity between pairs of embedding vectors to produce a similarity matrix. To benchmark the predictions against a subset of real similarity judgments, we used the Spearman correlation coefficient between the predicted and the actual similarity judgments.

6.3.1 Datasets

We collected tags for three modalities: images of objects (vegetables, furniture, and animals) [357], audio recordings of emotional prosody [44], and video clips of everyday activities [358].

6.3.2 Validating STEP

To benchmark STEP, we compared it against several baselines on the video dataset. In the first comparison, we randomly selected only a single high-rated tag from the last iteration per stimulus. We showed that using a single tag greatly decreases the correlation with human similarity (from $r = 0.74$ using all tags to $r = 0.35$ using a single tag).

In the second comparison, we compared tags from the first iteration of STEP (equivalent to non-adaptive tag collection) to tags from the last iteration. We showed that the correlation with human similarity greatly decreased when using tags from the first iteration (from $r = 0.74$ to $r = 0.44$).

Finally, we extracted the activity labels for each video given by the corpus and compared them to the tags from STEP. We showed that using labels decreased the correlation with human similarity (from $r = 0.74$ to $r = 0.64$).

All in all, we showed that: (i) tags produced after multiple iterations of STEP outperformed all three baselines and (ii) that the iterative nature of STEP leads to higher quality tags.

6.3.3 Results

In Figure 6.4, we show the mean correlation for each method type (DNN, captions, and tags) for each modality. We found that DNN are always outperformed by tags and captions (but not by a large margin), tags are usually better than captions (except for video), and across all modalities, combining textual embeddings (either caption or tags) with DNN embeddings (called “stacked”) leads to the best performance. This indicates that the tags and captions capture information that is not captured by the DNN embeddings but is relevant for predicting similarity judgments.

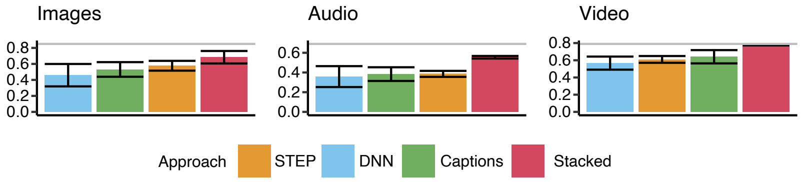


Figure 6.4: Correlation to human similarity Mean correlation for each method type for each modality. The error bars are standard deviations. The gray line indicates the inter-rater reliability, indicating the upper bound of the correlation.

6.4 Grounded semantic networks

Previous research has shown that many biological (such as the neural network of a worm [359]), technological (such as the US power grid [360]) and social networks (such as film actor networks [361] and collaborations between scientists [362]) exhibit scale-free and small-world properties. In scale-free networks, the degree distribution follows a power-law distribution, meaning that there are a few nodes with many connections and many nodes with few connections. As shown in Figure 6.5 (top) this is in contrast to random networks, where the degree distribution follows a normal distribution, meaning that most nodes have a similar number of connections. Small-world networks are characterized by sparse connectivity (i.e., few connections between nodes), a high clustering coefficient (i.e., nodes tend to cluster together), and a low average path length (i.e., the average number of steps it takes to get from one node to another) [360].

Steyvers and Tennenbaum [363] showed that semantic networks also exhibit small-world and scale-free properties, by analyzing three large text corpora [364–366]. Here we used STEP to collect tags for grounded experiences (e.g., seeing a physical image, hearing a sound, or watching a video).

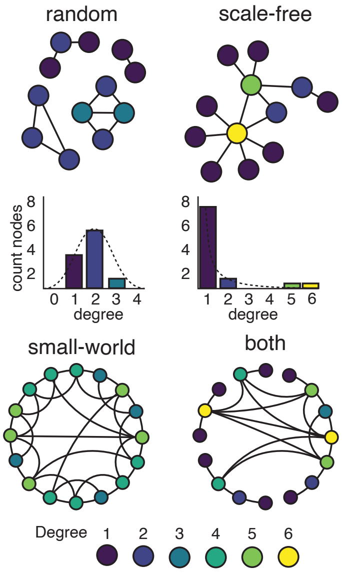


Figure 6.5: Types of networks Top: Example of a random and a scale-free network and their degree distributions. Bottom: Example of a small-world network and a network with both small-world and scale-free properties.

6.4.1 Datasets

- **RAVDESS** [44]: The dataset comprises a set of sentences spoken by 24 US American actors to convey a specific target emotion (“neutral”, “calm”, “happy”, “sad”, “angry”, “fearful”, “disgust”, and “surprised”). Out of the 1,440 recordings, we selected 1,000 by selecting three emotions per speaker per sentence, randomly omitting 104 emotional stimuli, and including all 96 neutral recordings. The co-occurrence graph is shown in Figure 6.2.
- **BOLD5K** [367]: The dataset consists of 4,916 images of indoor/outdoor activities as well as natural scenes. The images can be further grouped into i) 1,000 hand-curated indoor and outdoor scenes from 250 categories [368], ii) 2,000 images from the COCO dataset [337] depicting multiple objects (both inanimate and animate) and interactions between them (e.g., everyday human social interactions), and iii) 1,916 images from ImageNet of individual objects [336]. All images were tagged with STEP. The co-occurrence graph is shown in Figure 6.6.
- **Mini-Kinetics** [358]: The dataset consists of short video clips of everyday activities from 200 activity classes taken from the Mini-Kinetics-200 dataset. For each of the 200 activity classes, 5 random videos were sampled, totaling 1,000 video clips. The co-occurrence graph is shown in Figure 6.7.
- **WikiArt** [369]: The dataset consists of 4,105 artworks from WikiArt that received annotations for the emotions they evoke. We randomly sampled 1,000 artworks from the WikiArt Emotions dataset [369], by

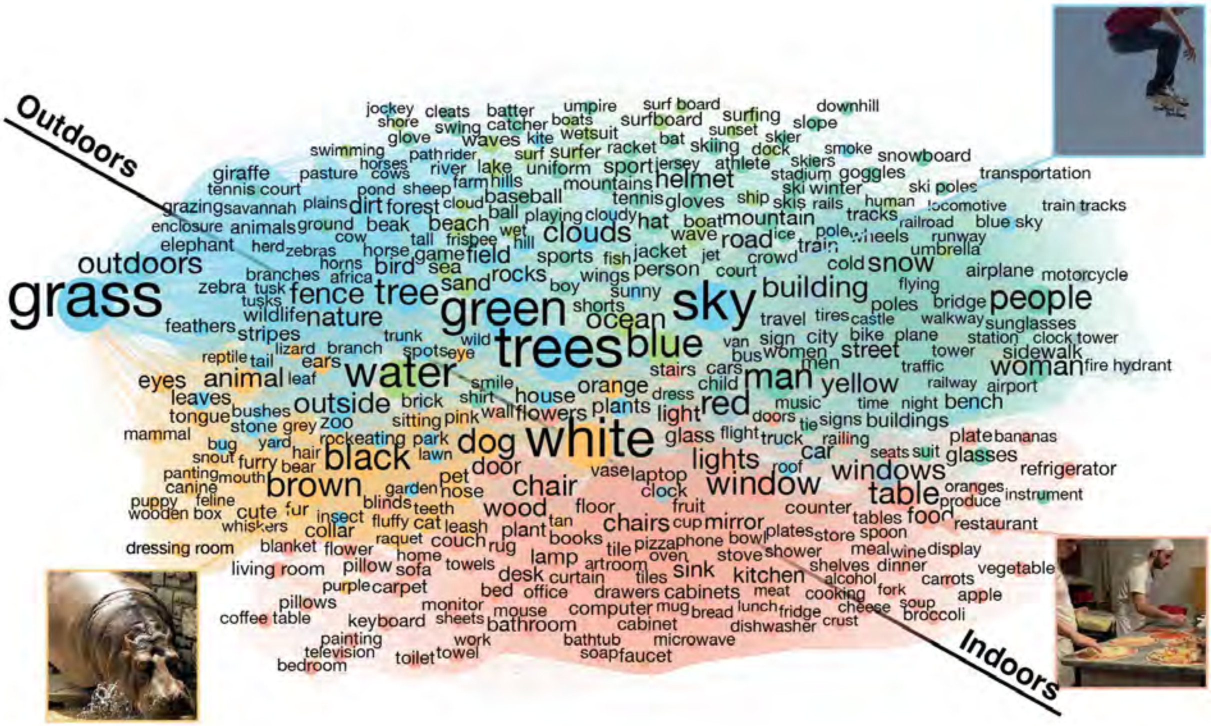


Figure 6.6: Co-occurrence graph of Bold5000 images Corpus of 4,916 images of activities and natural scenes. The color of the nodes is the modularity class. Modularity: 0.35. We added speculative axes (indoor-outdoor) based on our impression of the communities' content. Images from [367] (public domain license).

sampling 100 images from each of the following categories: "Impressionism", "Neo-Expressionism", "Post-Impressionism", "Cubism", "Abstract Expressionism", "Minimalism", "Color Field Painting", "Art Informel", "Abstract Art", and "Lyrical Abstraction". The co-occurrence graph is shown in Figure 6.8.

6.4.2 Participants

All participants were recruited through Amazon Mechanical Turk recruitment platform (MTurk), had to reside in the United States, be at least 18 years old, and have successfully completed at least 5,000 tasks on MTurk with an approval rate of 99%. Participants additionally had to pass an English proficiency test in order to participate [332]. Overall, N = 1,902 participants took part in the STEP paradigm.

6.4.3 Results

As shown in the co-occurrence graphs of each of the modalities (Figures 6.2, 6.6, 6.7, and 6.8), STEP reveals the rich semantics of each of the modalities. Where emotional prosody recordings roughly follow the circumplex model of emotion [164] (Figure 6.2), the BOLD5K images show a clear separation between indoor and outdoor activities (Figure 6.6), the Mini-Kinetics videos show a

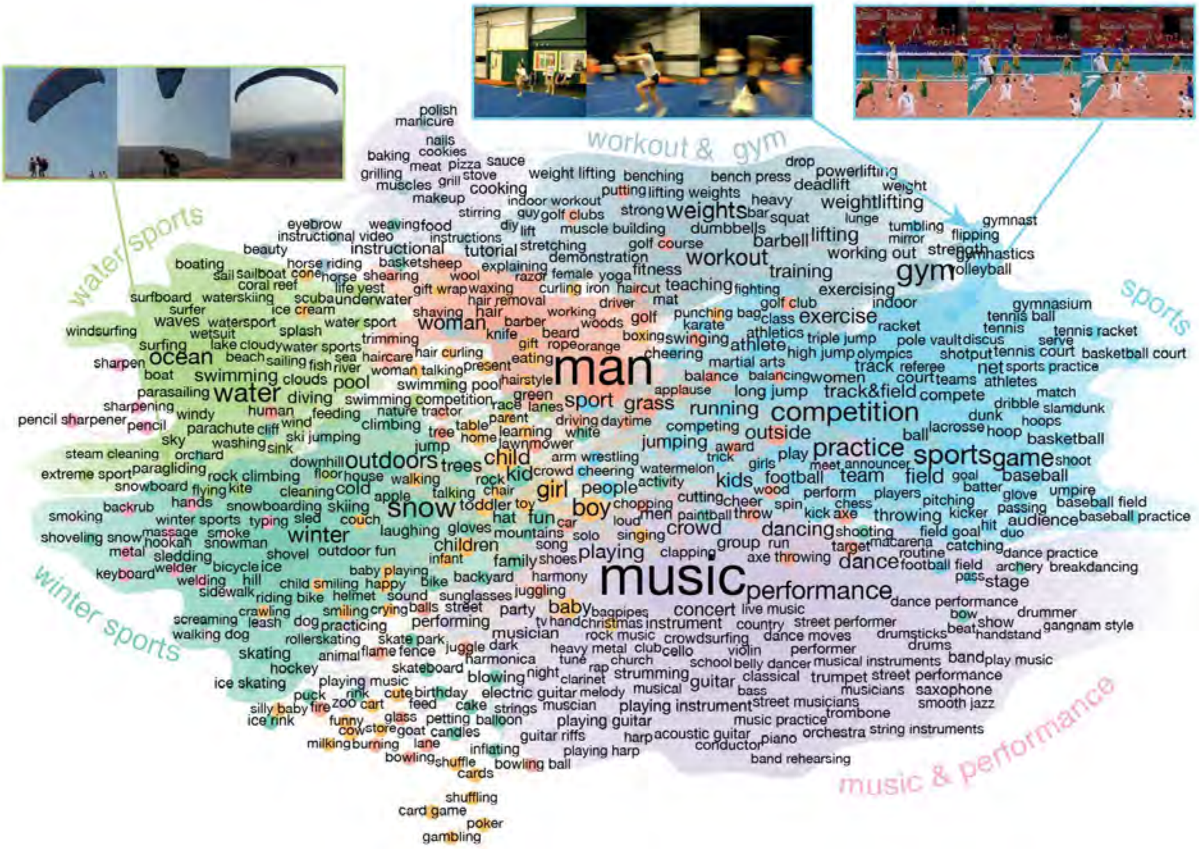

workout & gym
water sports
sports
winter sports
music & performance
polish
manicure
nails
baking cookies
grilling meat pizza sauce
muscles grill stove
makeup cooking
weight lifting benching bench press
drop powerlifting deadlift weight
putting lifting weights heavy weightlifting
indoor workout
stirring guy golf clubs strong weights bar squat lunge
eyebrow weaving food diy lift muscle building dumbbells barbell lifting
tumbling mirror flipping gymnast
instructional video instructions
beauty instructional tutorial stretching golf course
demonstration workout working out strength gymnastics
boating horse riding basketsheep explaining
sail sailboat cone horse shearing wool razor female yoga fitness training gym volleyball
coral reef life vest gift wrap waxing curling iron haircut teaching fighting
surfboard waterskiing scuba underwater hair removal driver mat exercising indoor
surfer ice cream golf club
windsurfing waves watersport splash water sport shaving hair working punching bag class exercise
woman barber woods golf karate gymnasium
wetsuit trimming knife beard boxing swinging athletics racket tennis ball
surfing lake cloudy water sports hair curling gift rope orange athlete triple jump pole vault discus tennis serve tennis racket
sharpen ocean beach sailing fish sea haircare eating cheering high jump olympics shotput tennis court basketball court
boat swimming clouds river woman talking present man martial arts track referee net sports practice
pool swimming pool hairstyle balance balancing women court teams athletes
parasailing water diving swimming competition race lanes green sport applause long jump track&field compete match
sharpening windy human feeding nature tractor grass running competition dribble slamdunk
pencil sharpener pencil parachute cliff wind climbing parent driving daytime dunk hoops
sky washing sink ski jumping tree table learning white competing outside lacrosse ball hoop basketball
steam cleaning orchard jump home lawnmower jumping award
extreme sport paragliding downhill outdoors trees child arm wrestling trick girls play practice sportsgame shoot
snowboard flying kite rock climbing floor house walking rock kid crowd cheering watermelon meet announcer team field goal baseball
backrub hands cleaning cold apple talking chair girl people activity kids football batter glove umpire
smoking snowboarding skiing snow toddler toy boy chopping cutting cheer wood perform players pitching baseball field
winter sports typing sled couch hat fun car loud men paintball throw spin chess kick axe throwing kicker passing audience baseball practice
shoveling snow massage smoke winter laughing gloves mountains solo singing crowd dancing shooting field goal hit duo
hookah snowman children song group run target macarena catching
metal sledding shovel outdoor fun infant family shoes playing clapping axe throwing dance routine dance practice
keyboard welder bicycle ice baby playing football field archery breakdancing
welding hill child smiling happy bike backyard harmony music pass stage
sidewalk riding bike helmet sound sunglasses juggling performance
crawling smiling crying balls street party baby bagpipes concert live music dance performance
screaming leash dog practicing performing tv hand christmas instrument country street performer bow drummer
walking dog rollerskating skate park musician rock music crowdsurfing dance moves drumsticks beat show handstand gangnam style
skating animal flame fence juggle dark heavy metal club cello violin performer drums
hockey skateboard harmonica tune church school belly dancer musical instruments band play music
ice skating playing music blowing night rap clarinet strumming guitar classical trumpet street performance
puck rink cute birthday electric guitar melody musical bass musicians saxophone
silly baby fire zoo cart feed cake strings musician playing instrument street musicians smooth jazz
ice rink funny glass petting balloon playing guitar music practice trombone
cow store goat candles guitar riffs harp acoustic guitar piano orchestra string instruments
milking burning lane inflating playing harp conductor
bowling shuffle cards bowling ball band rehearsing
shuffling
card game
poker
gambling

Figure 6.7: Co-occurrence graph of Mini-Kinetics videos Videos of everyday activities from the Mini-Kinetics-200 dataset [358] (CC-BY). Modularity: 0.55. We added speculative labels to the modularity classes based on our impression of the communities' content.

clear separation between activities that involve people and those that do not (Figure 6.7), and the WikiArt images show a clear separation between abstract and representational art (Figure 6.8).

When inspecting the graphs, one can see that a few tags are big (i.e., high degree) and the majority of tags are small (i.e., low degree), indicating that the graphs are scale-free. Also, we can see that the tags are clustered together, indicating that the graphs are small-world. To quantify those observations, we computed:

- **Average sparsity** $\bar{s}$: The average number of connections a node has in relation to the maximum number of connections a node can have (i.e. $\bar{d}/N$, where $\bar{d}$ is the average number of connections a node has and N is the number of nodes).
- **Average shorted path length** L: The average number of steps it takes to get from one node to another (shortest path).
- **Clustering coefficient** C: The average of the local clustering coefficients of all nodes in the graph. Intuitively, this is the fraction of a node's neighbors who are themselves neighbors.
- **Small-worldness** σ: The ratio of the clustering coefficient and the average shortest path length (i.e., $\sigma = (C/C_r)/(L/L_r)$). A graph is said to be small-world whenever $\sigma > 1$.

We found that all graphs were sparse (each node is only connected to < 3.5% of all other nodes), had a low average the shortest path length (it takes on average < 3 steps to get from one node to another), and had a high clustering coefficient

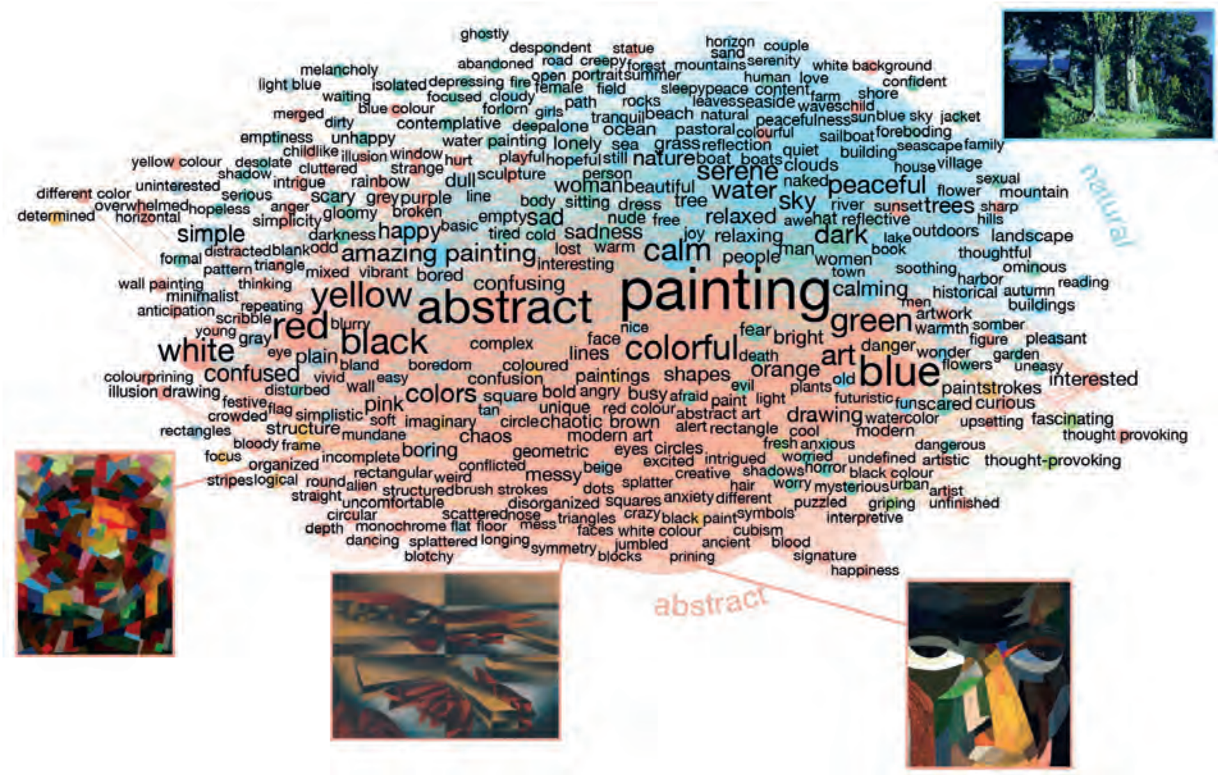

ghostly
despondent statue horizon couple sand
melancholy abandoned road creepy forest mountains serenity white background
light blue isolated depressing fire open portrait summer human love confident
waiting focused cloudy female field sleepy peace content farm shore
merged blue colour forlorn girls path rocks leaves seaside waves child
dirty contemplative deep alone tranquil beach natural peacefulness sun blue sky jacket
emptiness unhappy water painting lonely sea ocean pastoral colourful sailboat foreboding
childlike illusion window grass reflection quiet building seascape family
yellow colour desolate cluttered strange hurt playful hopeful still nature boat boats clouds house village
shadow intrigue rainbow dull sculpture person serene
different color uninterested serious scary grey purple line woman beautiful water naked peaceful flower sexual mountain
determined overwhelmed hopeless anger gloomy broken body sitting dress tree sky river sunset trees sharp
horizontal simplicity darkness happy basic empty sad nude free relaxed awe hat reflective hills
simple tired cold sadness joy relaxing dark lake outdoors landscape
formal distracted blank odd amazing painting lost warm calm people man women book thoughtful
pattern triangle mixed vibrant bored interesting town soothing harbor ominous
wall painting thinking confusing calming historical autumn reading
minimalist yellow abstract painting men artwork buildings
anticipation repeating scribble blurry nice fear warmth somber
young red black face bright green figure pleasant
white gray eye plain complex colorful death danger wonder garden
bland lines orange art flowers uneasy
colourpriming confused vivid easy boredom coloured old blue interested
illusion drawing disturbed wall confusion paintings shapes evil plants paintstrokes
festive colors square bold angry busy afraid paint light futuristic fun scared curious
crowded flag simplistic pink tan unique red colour abstract art drawing watercolor upsetting fascinating
rectangles structure soft imaginary circle chaotic brown alert rectangle modern thought provoking
bloody frame mundane chaos modern art cool
focus organized incomplete boring geometric eyes circles fresh anxious dangerous
conflicted excited intrigued worried undefined artistic thought-provoking
stripes logical rectangular weird messy beige creative shadows horror black colour
round alien structured brush strokes splatter hair worry mysterious urban artist
straight uncomfortable dots disorganized squares anxiety different puzzled griping unfinished
circular scattered nose crazy symbols interpretive
depth monochrome flat floor mess triangles black paint
faces white colour cubism
dancing splattered longing symmetry jumbled ancient blood
blotchy blocks prining signature
happiness
natural
abstract

Figure 6.8: Co-occurrence graph of WikiArt Sample of 1,000 artworks from WikiArt [369]. Sample artworks are reproduced from WikiArt under a public domain license. Modularity: 0.31. We added speculative labels to the modularity classes based on our impression of the communities' content.

(neighboring nodes are likely to be connected, $C > .7$). The small-worldness was > 20 for all graphs, indicating that the graphs are small-world (all $\sigma > 1$).

This shows that the grounded semantic networks obtained via STEP exhibit small-world and scale-free properties, similar to other biological, technological, and social networks.

6.5 Discussion

6.5.1 Summary

In this chapter,

- We reviewed three paradigms for HITL annotation pipelines: GWAP, Cascade, and Visual Genome that all use consensus across participants to obtain a validated set of tags describing stimuli (see Table 6.1).
- We introduced STEP, a novel adaptive tagging pipeline that involves the creation of new tags and rating of existing tags.
- STEP is modality-agnostic, in contrast to Visual Genome, is less costly than Cascade and Visual Genome, and improves over existing work by combining multiple steps in once (stimulus rating and creating in the same step), by using ratings instead of categorization, and by applying an iterative approach to improve the quality of the tags.
- We showed that STEP can be used to predict similarity judgments and that STEP outperforms class labels given by the corpus and tags from

the first iteration of STEP (indicating that the iterative nature of STEP leads to higher quality tags).

- We demonstrated that grounded semantic networks obtained via STEP exhibit small-world and scale-free properties, similar to other biological, technological, and social networks.

Problem	GWAP	Casc.	VisG	STEP
Consensus	✓	✓	✓	✓
Modality agnostic	✓	✓	✗	✓
Costly	✓	✗	✗	✓
Multiple steps at once	✗	✗	✓	✓
Soft-labels	✗	✗	✗	✓
Iterative	✗	✗	✗	✓

Table 6.1: STEP vs other HITL algorithms VG = Visual Genome, C = Cascade Checkmark indicates the paradigm addresses the problem, a cross indicates the paradigm does not address the problem, and a tilde indicates the problem is only solved partially.

The results show that STEP can be used to collect high-quality tags involving lay participants in a cost and time-efficient manner.

6.5.2 Contributions to Emotional Prosody

When STEP is applied to the domain of emotional prosody, it allows to solve the following identified problems (see Table 3.1):

- **Assumption of a single emotion**: unlike forced-choice emotion recognition experiments, STEP does not force participants to choose a single emotion, but allows them to provide multiple tags describing the full emotional experience.
- **Assumption of existing emotion taxonomy**: STEP does not present the participant with a predefined list of emotions, but allows them to provide their own emotion tags they think are relevant.
- **Unclear alignment of subtypes**: the co-occurrence of tags across emotional recordings allows to identify subtypes of emotions, that are likely to co-occur (e.g., “happiness” and “joy”).
- **Lost-in-translation**: When STEP is applied to different languages and a subset of stimuli is presented to all participants, this allows to align the semantic spaces across languages and allows to measure the alignment of emotional concepts across languages (e.g., if the tag “traurig” in German always co-occurs with “sad” in English they are likely to refer to the same concept).

6.5.3 Limitations and Outlook

Future research can improve the following aspects of STEP:

- **Quality of tags**: The quality of the tags can be further improved. Enhanced checks can be put in place to ensure that the tags are valid words (such as spell checking or lemmatization). Another problem is that not all participants followed the instructions closely. In the emotion dataset, when asking participants about the emotion (Figure 6.3), some participants provided tags related to the speaker (e.g., “male” or “female”). Ways to reduce these problems could be to incentivize flagging of such tags more (e.g., a financial bonus) or to filter such tags posthoc (e.g., tags related to speakers tend to cluster together and could be assigned a different modality class).

- **Improved benchmarks**: Here we only compared STEP to the video modality. Future, work should extend this to different modalities and compare STEP to other HITL annotation pipelines, potentially also measuring differences in efficiency and cost across paradigms.
- **Improved efficiency**: Over the course of iterations, the number of tags accumulates. The initial tags, which tend to describe the most salient aspects of the stimulus, are rated more relatively to tags that are added later. One idea to overcome this would be to implement freezing tags after they received a certain number of ratings. Another idea would be to implement early stopping. In the current data collection, the chain can stop after 10 iterations, if the last iteration has at least 2 tags that were rated at least 3 times and had a mean rating of 3 stars. Those values were based on initial pilot data and future research can investigate more principled ways to stop the chain early.

© Elyse Butler

Part II
Infrastructure for Multilingual Research

It is imperative to study emotional prosody across languages. This is important for machine learning applications trained on multilingual data that are used across the globe and for studying how the expression of emotion in prosody varies across cultures and languages. In the first part of this thesis, we identified three problems of existing research. The second part of the thesis is mainly about the last problem – the lost-in-translation problem – about how to align emotional terms across languages. In this part, I developed an infrastructure to run massive multilingual online experiments across the globe and benchmark it on a well-established instance of grounded semantics study of the influence of language on perception. In Chapter 7, I describe my contributions to Psynet, a Python package I co-developed to conduct large-scale, complex online experiments in many languages across the globe. Running multilingual data collection requires verifying that participants are indeed native speakers of different languages. In Chapter 8, I solve this practical problem by developing an automated pipeline to create vocabulary tests that are used to assess online participants' linguistic backgrounds efficiently. In Chapter 9, we benchmark the infrastructure, by running it on a well-studied domain of color naming.

Chapter 7

Massive Online, Cross-Cultural Experiments

Studying emotional prosody across languages is essential because emotional expression varies significantly across linguistic and cultural contexts [11, 189, 370]. This variability often leads to uncertainty about whether emotion concepts align across languages, which we refer to as the "lost-in-translation" problem. Addressing this issue has practical implications, particularly for developing emotion recognition systems used globally [153, 154, 160]. It also has scientific importance, as the language a person speaks can influence nonlinguistic mental representations and cognitive processes [371, 372].

In particular, the field is hindered by the WEIRD recruiting bias [246], which limits research to an unrepresentative subset of the global population [247]. Moreover, most experiments are conducted in English [259], and emotion is typically studied using English terms (see Background Section 2.3.2), which means emotion research is biased towards English-speaking populations. For example, Prolific, the most widely used recruitment platform for online experiments, predominantly includes participants from Europe and North America, with an interface available only in English.

The solution is to conduct large-scale, cross-cultural online experiments, leveraging diverse recruitment platforms to reach participants from various countries who can engage with studies in their native languages. As of 2024, 66.2% of the global population has internet access, and 69.4% owns a smartphone, equating to approximately 5.35 billion devices [269]. This widespread connectivity indicates that a majority of the global population is accessible online. While this online cohort is more exposed to intercultural influences than remote societies (e.g., the Pirahã in the Amazon, the Himba in Namibia, or the Tsimane in Bolivia), it still exhibits significant linguistic and cultural diversity [260, 261]. Notably, even remote communities are becoming more connected through satellite phones, radios, and smartphones.

However, online experiments offer several advantages over traditional fieldwork. They enable the collection of large datasets at relatively low costs, requiring payment only for server infrastructure and participant compensation, which is often lower outside the Global North. Furthermore, online research allows for high levels of automation, streamlining recruitment, payment, and translation processes, which is not possible in fieldwork.

To address these challenges, I developed a research infrastructure for conducting massive online experiments on a global scale. In this chapter, we will review existing frameworks for running online experiments, introduce PsyNet—a Python package designed for advanced behavioral experimentation—and detail my contributions to the package, with a focus on enabling large-scale, cross-cultural research.

7.1 Background

Over the past decade, psychology, sociology, and economics have increasingly incorporated online participant pools into research, offering significant advantages. These pools enable experimenters to expand the size and diversity of participant groups while facilitating studies that would be nearly impossible in traditional lab settings, such as examining interactions among thousands of participants within social networks. Such studies include virtual worlds where participants engage in simulations of complex social systems [373–375], investigations into cultural transmission [376, 377], governance decisions within online communities [378, 379], large-scale perception research [380], and innovative methods combining humans with machine-learning algorithms to explore high-dimensional perceptual representations in the mind (Background Section 2.2, Chapter 4– 6). Such experiments can provide unique opportunities to study pressing issues of our time, such as the spread of misinformation [381] and the political instability within social networks [382].

Over the years, a number of frameworks have been proposed to run online experiments [383–386]. The frameworks differ in their focus, with some focusing on providing a graphical user interface (Gorilla [386], PsychoPy [383], or OpenSesame [387, 388]), others focus on complete front-end processing to enable maximal scalability since only static hosting is required (Pushkin [389] and jsPsych [384, 390]), and other focus on providing a server which can run multiple experiments in parallel (JATOS [385] or World-Wide-Lab [391]). However, all of those frameworks support limited complexity in terms of experiment design and processing capabilities.

One recent framework that addresses some of these challenges is Empirica [392]. It supports the development of multiplayer experiments by coordinating participants into rounds to complete tasks, such as solving a murder mystery. It includes utilities for real-time communication between participants and the server and provides a dashboard for monitoring participant progress. However, experimenters must largely implement the game logic themselves, often requiring advanced web development skills. In one case study, researchers hired a software developer to create a polished experiment, highlighting the complexity of this process.

Another framework, Dallinger [393], specializes in experiments structured around networks, such as those involving social networks or cultural evolution. Its key feature is a standardized database structure that maps onto Python objects using PostgreSQL, facilitating network-based experiment designs. Dallinger also simplifies web server provisioning, experiment deployment, and participant recruitment via services like MTurk and Prolific. Under the hood, it builds on the Flask web framework and Gunicorn WSGI server giving Dallinger powerful back-end processing capabilities, such as generating complex stimuli using generative AI and processing participant responses (e.g., audio recordings, and transcriptions). However, it provides limited support for organizing event timelines within participant sessions and leaves front-end interface design to the experimenter, making new experiment implementation time-intensive.

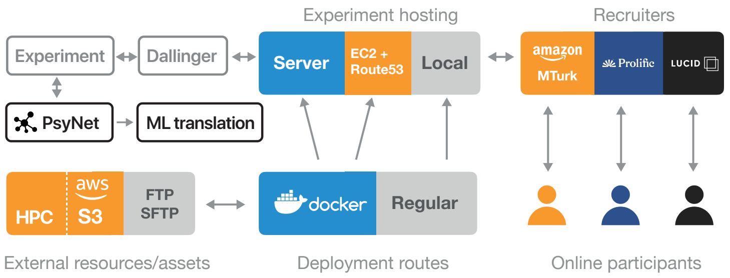


Figure 7.1: PsyNet Workflow PsyNet is a wrapper around Dallinger. Dallinger can be run by installing the dependencies locally (regular) or can be containerized using Docker. Experiments can be deployed to a server by pushing the experiment to Heroku or via Docker. The deployed experiment requests participants through a recruiter. Currently, PsyNet supports MTurk, Prolific, and Lucid as recruiters.

7.2 Towards a virtual lab with PsyNet

To overcome these problems in Dallinger, PsyNet was developed as a wrapper around it (see Figure 7.1). The package was initialized by Peter Harrison and Nori Jacoby in 2019 and I have been a core developer of PsyNet since the beginning.

Dallinger largely misses a vocabulary of standardized concepts ranging from trials, to stimuli (e.g., audio, video), to paradigms (e.g., static experiment, iterated reproduction), which is introduced in PsyNet.

In Dallinger, a `Participant` creates an `Info`, which is part of a `Node` in a `Network`. An edge between two nodes is called a `Vector`. Vectors control the flow of the network. In most experiments, the structure of the network is a chain, where the infos of a given node is summarized in the next node, which then serves as the input for the next infos (see Figure 7.2).

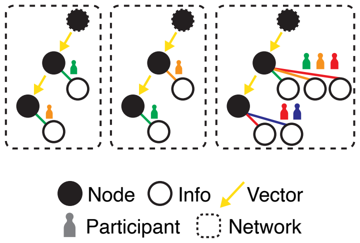


Figure 7.2: Dallinger classes Example of three networks, where individual participants are assigned to Infos that are part of a node which is part of a network.

PsyNet improves this by introducing the following concepts (the relationship of the most important classes is shown in Figure 7.3):

- **`Timeline`**: A timeline is a sequence of events that the participant will experience. It supports all basic operations such as conditional logic, branching, and looping.
- **`ModularPage`**: To render a page in Dallinger, one has to pass variables to the Jinja template, which then renders an HTML page. In PsyNet, a `ModularPage` accepts both a `Prompt` and a `Control`. A `Prompt` is a display (e.g., video, audio, image), and a `Control` is a response (e.g., text input, a slider, microphone recordings). This separation allows us to easily change the modality of a particular experiment by only making minor changes to the experiments.
- **`Asset`**: PsyNet introduces the notion of an `Asset`. Assets are used to store files (e.g., images, audio, video) that can be used in a `ModularPage`. When exporting the experiment, the assets are automatically downloaded from the server or stored on a cloud storage (e.g., S3). An asset is a file that can be used in a `ModularPage`.
- **`Trial`**: A `Trial` is a subclass of `Info`. The word “trial” aligns more with the vocabulary used in psychology. Using the `show_trial()` method a `ModularPage` (or a sequence of ModularPages) is shown to the `Participant`.
- **`TrialMaker`**: In Dallinger, the creation of a `Network` is done manually in the `Experiment` class, as well as adding a `Node` to the `Network`, which makes it hard to reuse the same structure in different experiments and cumbersome to change the experiment logic. To solve this, PsyNet introduces the concept of a `TrialMaker`. A trial maker

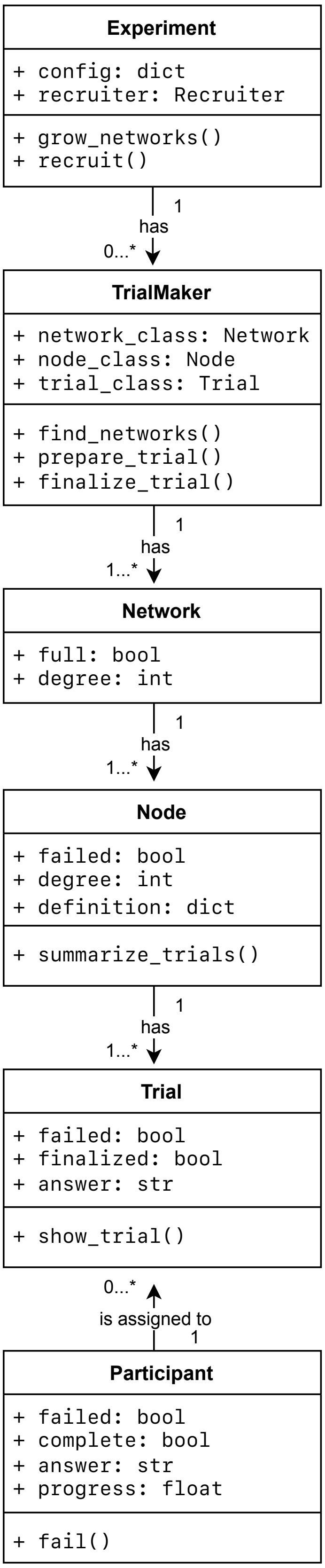


Figure 7.3: Truncated PsyNet UML UML diagram of most important classes in PsyNet.

is responsible for assigning a network (and thus a node) to a participant (`find_networks()`), creating the trial, and assigning it to a node and a participant (`prepare_trial()`), and finalizing a trial (`finalize_trial()`), e.g., asynchronously wait for the post-processing of the trial to be done, such as analyzing a voice recording. If this process succeeds the trial is marked as `finalized` and if it fails the trial is marked as `failed`. Participants can be excluded from the experiment if they fail a certain number of trials (which will also fail the participant). Once a `Node` reaches the targeted number of finalized trials, it will call `summarize_trials()` which will prepare the next node based on the trials in the current node. The information of the next node in `definition` and the `degree` indicates the iteration of the node in the network. What makes trial makers powerful is that they can implement a paradigm (e.g., GSP, Create and Rate, or STEP), but the trial, node, and network class can be overridden such that the presentation can be changed but the logic can be reused. This allows experimenters to easily change experiments, e.g., change an image-based experiment to an audio-based experiment by only changing the stimulus display type.

- **`Experiment`**: The experiment class is the main class in Dallinger and PsyNet. It will periodically check if a network can grow (i.e., the last node of the network has reached the targeted number of finalized trials and the network is not already `full`). If a participant reaches the end of an experiment (`progress` is 1), the participant is marked as `complete` and is paid the amount specified in the experiment through the recruiter. If the participant `failed`, participants receive a partial payment if this is supported by the recruiter. Depending on the experiment settings (`config`), a new participant is recruited automatically if not all networks are `full`.
- **Prescreening tasks**: Prescreening tasks are essentially little trial makers that make sure that the participant meets the criteria of the experiment. For example, this can be an automated check to make sure people are wearing headphones [312], a language proficiency task [10], or a check of the microphone quality. PsyNet provides a large library of prescreening tasks that can be easily added to the experiment.
- **Questionnaires**: Quite often, experiments require participants to fill in a questionnaire. This can be demographic information, a standardized questionnaire (e.g., on musicality or personality such as the Big Five), or some standard questions after the experiments (e.g., feedback or technical issues). Psynet provides a library of questionnaires that can be easily added to the experiment.
- **Monitoring**: PsyNet provides various dashboards that allow to monitor the progress of the experiment (where the participants are within the timeline), the progress of the chains (by plotting the chains as a graph), and the status of the experiment (e.g. CPU usage, memory usage, number of participants, recruiter metrics). This makes it easy to see if the experiment is running smoothly and if there are any issues that need to be addressed.

7.3 Contributions

Over the years, I have made various contributions to PsyNet and Dallinger including the implementation of an audio JavaScript API, slider interfaces for

GSP, improvements to the dashboard for data monitoring, automated error reporting during data export, improved browser detection, and the implementation of the STEP and "Create and Rate" paradigm. In the following sections, I will focus on three main contributions: the internationalization API, the provisioning API, and the Lucid integration.

7.3.1 Internationalization API

In order to run an experiment in many languages, the experiment has to be translated. To facilitate this, we used the `gettext` utilities to extract translatable strings from the experiment and to translate them using the Google Translate or DeepL API. When marking a string as translatable, the string can either be translated in isolation or together with all text in a particular 'context', for example, all text on an instruction page would be translated together. Strings are marked as translatable by wrapping them in a function call (e.g., `_('Hello World')`). We now use `gettext` to extract all translations from all Python files in the experiment and the `pybabel` package to extract translatable strings from other files (e.g., HTML, JavaScript). The extracted strings are then translated using the DeepL API. If the language is not supported by DeepL, the Google Translate API is used. The translations are then checked for consistency, for example within the same context the same translation cannot be used twice and each string needs to have a non-empty translation. Variable replacement in fstrings can only be done after the translation has been looked up (e.g., `_('Hello {name}').format(name='James')`), this means that the variable placeholders have to be present in the translation. This can be challenging because translators sometimes also translate the variable placeholders or HTML tags (like `<strong>`). To check for these issues, we developed a consistency checker that checks for these issues and reports them to the experimenter. Finally, variable insertions are tested to avoid runtime errors. The translations are then stored as a `.po` file and are compiled during runtime. Each `participant` has a `locale` attribute that is set to the language of the participant and attaches the translator to the participant. PsyNet is now available in 40 languages: Arabic, Belarusian, Bulgarian, Chinese, Croatian, Czech, Danish, Dutch, Estonian, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Lithuanian, Malay, Norwegian Bokmål and Nynorsk, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swahili, Swedish, Tagalog, Thai, Turkish, Ukrainian, Urdu, Vietnamese.

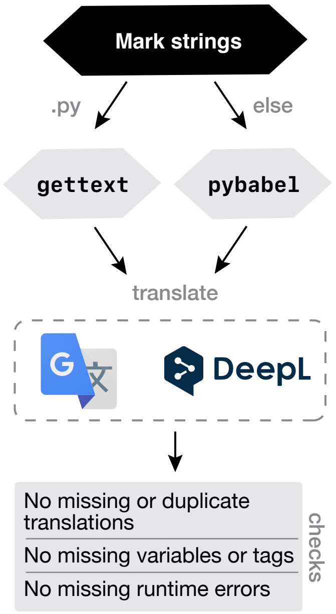

Mark strings
.py
else
gettext
pybabel
translate
DeepL
No missing or duplicate translations
No missing variables or tags
No missing runtime errors
checks

Figure 7.4: Internationalization API Four steps of the internationalization API: mark translatable strings, extract translations, machine translation, and automated checks.

7.3.2 Lucid Integration

The majority of online studies are currently conducted on Prolific. However, Prolific has a limited reach in terms of countries and languages (see Figure 7.5), the interface is only available in English, and captures a biased sample of demographics (mainly younger, educated, and highly fluent speakers of English).

To overcome this recruiter bias, we contributed to the integration of Lucid Marketplace recruitment platform (Lucid) in PsyNet. The marketplace, owned by Cint, connects businesses with a diverse network of respondents for survey-based research. Unlike Prolific, Lucid is a marketplace in which suppliers (e.g., researchers) propose a survey or experiment for a certain compensation. Bidders, which are usually market research companies, then bid on the survey or experiment and contract the participants. This means as a researcher you

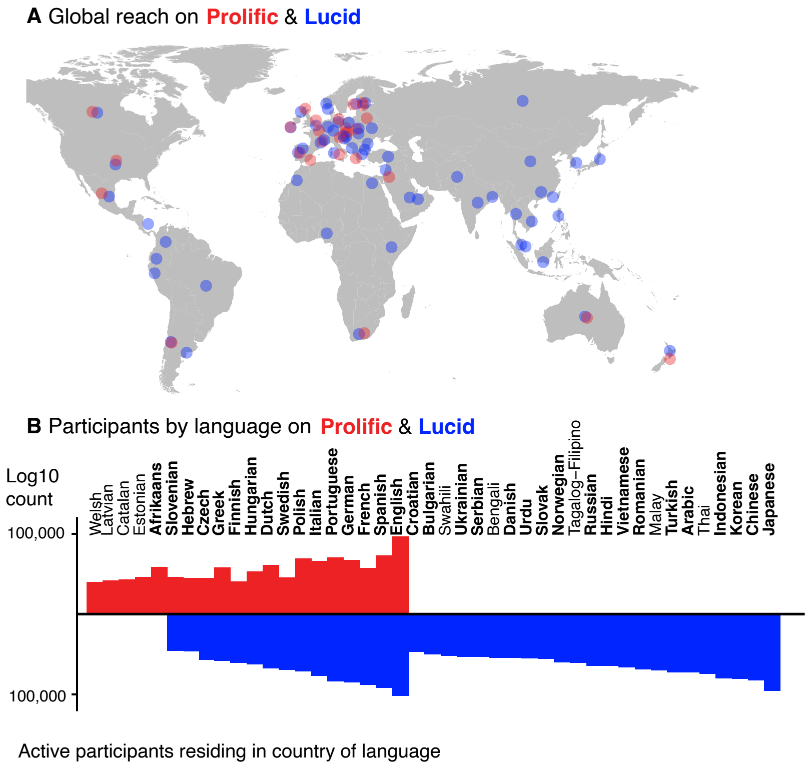


Figure 7.5: Lucid vs. Prolific **A** Recruitable countries in Prolific and Lucid. **B** Number of participants per language on Prolific and Lucid. For a country or language to be listed, it must have at least 100 active participants.

only interact with the participant during the experiment (e.g., there is no direct messaging before or after the experiment possible).

This made it quite challenging to integrate Lucid into PsyNet because participants would not arrive at the experiment or progress through the experiment as expected, which makes it hard to debug why this happens. To overcome this, we developed a monitoring system that polls the Lucid API to backmatch registered submissions on the Lucid platform and in the experiment, load and display all recruiter metrics (e.g., what is the median duration of the experiment for completed and failed participants) and log all interactions between the recruiter and the experiment.

7.3.3 Provisioning API

In order to run experiments across the globe, it is important to have a server that is close to the participants to reduce latency. AWS provides on-demand servers using EC2 in a large number of regions. We have developed a command line API to automatize setting up a new server in a region of choice (provisioning), pause and resume the server (to cut costs when the data collection is paused, e.g. during multi-day recruitments where the experiment is paused overnight), and to destroy the server when the data collection is done (teardown).

To support custom subdomains, we have implemented support for the Route 56 service, which allows users to set up a custom domain and wildcard SSL certificate for the domain.

7.4 Discussion

7.4.1 Summary

In this chapter,

- We have introduced Dallinger and PsyNet, two Python packages I actively contributed to, that allow me to run advanced behavioral experiments online.
- We have described the internationalization API that allows one to quickly translate an experiment into many languages.
- We have described the Lucid integration that allows to recruit participants from a more diverse set of countries and languages than Prolic.
- We have described the provisioning API that allows to automatically set up a server in a region of choice to reduce latency for the participants.

7.4.2 Outlook

- **Internationalization API**: I am currently working on an integration of ChatGPT translation to cover a larger number of languages and potentially enhance the translation quality, measure translation quality (e.g., by using BLEU scores), use an LLM to mark translatable strings in the experiment and develop a command line client to easily perform translations (e.g. `psynet translate fr de`).
- **Lucid integration**: I am in direct contact with the Lucid API team to improve the integration, e.g., find a way to enable variable or bonus payment and improve quality metrics to identify low-quality participants.
- **Provisioning API**: I am developing a pipeline to easily provide a certain service (e.g., a TTS or a transcription model) as a Flask app and to dockerize and deploy it as a RESTful API on a provisioned server.

Chapter 8

Language proficiency test for 1,939 languages

Based on
Pol van Rijn, Yue Sun, Harin Lee, Raja Marjieh, Ilia Sucholutsky, Francesca Lanzarini, Elisabeth André, and Nori Jacoby. 2023. 'Around the World in 60 Words: A Generative Vocabulary Test for Online Research'. *CogSci*.

When running multilingual experiments, it is difficult to know the language of the participants. This is different from lab experiments where the experimenter meets the participants in person or a company running a survey on a known pool of respondents. With PsyNet you can deploy online experiments anywhere in the world within minutes, but as an experimenter, you have little control over the language of the participants. For example, participants might report they speak a certain language to participate in a study or use automatic translation to participate in a study in a different language.

So with increased global access to the internet and increased possibilities to recruit online participants from all over the world [12, 75, 394–397], it is important to have a quick and reliable way to assess the language proficiency of participants. It is necessary to objectively verify participants' linguistic background beyond self-report in online studies [398], because (i) online participants may have diverse multilingual backgrounds [399], (ii) may provide noisy responses [400, 401], and can be less motivated or honest than lab participants [402].

To quickly screen the linguistic background of a participants, one needs a language proficiency test which:

- is fast to administer,
- is available in many languages,
- can be administered online,
- can be created automatically (to support repeated participation), and
- is reliable and valid.

In this chapter, we present an automated pipeline to create language proficiency tests for any language with sufficient text data. I use the pipeline to create vocabulary tests for 1,939 languages and validate them in three large-scale online experiments.

Connected more broadly, the language test allows studying more diverse populations and reducing the WEIRD bias in psychology [246] (see Background 2.3.2.1).

8.1 Background

Language proficiency is commonly measured using vocabulary tests. There are different types of tests, each developed for specific purposes. For example in DIALANG [403] was developed to assess language proficiency in 14 European languages. Vocabulary size is assessed by asking participants to fill the gap in a sentence, both measuring receptive (fill the gap using forced choice) and productive vocabulary (type a word). The Peabody Picture Vocabulary Test [404] on the other hand was developed to orally measure vocabulary size by presenting participants with four pictures and asking them to select the one that matches the word spoken by the experimenter.

8.1.1 LexTALE

LexTALE [332] has been developed to quickly measure the receptive vocabulary knowledge of a participant before taking part in a psycholinguistics experiment, being substantially more time-efficient than other language proficiency tests [405]. In this test, participants identify the real words from a list containing also fake words (pseudo-words). A test is designed by manually selecting rare real words and creating fake words with matching structural similarities to the selected real words. Native speakers find this test easy, while non-native speakers find it difficult. While the test was initially developed for English, it has been extended to 13 other languages [406–417].

However, LexTALE is limited in its generalizability. Manually creating the word list requires human domain experts and, therefore, introduces subjective biases. This reliance on human labor also limits the number of words in the list, which restricts the possibility of repeated testing. Furthermore, LexTALE requires a word frequency database that is not available for low-resource languages. Indeed, LexTALE is currently available in 14 languages, whereas in comparison, online recruiting platforms can provide access to speakers of over 90 languages [418, 419].

8.2 Pipeline

To overcome these limitations, I developed an automated pipeline to create language proficiency tests for any language with sufficient text data, that is comparable to LexTALE and allows to distinguish native speakers from speakers of closely related languages. The proposed pipeline consists of five steps (Figure 8.1): (1) collect the text, (2) clean the text, (3) create pseudo-words, (4) select real words, and (5) create pairs.

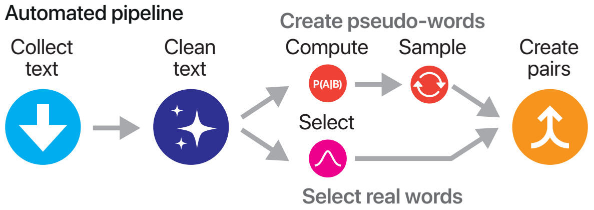


Figure 8.1: Pipeline Summary of the automated pipeline consisting of the following steps: Collecting and cleaning the text, creating pseudo-words, selecting rare words, and matching the real and pseudo-words.

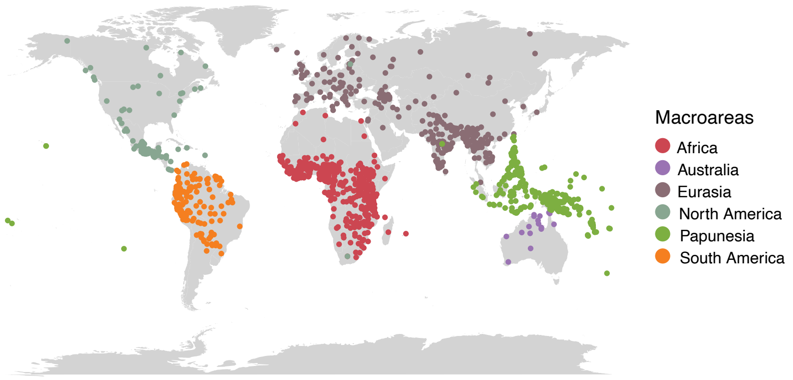


Figure 8.2: All 1,939 BibleVocab languages mapped onto the globe Each dot represents a language for which we created a vocabulary test. Dots are colored according to the UN macro areas. The location of the dots originates from glottolog [424].

8.2.1 Collect the text

I collected Wikipedia articles for 60 languages (WikiVocab) and Bible translations in 1,939 languages (BibleVocab).

8.2.1.1 Wikipedia

Wikipedia has been used for various research projects ranging from a database of notable people [420] to studying governance in online communities [421]. Due to its size, it is a suitable text source for estimating vocabulary frequency in multiple languages. For each language, we downloaded all Wikipedia articles using `wiki40b` [422] and `wikipedia` [423], which allows to directly download the Wikipedia dump for a language. Both packages implement basic pre-processing, e.g., removing non-content sections. To obtain a large enough vocabulary we ranked all Wikipedia articles in a language by the article length and processed the longest 10,000 articles.[1] For three languages (Gothic, Northern Sami, and Wolof), there were less than 10,000 articles. Since article length differs per language, we include the first 100,000 longest articles if they had less than 5 million valid words to reliably estimate the word frequency of infrequent words occurring down to only once per million.

8.2.1.2 Bible

The Bible is one of the world's most translated texts, making it a good textual source for creating vocabulary tests despite its relatively small size. It has shown to be a valuable resource for low-resource languages by training Natural Language Processing (NLP) models on the Bible, including machine translation [425], Part-Of-Speech (POS) tagging [426], and multilingual TTS and ASR using Bible recordings [427]. Here, we downloaded the Bibles from: https://www.Bible.com/. This website includes Bible translations for 2,068 languages. However, some languages only consist of a small portion of the Bible (e.g., only a few chapters) and thus are limited in the number of words. We removed languages that produced a final vocabulary list of less than 30 items, resulting in 1,939 languages, densely covering all UN macroareas of the world (Figure 8.2).

[1] We have originally considered various ways of sampling Wikipedia articles. Initially, we considered sampling the articles on the same topic to equate them across the vocabulary tests. However, there is only a limited number of articles on the same topic that exist for all included languages. This constraint would lead to too few articles per language. Another concern is that the length per concept strongly differs across languages (e.g., Finnish Wikipedia might have a long article about Finland, whereas the articles in other languages are shorter).

8.2.2 Clean the text

8.2.2.1 General text-cleaning

The following processing steps are general and are applied both for Wikipedia and the Bible:

General pre-processing. To improve the quality of the generated words, we implemented the following checks: (i) we rejected tokens that are written in exclusively capital letters since they are potential acronyms or anomalies in the text, (ii) we excluded one-letter words in the Latin alphabet as they tend to be used as indexes (e.g., "may *n* be the number of participants"), and (iii) we removed words containing letters or punctuation. All words are then changed to lowercase.[2]

Remove words with foreign characters To avoid typos, foreign words, or proper nouns, we removed words that contain characters that are not part of the writing system (e.g., a Chinese character in an English word). The removal is done using Unicode, a standardized text encoding to support most of the world's writing systems. Characters in Unicode are organized into blocks. For example, there is a block for Cyrillic, Arabic, or Hebrew characters (see Figure 8.3). We obtained a histogram of Unicode blocks for all characters in each of the accepted tokens. Based on this distribution, we removed words that are not part of the writing system. For WikiVocab, we manually identified the Unicode blocks (based on the expected language). We kept including the largest Unicode blocks for the Bible texts until the cumulative percentage exceeds 50 %.

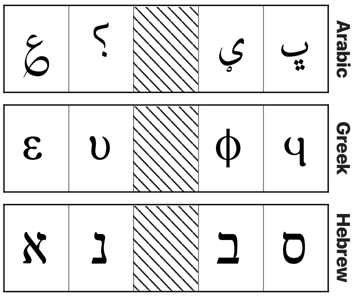


Figure 8.3: Unicode blocks Writing systems occupy distinct Unicode blocks. For example, the Latin script is in the block 0000-007F. Words that contain characters from unfrequently used blocks are likely to be foreign words.

Detect compound words Compound words are made by combining multiple existing words in a language. Languages differ in their usage of compound words. Certain languages, such as Dutch and German, allow to spontaneously create new words by combining two existing words. This is problematic in the context of the test for two reasons: First, compound words tend to occur much less frequently than each of its components, however, the infrequent compound word is not more difficult than each of the components. This means that compound words are likely to be selected as difficult words, whereas they are not. Second, without removing compound words, the created pseudo-words are more likely to be compound words themselves. Since the text corpus will not contain all possible compound word combinations, the generated word is likely to be marked as a fake word where actually it is an uncommon but real compound word.[3]

Autobahnraststätte

Figure 8.4: Compound words In some languages – especially in West Germanic languages like German and Dutch – indefinitely long compound words can be created by concatenating words. Quite often the compound words fall into subwords, which again can be compound words. We use a model to detect likely word boundaries. For example, if one were to cut the German word "Autobahnraststätte", it would most likely be split into "Autobahn" and "Raststätte", but both subwords are also compound words.

To remove compound words, we trained `charsplit` [428] for each language, which is a model to detect likely word boundaries. The model is trained on all cleaned words. We considered a word to be a compound word if the boundary is likely (> 0, threshold proposed by the author) and if the last segment is a valid word (part of the word list) in that language. For example in German, the model would split the word "Autobahnraststätte" into "Autobahn" and "Raststätte" (see Figure 8.4), because this is the most likely split and "Raststätte" is a valid word in German. We did not use compound word detection for Chinese, Japanese, and Korean (CJK), as each word is a chain of different characters, and most characters can occur in isolation, so almost all words would be flagged as compound words.

Character to letter conversion For both tests, we converted characters to a letter-like representation from which one can obtain n-grams. In WikiVocab,

[2] While some languages, like German, capitalize nouns and thus case can be a vocabulary marker, the pipeline is designed to be language-agnostic and thus does not take into account language-specific rules.

[3] There are also some disadvantages to removing compound words. One disadvantage of compound word detection is that lexicalized compound words also tend to be flagged as compound words. For example, the lexicalized compound word "doodskist" in Dutch consists of two parts "dood" (death) and "kist" (box), and has a meaning beyond the two compounds, namely coffin. While the compound word removal might lead to excluding some potential real lexicalized compound words, the removal of compound words is necessary to avoid that the pseudo-words are compound words.

we used custom packages to convert CJK languages to a letter-like string. For Chinese, we use Pinyin [430], Hiragana for Japanese [431], and Jamo for Korean [432]. For the Bible, we checked if the median word length was less than four characters. If this was the case, we assume it is a character-based language, and we converted the characters to Roman letters using uroman [429]. We store the mapping between characters and letters to convert the letters after the sampling back to characters.

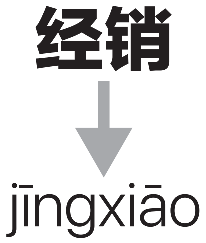


Figure 8.5: Character-to-letter conversion Characters are converted to a letter-like representation. For WikiVocab this is done using language-specific packages, for the BibleVocab this is done uroman [429] if needed.

8.2.2.2 Wikipedia text-cleaning

The following processing steps were applied only to Wikipedia data. Here, we used NLP tools available for the 60 languages but may not be available for low-resource languages. If the language does not have this resource, this process can be skipped.

General pre-processing In addition to the cleaning procedure described in Methods 8.2.2.1, we use the `isalnum` function in Python to filter alpha-numeric strings (note that this is not supported by all writing systems, e.g., Sanskrit, in only 6 of the 60 languages for which we created the task this function was not available).

Avoid jargon Jargon words such as "hippocampus" are problematic because they occur infrequently, but they are not known to all native speakers (only known to domain experts). So to avoid marking jargon words as real difficult words, we removed them automatically. To detect possible jargon words, we computed the ratio between the number of occurrences of a word and the number of articles the word occurs in (Figure 8.6). Jargon words tend to be used frequently in a small number of articles. We, therefore, only kept tokens in the 95 % percentile of the ratio.

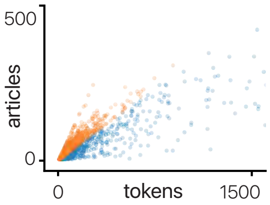


Figure 8.6: Token article ratio Jargon words tend to be used frequently in a small number of articles. One way to detect jargon words is to compute the ratio between the number of occurrences of a word and the number of articles the word occurs in. The blue points are above the the 95 % percentile.

Lemmatizer and POS-Tagger Proper nouns (e.g., "Obama" or "Paris") are problematic in the context of the vocabulary test since they are names and not words. A POS-tagger labels the Parts Of Speech of all words in the sentence. To make the test more comparable across languages, we only select nouns. We also use a lemmatizer to find the lemma of a given token (e.g., "shoes" → "shoe"). Lemmas are better suited for a vocabulary test than tokens since irregular word forms of tokens can obfuscate the lexical item. Sometimes the lemmatization can lead to a misspelled word form (e.g., "ponies" might be lemmatized as "poni", where "pony" would be the correct lemma). We, therefore, only include lemmas which also exist as tokens. There are various POS-taggers, such as Natural Language Toolkit [433] or `spacy` [434]; however, they mainly support English and a limited set of majority languages. We, therefore, use UDPipe 2.0 [435], which supports more than 60 languages and provides additional meta-information, for example, if the token is a foreign word, an abbreviation, or if it is a typo. We used this tool since it is available in many languages.

Spellchecker The generated pseudo-words should follow the regularities of the language. It is, therefore, key to only include typical (i.e., no foreign words) and correctly spelled words. We use the multilingual language-embedding model `fasttext`, which was trained on 176 languages [436], to predict the language from a given word, which filters out most foreign words. Where available, we use open-source dictionaries from LibreOffice [437] to assess if the word is correctly spelled using the Python packages `guess_language-spirit` [438] and `pyenchant` [439]. Certain languages were too close to other languages, leading to a wrong prediction. For example, Western Armenian was flagged as Armenian for 88 % of the tokens. In total, for 13 languages, we could not detect the targeted language in at least 35 % of the tokens. In these cases, we did not exclude words based on the spellchecker. The 13 languages are Faroese, Irish, Scottish Gaelic, Galician, Gothic, Western Armenian, Latin, Maltese, Norwegian Nynorsk, Sanskrit, Northern Sami, Uyghur, and Wolof.

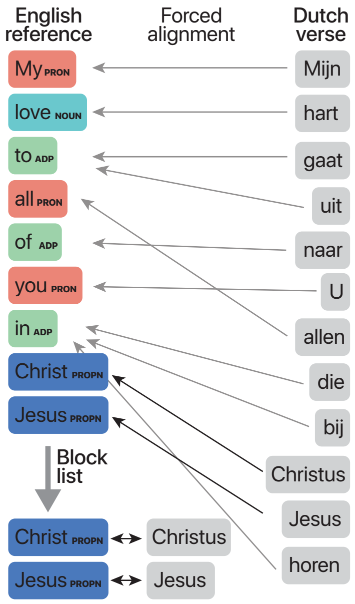


Figure 8.7: Verse alignment Align all verses overlapping with an English reference Bible. Use the Spacy POS tagger to mark proper nouns. Search for proper nouns in the English translation from a curated list.

8.2.2.3 Bible text-cleaning

The following processing steps were only applied to the Bible data. We explicitly did not rely on language-specific resources like dictionaries that are not available in all Bible languages.

Verse alignment: We first align all verses overlapping with an English reference Bible ("New Living Translation") using a universal text alignment tool `fast-align` [440, 441]. We used the Spacy POS tagger to mark proper nouns [434] due to the high performance on English and searched for proper nouns in the English translation from a curated list [442].

Stopword removal: We then aggregated over all occurrences of a stopword (e.g., "Jesus"). For each reference stopword, we retrieved all aligned target words. Since the alignment is not perfect, the same stopword is not always aligned to the same target word. We only included the most common target word if it's used in more than 20 % of the alignments. Since the word might be spelled slightly differently due to inflection, we included words in the stop list that are similar to the top match (> 80 % fuzzy match). We excluded those stopwords obtained from the previous procedure to reduce proper nouns in the translations.

Further processing: For all languages without character conversion, we used compound word detection and removed words of an untypical length (Methods 8.2.2.1).

8.2.3 Pseudo-words generation

Both WikiVocab and BibleVocab used the following steps.

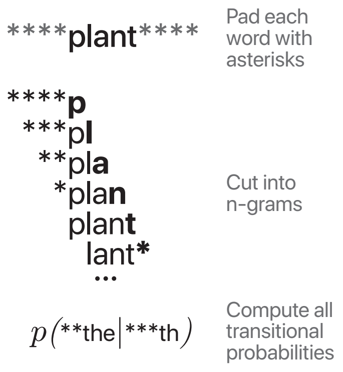


Figure 8.8: Compute n-grams Pad each word in the corpus by asterisks. Cut each word into 5-grams. Compute all transitional probabilities of the 5-grams.

Compute n-grams Existing linguistic work on spoken lexicons of multiple languages has shown that pseudo-words generated from a 5-phone model capture most phonotactic regularities across the real words of most languages [443, 444]. We, therefore, used 5-gram transitional probabilities to create pseudo-words since they are the closest equivalents to 5-phone transitional probabilities for the written language. To track different transition probabilities at the beginning and end of each word, we padded the beginning and end of each word with asterisks (the symbol for word termination).

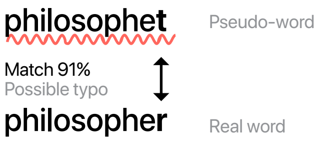


Figure 8.9: Possible typo Pseudo-words that are too similar to real words (> 90% overlap) are likely to be read as typos and thus are rejected.

Sample from n-grams We begin by choosing a sequence of five characters, starting with four asterisks ('****') to signify the word's beginning (see Figure 8.8). For each subsequent sequence, the initial four characters match the final four of the preceding one. This process continues until we select a sequence ending in an asterisk, signaling the word's end. After removing the asterisks, we checked the resulting letter string. Using this padding method, words with fewer than five letters can be created if the termination symbol occurs earlier. We continued this process until 1,000 unique pseudo-words were generated.

Validate pseudo-words We rejected generated pseudo-words that correspond to real words in the language (tested using our corpus). We also rejected pseudo-words that contain too few or too many letters based on the range of word length of the real words in our list for each language (± 2 SD from the median word length). For character-based languages like Chinese, we convert each pseudo-word from the letter-based representation back to the character representation. We do this by replacing all characters. To replace the longest letter sequences first, we sort the letter-character mapping by the length of the letter string. If not all letters in the pseudo-word can be replaced by characters, the word is rejected. For all other languages, we checked if the created pseudo-words are likely to be compound words (as explained in section 8.2.2.1). We rejected the word if this is the case. To avoid the creation of pseudo-words that looked similar to existing words and are potential typos, we compute a fuzzy search using `thefuzz` [445] (see Figure 8.9). Since the total number of words is extremely large, we limited the search to words that start with the first and last three letters and are of a similar length (10 % difference in length allowed). We stored the maximum match between the pseudo-words and any word in that language.

8.2.4 Select real words

To match the task difficulty of our task to LexTALE, thus making them comparable for experiments, we first identified the real LexTALE items in our word frequency distribution (log10-scale, see Figure 8.10). We then compute the mean and standard deviation of the LexTALE items per language. We select real words by finding words with frequencies that are randomized from a normal distribution centered at the average LexTALE word frequency, with the standard deviation being computed over all languages with LexTALE. The same mean and standard deviation were used for all languages, as well as for languages without LexTALE.

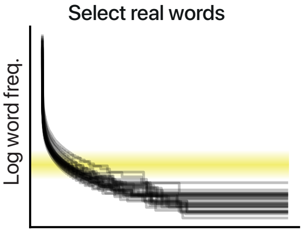


Figure 8.10: Equating test difficulty To establish a consistent test difficulty, we computed the abundance of words with corpus and selected word frequencies that were similar to the one used in previous non-automatic tests [332] (shaded yellow area).

8.2.5 Creating pairs or words and pseudo-words with matched difficulty

In the actual tests, it was important to balance the difficulty of real and pseudo-words. We, therefore, created pairs of tests with similar expected difficulty and structure. Of course, in the actual experiment, the words were presented in random order, so participants could not take advantage of this pairing. Out of the 1,000 created pseudo-words, we selected 500 that best matched the words in that language. To do so, we obtained the logarithm of the transitional probabilities of the letters' 5-grams for both the words and the pseudo-words. We then computed the average absolute difference between the words and pseudo-words that have the same number of 5-grams. On the resulting distances, we performed greedy matching, where we kept matching the word and pseudo-words with the smallest distance. We then matched words and pseudo-words with a similar 'rarity' at the same position (see Figure 8.11). We repeat this procedure until we match all pseudo-words. From the matched list, we only included 500 matched pseudo-words that have the smallest fuzzy match ratio to any of the words in that language. By doing so, we selected the 500 pseudo-words that are least likely to be typos (as typos are hard to detect even for native speakers, especially when words are presented quickly). This procedure tries to ensure that the low-level statistics of characters of words

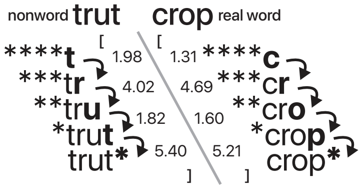


Figure 8.11: Pair word For each pseudoword, the closest real word is selected in terms of log10 conditional probability.

and pseudo-words are similar, thus preventing participants from using this knowledge to resolve the task without real lexical knowledge.

All tests can be administered online: https://vocabtest.org/. The pipeline is available on GitHub https://github.com/polvanrijn/VocabTest and can easily be extended to other languages by providing a text corpus.

In the next section, we will describe the interface shown to the participants.

8.3 Methods

8.3.1 Interface

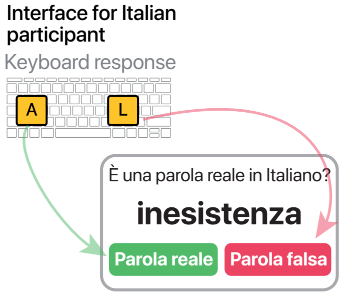

Interface for Italian participant
Keyboard response
A
L
È una parola reale in Italiano?
inesistenza
Parola reale
Parola falsa

Figure 8.12: Interface Interface of the validation experiment for an Italian participant tested in Italian. The word is presented as an image to avoid copy-pasting.

The word lists are presented in a web-based interface (Figure 8.12). Each word is displayed in the center of the screen for two seconds. Participants then have to indicate using their keyboard if the word is real or fake. Words and pseudo-words are presented in random order. To reduce the chance that a participant will search for the word on the internet, we displayed the word as an image (thus, the participant could not copy it as text). To estimate the reliability of the test, participants did two batches of trials per language and test. Each batch contained 30 trials except for the last experiment, where it's 20 trials and 20 repetitions of the same trials. All texts in the interface of the experiment (e.g., buttons, instructions, etc.) were presented in the participant's native language. Non-English texts were automatically translated using DeepL or, if the language was not supported, then by Google Translate.

8.3.2 Participants

All participants provided informed consent according to an approved protocol (Max Planck Ethics Council #2021_42) and were recruited through Prolific (N = 543, 8 languages) and Lucid (N = 3,228, 35 languages). Participants on Prolific received at least 9 GBP. On Lucid, the participant compensation was matched to the local minimum wage as closely as possible.

In the next section, we will extensively validate the pipeline, by conducting three large-scale online experiments involving 3,771 participants speaking 35 languages, resulting in 795,610 total word judgments. In the first experiment, we applied the pipeline to Wikipedia articles (WikiVocab) and showed that the resulting test can accurately identify native speakers, even among non-native speakers of closely related languages on Prolific [418], a common recruiting tool. In the second experiment, we then utilized the pipeline to conduct large-scale language proficiency testing in 35 countries, assessing proficiency in each country across all 35 languages. We show how language proficiency is predicted by participants' demographics, linguistic distance to the native language of the participant, and self-reported language proficiency. In the third experiment, we applied our pipeline to the Bible and provided vocabulary tests for 1,939 languages (BibleVocab). We show that the resulting tests, while created on smaller text corpora and minimal preprocessing, can still distinguish between native speakers in typologically similar languages. This finding demonstrates that the pipeline can be applied to an open-ended number of languages, even ones with relatively low resources.

8.4 Results

8.4.1 Distinguish between close languages on two platforms

To measure the quality of WikiVocab, we benchmark it with eight existing LexTALE tests [332, 407, 410–412, 416] and examine whether the test would be able to distinguish between native speakers of closely related languages. We selected the eight languages to have linguistically remote (e.g., French and Chinese) and close pairs (e.g., Dutch and German). Six languages are Indo-European languages that can be divided into Germanic and Italic subfamilies, each consisting of three languages (English, German, Dutch and Spanish, French, and Italian, respectively). In addition, we also add two typologically distinct languages (Chinese and Finnish, Figure 8.13).

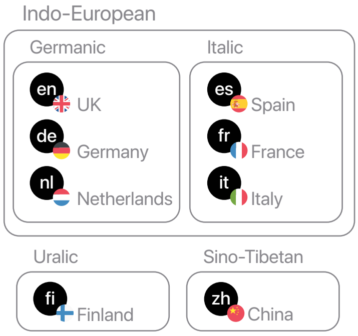


Figure 8.13: Balanced validation design Selected typologically different languages used in a fully crossed study design. At least 35 monolingually-raised participants from the United Kingdom, Germany, the Netherlands, Spain, France, Italy, China, and Finland participate in a language test in their native (L1) and in a randomly selected foreign language (L2) recruited from Prolific and Lucid.

Each participant did LexTALE and WikiVocab in their native language and in one of the other seven foreign languages selected at random. The order of the languages is random. For each test and language, there are two blocks of each 30 items. For comparability with the LexTALE tests, we only included the first 60 items in the test, indicating that all people saw all items in the two languages and tests exactly once.

We run the tests first on Prolific and then replicate them on Lucid. For both recruiters, we invited participants from the same country-language pairs (e.g., Spanish in Spain and not in Mexico), except for Chinese and Finnish since there were not enough participants in Mainland China and Finland on Prolific. Approximately 40 participants were recruited per language, per country, and per platform, which corresponds to the number of participants in LexTALE validation studies [332].

To measure consistency across tests or blocks of the same test, we use Pearson correlations (the 95 % confidence intervals are obtained via bootstrapping, $n = 1{,}000$). Figure 8.14 shows that both WikiVocab and LexTALE are highly reliable, the correlation between the performance on the first and in the second block for WikiVocab ($r = 0.82$ [0.78, 0.85], $p < .001$) was nearly as high as the one of LexTALE ($r = 0.87$ [0.85, 0.89], $p < .001$). The performance of the two tests was highly correlated ($r = 0.85$ [0.83, 0.88], $p < .001$, see Figure 8.15).

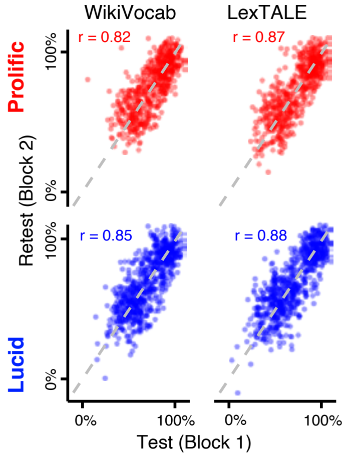


Figure 8.14: Test-retest reliability Pearson correlation between two blocks for WikiVocab and LexTALE in both recruiters.

The heatmaps in Figure 8.15 depict the performance on LexTALE, WikiVocab, and the language self-report. For both vocabulary tests, there is a prominent diagonal, indicating that the native language obtained a higher accuracy compared to the other languages. The average performance on the main diagonal was 88 [87, 89] % in WikiVocab and 89 [88, 90] % for LexTALE, whereas on all other languages, it was 62 [60, 63] % in WikiVocab and 57 [55, 58] % for LexTALE, this difference was significant for both WikiVocab ($d = 2.6$) and LexTALE ($d = 3.2$).[4] This indicates that both tests significantly distinguish between natives and non-natives, but on LexTALE, native speakers reached slightly higher scores and non-natives slightly lower scores.

Importantly, as shown in Figure 8.16, the performance in the L1 was higher than the other languages in the same subfamily (WikiVocab: 68 [65, 70] %, LexTALE: 62 [58, 65] %), was lower even in a language within the same language family (WikiVocab: 64 [62, 66] %, LexTALE: 57 [55, 60] %) and the score for different language families was approximately at chance level (WikiVocab: 56 [55, 58] %, LexTALE: 53 [52, 55] %).

[4] 4: Cohen's d is a measure of effect size, which quantifies the difference between two groups: $d = (M_1 - M_2)/S$ with $S = \sqrt{((\sigma_1^2 + \sigma_2^2)/2)}$

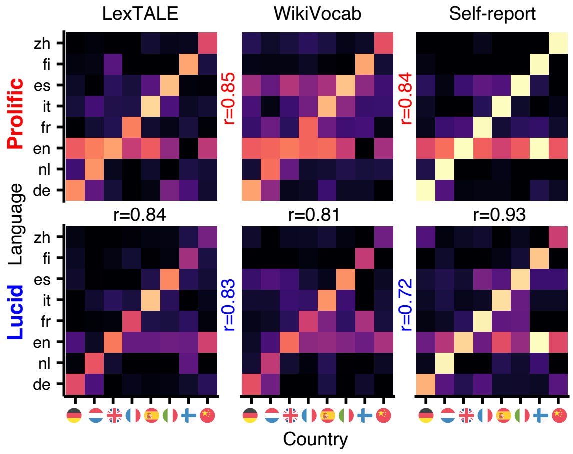


Figure 8.15: Average performance score for LexTALE, WikiVocab, and self-report Average performance score for LexTALE, WikiVocab, and self-report in Prolific (upper panel) and Lucid (lower panel). The x-axis represents the eight countries, and the y-axis the eight languages. The color-fill indicates the accuracy. Vertical Pearson correlations reflect the correlations of the same test in different recruiters. Horizontal correlations show the correlation between WikiVocab and LexTALE and self-report within the same recruiter.

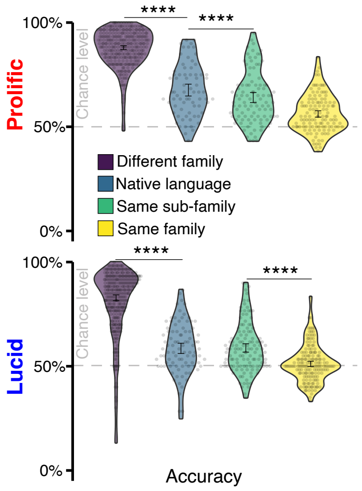


Figure 8.16: Violin plots Violin plot of test performance for native language, same sub-family, same family, or different family. Dots represent the test scores of single participants. The error bar is the standard error of the mean.

The test performance was correlated with self-reports for WikiVocab (r = 0.84 [0.82, 0.86], $p < .001$). Interestingly, all matrices show a horizontal line for English, indicating that most participants are quite fluent in this global language. Overall, in terms of performance, LexTALE slightly outperformed WikiVocab, but considering that WikiVocab was created through an automated procedure, it is quite remarkable that it achieved almost the same high level of performance without utilizing domain experts' knowledge.

As shown in the lower panels of Figure 8.14–8.16, the performance in the two tests was comparable to Prolific. The test-retest reliability was slightly higher on Lucid (LexTALE: r = 0.88 [0.86, 0.9], WikiVocab: r = 0.85 [0.83, 0.87]) than on Prolific. The heatmaps in Figure 8.15 show a similar structure of performance compared with Prolific (LexTALE: r = 0.77 [0.68, 0.84], WikiVocab: r = 0.79 [0.71, 0.86], Self-report: r = 0.90 [0.84, 0.93]). Again, for both vocabulary tests, there is a prominent diagonal (average performance on the main diagonal; WikiVocab: 83 [81, 84] % LexTALE: 85 [83, 86] %), indicating that the native language compared to the other languages (average performance off-diagonal; WikiVocab: 55 [54, 57] %, LexTALE: 54 [53, 55] %) obtained a higher accuracy (WikiVocab: d = 2.1, LexTALE: d = 2.3).

Figure 8.16, again shows that the performance in the L1 was higher than the other languages in the same subfamily (WikiVocab: 59 [56, 61] %, LexTALE: 56 [54, 58] %), was lower or the same in a language within the same language family (WikiVocab: 59 [57, 61] %, LexTALE: 54 [52, 56] %) and the score for different language families was approximately at chance level (WikiVocab: 51 [50, 52] %, LexTALE: 53 [51, 54] %). The results on Lucid differ from Prolific in two ways: The diagonal is slightly weaker on Lucid than on Prolific – indicating the scores on the L1 were not as high as on Prolific – and the horizontal line for English is not as strong as on Prolific, indicating that the English proficiency is lower on Lucid (Lucid: 68.5 CI = [62.2, 74.7] %, Prolific: 77.6 CI = [69.3, 86.0] %). These differences suggest that Lucid participants capture a much wider performance diversity, particularly including participants with lower English performance than Prolific. This is likely due to a higher degree of socio-economic diversity and lesser exposure to English.

8.4.2 Language proficiency survey across the globe

In the previous section, we have shown that WikiVocab can distinguish between closely related languages in two different recruiting platforms. To test if this also holds for a larger set of languages, we conducted a second experiment on Lucid involving 2,798 participants from 35 countries and 35 languages. Those 35 languages span 9 different writing systems (e.g., Cyrillic, Devanagari, Korean) and 15 language subfamilies (e.g., Slavic, Malayo-Polynesian, Indo-Iranian).

Before the main experiment, participants were asked how well they spoke each of the 35 languages and if they learned it in school. Participants were always tested in their first language and in three randomly chosen languages (prioritizing foreign languages spoken by participants).[5]

Consistent with Experiment 1, the performance in the L1 (79.0 %) was higher than other languages in the same subfamily (60.3 %; d = 1.38), was lower even in a language within the same language family (56.2 %, d = 1.79) and the score for different language families was approximately at chance level (51.7 %, Figure 8.17). So, on average, a native German speaker doing the Dutch test (same language subfamily) should perform better than a Spanish native speaker doing the Danish test (same family), and participants doing a language from a different family should do even worse (e.g., an Arabic native speaker doing the Chinese test).

Figure 8.18 shows that the score in the native language was always higher than in the second language, except for Urdu.

The heatmap in Figure 8.19 shows that most languages have a prominent diagonal, indicating the performance in the first language was better than in any of the other languages (d = 1.92). In some cases the performance is symmetric. For example, Slovakian participants do well in linguistically close Czech (79 [74, 84] %) and vice versa (83 [80, 86] %). Similarly, Russians understand Ukrainian (72 [68, 77] %), and Ukrainians understand Russian (82 [75, 87] %). A similar pattern is found for Serbian speakers, who show increased scores for Slovenian (73 [67, 80] %) and Croatian (78 [67, 87] %) compared with their native language (85 [83, 87] %). However, not all relations are symmetrical. For example, Portuguese participants perform better in the Spanish test (64 [59, 69] %) compared to Spanish participants doing the Portuguese test (56 [50, 62] %).

While most languages have a prominent diagonal, some languages have a weak diagonal (Urdu, Norwegian, Hindi, and Hebrew) and are also very close to the diagonal in Figure 8.18, indicating the distance between performance on L1 and L2 is small. One explanation for this difference might be that India, Pakistan, and Israel are highly multilingual societies, which might lead to a different view on language proficiency [424, 446]. However, when analyzing the self-reports, participants indicated that they are most proficient in the indicated L1 and not quite as proficient in the other languages of that country. For Pakistani (Urdu), Norwegian (Norway), and Indian (Hindi) participants, we find a low average performance on all tests (52 %, 55 %, 57 %, and 57 %). Noticeably, Israeli participants performed well on the Russian test, potentially because of the large immigration wave from Russia to Israel in the '90s [447].

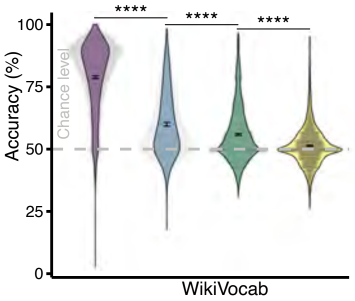


Figure 8.17: Violin plots Violin plot of accuracy in the test for different language comparisons. The same type of plot as in Figure 8.16.

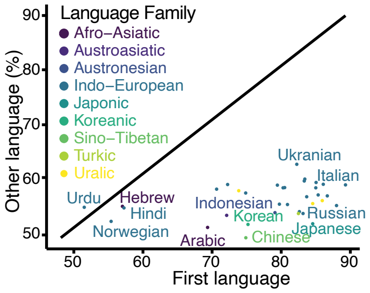


Figure 8.18: L1 vs L2 Performance on the first language vs. other languages. The diagonal line indicates chance level.

[5] This is because if three out of four languages are completely unknown to the participant, performance should be at a chance level, and the task would become random for the participant.

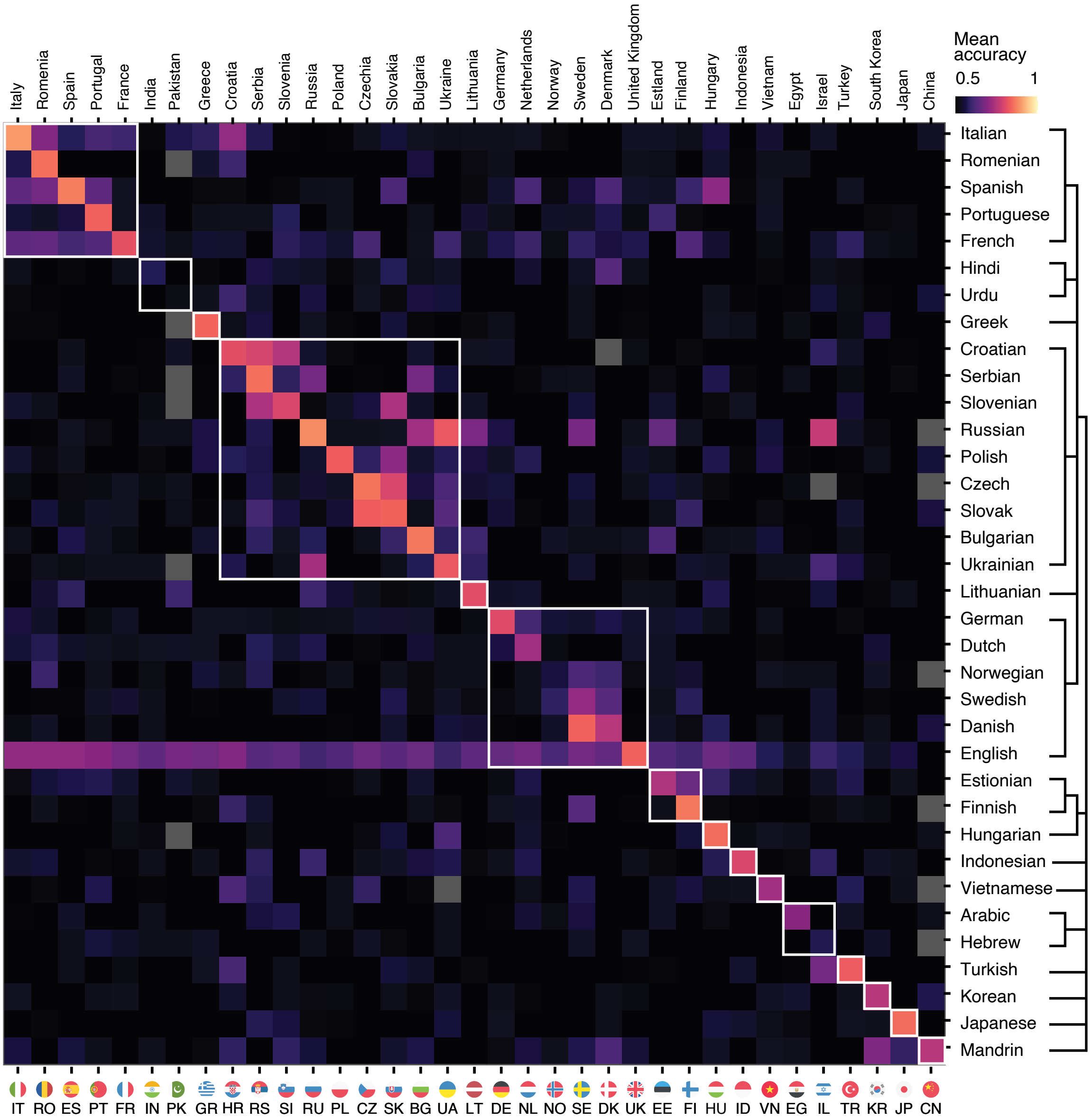


Figure 8.19: Heatmap WikiVocab accuracy for 35 × 35 languages. Languages are sorted by lexical distance [261]. The tree indicates the language family and subfamily.

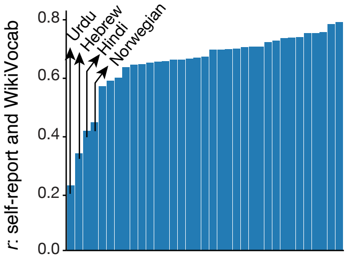


Figure 8.20: Correlation between self-report and WikiVocab Correlation between test-scores and WikiVocab outcome on L1 and L2

In addition, participants from these four countries have the lowest correlations between language self-report (Figure 8.20) and test performance and show low test-retest reliability (Figure 8.21). This leaves open the possibility that participants may have been dishonest in their self-reports.

8.4.3 Generalization to low-resource languages

In the previous two sections, we demonstrated that the vocabulary tests based on Wikipedia articles (WikiVocab) can distinguish between closely related languages and can be used to assess language proficiency across the globe. However, not for all of the worlds this amount of text is available, and quite often NLP tools are entirely missing. To assess if the pipeline can be applied to low-resource languages, we conducted vocabulary tests on the Bible in 1,939 languages using a bare minimum of preprocessing (BibleVocab).

In the experiment conducted on Prolific (n = 240), each participant does BibleVocab and WikiVocab in their native language and in one of the other seven foreign languages selected at random (same languages as in the first experiment). The order of the languages is random. For each test and language, there are two blocks of each 20 items. The items are presented at the end of the block again in a randomized order.

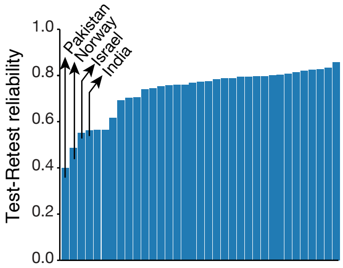


Figure 8.21: Test-Retest correlation Correlation between first and second test block for WikiVocab.

Similarly to the WikiVocab results from the first experiment (d = 2.1), we found that the performance in L1 is higher than all other L2s (d = 2.15, see Figure 8.22). We also found that the accuracy on BibleVocab is significantly higher in L1 compared to other L2s (d = 2.25). However, we found that the mean score on the L1s is slightly lower on BibleVocab (81.5 [80.5, 82.5] %) than in WikiVocab (89.7 [88.9, 90.4] %), which is caused by the minimal preprocessing, small size and the use of old and less familiar words in Bible.[6]

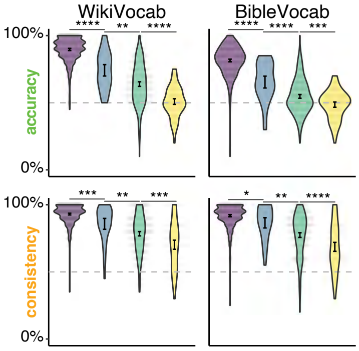


Figure 8.22: Violin plot Violin plot of accuracy in Bible- and WikiVocab for native language, same sub-family, same family, or different family. For accuracy (top) and consistency (bottom).

This also explains the lower test-retest reliability in BibleVocab (r = 0.68 [0.63, 0.73]) than in WikiVocab (r = 0.78 [0.74, 0.82], see Figure 8.23). Both scores were slightly smaller than previous experiments because of the reduced number of items. The reduced performance can come from misclassified items. One way to reduce false positives and negatives is to use item selection, namely, by presenting the same test items to a larger pool of participants and selecting those that allow distinguishing between the majority of participants. Item selection can significantly improve the test reliability while also avoiding other issues that our approach does not currently cover.

If one wants to avoid running an additional experiment, one can also use item repetition instead. Namely, to show the same item twice and measure response consistency. The hypothesis is that non-native participants are less consistent in their choices because it is difficult to remember their previous responses to the same item. Native speakers, on the contrary, should be able to determine the lexicality of an item and thus should be more consistent. To test this hypothesis, we use the fact that we repeated this experiment every item twice (for each participant), we found that native speakers are significantly more consistent in their choices than non-natives (native: 93.0 [92.4, 93.7] %, non-native: 77.1 [75.6, 78.5] %, d = 1.28) and BibleVocab (native: 91.8 [91.1, 92.6] %, non-native: 76.1 [74.6, 77.6] %, d = 1.17) (Figure 8.23, lower panel). This suggests that item repetition can be used as an alternative method to use the test, even if some items are mislabeled.

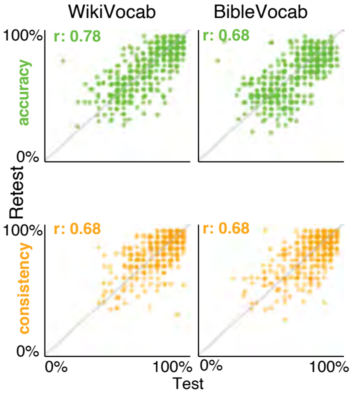


Figure 8.23: Test-retest reliability Test-retest reliability between two blocks for Bible- and WikiVocab regarding accuracy and consistency.

8.5 Discussion

8.5.1 Summary

In this chapter,

- We proposed a fully automated pipeline to create vocabulary tests for an open-ended number of languages only requiring a text corpus.
- We suggested two versions: one uses Wikipedia articles for 60 languages, and the other uses the Bible for 1,939 languages.
- We benchmarked the pipeline with eight existing LexTALE tests and showed that the test can accurately identify native speakers, even among non-native speakers of closely related languages on Prolific, and replicated these findings on Lucid.

[6] For example, many biblical names – like “Jehoshaphat” or “Mephibosheth” – are marked as real words and real words that don’t occur in the Bible – like “hash” or “twinkle” – are marked as pseudo-words.

- We validated the vocabulary tests in 35 languages spanning 9 different writing systems and 15 language subfamilies and showed that the pipeline can still distinguish between native speakers in typologically similar languages.
- We showed that the pipeline is fairly robust to smaller-sized text corpora and less preprocessing of the source material.

8.5.2 Limitations and Outlook

While the pipeline is a powerful tool for assessing language proficiency, it has some limitations:

- **Necessity of objective tests**: The findings show that most participants honestly self-report their language proficiency, which raises the question of whether an objective test is necessary at all. However, when doing research, one often needs to corroborate self-reports with objective proficiency tests. This is particularly important for multilingual societies, the language tests allow screening of the linguistic background of a whole participant population. For example, for Israeli online participants, we observed they were most fluent in Russian, suggesting that test participants include immigrants.[7] While self-report and WikiVocab performance are correlated in most countries in both recruiters (Prolific: r = 0.84 [0.81, 0.87], Lucid: r = 0.72 [0.67, 0.77]), we also find strong differences across countries. In particular, we found participants who report speaking Hebrew, Hindi, Norwegian, and Urdu have low test performance in their native language but also have low correlations between self-reported languages and test performance in general, which indicates self-reports are misaligned with the test scores on the different languages. In those four languages, we also observed substantially lower test-retest reliability (r = 0.40–0.55) from the other languages (around r = 0.80, Supplementary Figure 8.21), highlighting the relevance of testing language proficiency before the experiment and not solely relying on self-reports.
- **Lab studies**: The pipeline has been developed with running online studies in mind, emphasizing a high degree of automatization, and modularity, and making the pipeline as language-agnostic as possible. However, this comes at a cost, since not all possible orthographic and grammatical rules of the language are followed. For example, our pipeline produces lowercase nouns in all languages, even though the first letter of German nouns should be capitalized. Another example comes from Hebrew, where prefixes or suffixes integral to the word (such as the Hebrew letter "Vav" and "Yod") are not removed, though a domain expert creating the test would likely exclude them (because they can confuse native speakers in a similar way to typos, which we do eliminate from our test). In the context of online experiments, this is acceptable since these rules are usually not known by less proficient speakers or give no information about the lexicality of test items (e.g., all words are lowercase). However, this can be a problem for lab studies that require more control over the stimuli. To have the best from both worlds of manual and fully automated vocabulary test creation, one can use the pipeline to propose items for the vocabulary test and then have domain experts dismiss items based on a predefined set of rules. This is particularly important for multilingual projects because (a) the tests exist in a large number

[7] Approximately 12.4 % of the Israelis are Russian-speaking [448].

of languages and can easily be extended to new languages, (b) all items are generated from the same language-agnostic rules, and (c) linguistic rules can be specified before filtering which removes the subjectivity inherent to LexTALE and its variants.

- **Only one aspect of language proficiency**: Language proficiency is a multifaceted construct that includes vocabulary, grammar, pragmatics, and spoken language skills. With the test, we only assess receptive vocabulary knowledge, since the goal was to have a psychophysically reliable and rapid test for language proficiency. Future research can extend the pipeline to assess other aspects of language proficiency or apply the pipeline to non-linguistic proficiency. For example, the same pipeline can be used to distinguish between domain experts and non-experts by creating real and pseudo-words from a large selection of technical terms and asking experts and non-experts to judge which term exists and which does not.
- **Bots and automated quality control**: Large Language Models and their recent extensions to audio and vision pose a threat to online experiments [449]. The vocabulary tests developed here can serve as a kind of CAPTCHA to distinguish between human participants and bots [347]. To avoid cheating, items of the test are presented as images and the URL of the image is obfuscated. However, the advent of multimodal foundational models like GPT-4 [450] pose a threat to this approach, as they can easily read text from images. Future research can try to add noise to the images to make it harder for the model to read the text or use a TTS model to present the items orally, which are also becoming increasingly available for low-resource languages [427]. An oral version of the test will also enable extending the pipeline to test participants in less literate societies.

Chapter 9

Color naming across the globe

Based on
Jakob Niedermann, Ilia Sucholutsky, Raja Marjieh, Elif Celen, Thomas L Griffiths, Nori Jacoby, and **Pol van Rijn**. 2024. 'Studying the Effect of Globalization on Color Perception Using Multilingual Online Recruitment and Large Language Models'. *CogSci*.

To put the global recruitment infrastructure to the test, we conducted a large-scale, cross-cultural online experiment on color naming. This is a particularly interesting instance of grounded semantics to study, because (a) it is low-dimensional (only requiring three dimensions to describe) and much less high-dimensional than emotional prosody [123] and (b) it is well-studied [84–98].

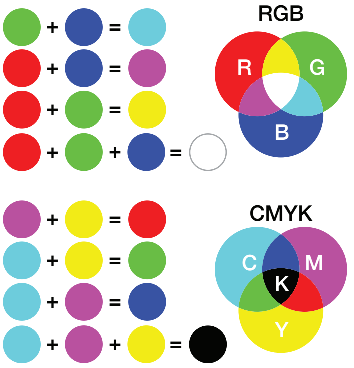


Figure 9.1: Additive and subtractive color spaces Top: additive Red, Green, Blue (RGB) space and subtractive Cyan, Magenta, Yellow, Key (Black) (CMYK) space (Cyan, Magenta, Yellow, Key/Black).

9.1 Background

9.1.1 Language shapes thought?

Research on grounded semantics has often been centered around the concept of linguistic relativity. This is the question if language influences thought, and, if so, how [451]. The study of how people name colors has been the most influential modality to study this question [86–95] because color perception has both linguistic (color words) and biological aspects (e.g., visual systems or evolutionary constraints).

Before we will present both sides of the color naming argument, we will speak about how we can define colors.

9.1.2 Color spaces

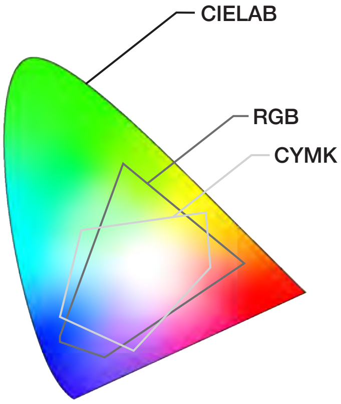


Figure 9.2: CIELAB color space Color space with three components: L* (lightness), a* (green to red), and b* (blue to yellow). The colors expressible with CMYK and RGB are a subset of the colors expressible in CIELAB.

Colors can be described as tuples of numbers, typically three or four values. The space of colors can be parametrized differently, each tailored to a different use case. There are two main types of spaces: additive and subtractive color spaces (Figure 9.1).

The Red, Green, Blue (RGB) color space is additive and combines different intensities of red, green, and blue light, where each channel is represented by a value ranging from 0 to 255, corresponding to the intensity of the color. Thus, if all channels are set to 0, the color is black, and if all channels are set to 255, the color is white. This space is commonly used for digital displays like computer monitors and digital cameras, where each pixel consists of three subpixels, each emitting one of the three colors.

There are also subtractive color spaces like Cyan, Magenta, Yellow, Key (Black) (CMYK) developed for printing, in which colors are created by subtracting varying percentages of cyan, magenta, yellow, and black inks from white paper.

However, both color spaces do not capture all perceivable colors. The CIELAB color space (Figure 9.2) is a device-independent color space that includes all perceivable colors. It is designed to be uniform with human vision and consists of three components: L* (lightness), a* (green to red), and b* (blue to yellow). The CIELAB color space is used in color management and to ensure color consistency across different devices. One can also use CIELAB color differences to quantify the perceptual difference between two colors.

9.1.3 Color naming studies

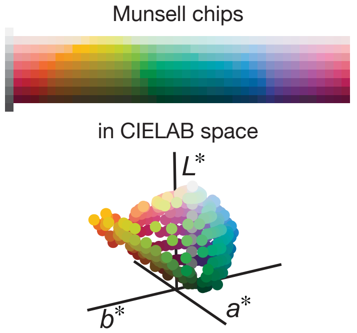


Figure 9.3: Munsell color chips Top: Palette of all 330 Munsell color chips. Bottom: All chips are projected into the CIELAB color space.

Berlin & Kay's (1969) Basic Color Terms: Berlin & Kay (1969) collected color naming data on speakers of 20 languages in the San Francisco Bay Area [452]. Each participant was shown 320 chromatic – subdividing the hue axis into 40 steps and 9 levels of brightness – and nine non-chromatic Munsell chips, ranging from white to black (see upper part of Figure 9.3) [453]. Note that the Munsell chips are not perfectly perceptually uniform (when projecting the chips into CIELAB space the locations of these color values do not yield a perfect symmetric shape, see Figure 9.3). The participants were asked to name each chip using a Basic Color Term (BCT) [452], which are defined as:

- **Monolexemic**: The term must be a single word, not a compound or descriptive phrase. For example, "red" is a basic color term, but "light red" is not.
- **Non-specialized**: The term should not be restricted to specific contexts or objects. In other words, it must be applicable to a broad range of objects and not limited to specific items or materials. For instance, "blonde" (used only for hair or beer) does not fulfill this criterion.
- **Psychological salience**: The term should be psychologically significant, easily recognized by speakers of the language, and should partition color space exhaustively.

Berlin & Kay (1969) show that there are substantial differences in color naming across languages, including the number of color terms across and their position in the color space. However, the study was also criticized for its methodology, because it covered ninety-eight languages with only twenty participants, so many of these participants were highly multilingual. Furthermore, all speakers resided in San Francisco, which may not accurately reflect the native linguistic environments.

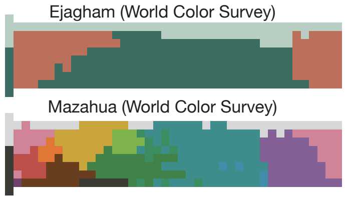


Figure 9.4: Two color maps from the World Color Survey Maps of majority colors in languages Ejagam and Mazahua from the WCS. The colors are the mean RGB color of all chips with the same majority color.

World Color Survey: To overcome these flaws, they conducted a large follow-up study, the World Color Survey (WCS) conducted from 1976–1980, in which color naming patterns were studied for speakers of 110 unwritten languages, with low exposure to other cultures [226] (see Figure 9.4 for two maps). In this study, the same methodology was used as in Berlin & Kay (1969), but a purely white chip was added to the palette of Munsell chips (now 330 chips). The data was collected in a twofold manner. First, participants name each chip using a BCT (colors presented in a random order). Second, based on the elicited color terms a list of BCTs was created, and for each color participants were asked to select the color term that best represents the color chip.

9.1.4 Universalist Theory

The following findings have been used to support the universalist theory:

- **Berlin & Kay (1969)** have claimed that there are maximal eleven color categories that align with the English color terms (black, white, red, yellow, green, blue, brown, orange, pink, purple, and gray) and other languages would either use these categories or a subset of them down to 3 color terms. They argue that the color naming pattern is constrained by evolution, such that languages with the same number of color terms have similar color maps.
- **Lindsey & Brown (2006)** reanalyzed the WCS data using k-means clustering on single participant color naming data and found that the optimal clustering result (8 clusters) closely resembled the actual color categories present in English [454].
- **Regier et al. (2007)** used irregular color samples from the CIELAB space of Munsell color chips (see Figure 9.3) to explain the location and shape of color categories in WCS [455].
- **Zaslavsky et al. (2018)** used the information bottleneck theory to explain the evolution of color terms in the WCS languages, showing that the WCS languages achieve nearly optimal communicative efficiency [456] and are compatible with iterated learning experiments on color [67] (see Section 2.2.2).

Taken together, these studies suggest that while languages may differ in the number of color terms, the process of color naming is constrained by universal principles.

9.1.5 Linguistic Relativity

The following findings have been used to support the linguistic relativity theory:

- **Linguistic categories influence perceptual boundaries**: Languages differ in the number of BCTs. For example for Russian [88], Greek [91], Turkish [89] and German [87] have different color categories than English. It has been shown that speakers of languages with different BCTs differ in their speed and accuracy in distinguishing between colors of BCTs in their own vs a foreign language [90, 239], showing that linguistic categories (BCTs) influence perceptual boundaries (differentiating between colors).
- **Linguistic categories influence memory**: Lowry and Bryant [238] have shown that speakers of Russian and English are influenced by their color representation of blue when remembering the color of objects that are shown in a range from blue to gray.

9.1.6 Implications of the color naming debate

The color naming debate has shown that languages differ in the number of color terms, but that the placement of these color terms is not random and languages with varying numbers of color terms are near-optimal communicative efficiency.

On the other hand, it has been shown that linguistic categories have downstream effects on color discrimination and memory. A larger literature has shown that the language and cultural background affect cognitive abilities not only for language-related tasks, such as emotion semantics [60], but also for ostensibly nonlinguistic abilities, including memory [263], space [237], time [457, 458] and sensory-perception [99, 243, 259, 459].

Studies have shown that the number of color categories in a language can rapidly emerge over time [92, 93] indicating that new color categories can be rapidly adopted and used.

These findings indicate that color naming evolves over time and may also be influenced by cultural factors, such values [460, 461], economy [462, 463], and more recently, globalization [464, 465].

To study if cross-lingual differences in color naming still persist in the age of globalization, we conducted a large-scale, cross-cultural online experiment on color naming. We compare the results against a highly globalized agent, namely LLMs trained on multilingual digital content from around the world. LLMs are particularly interesting from the perspective of the study of the interaction between perception and language [466], as they are trained on a substantial chunk of human language and can be used to interrogate the limits of perceptual information that can be extracted from language [7, 18, 467]. Still, it is unclear if LLMs will capture the variability in color naming across languages as it is mainly trained on English data [468].

9.2 Methods

9.2.1 Participants

Participants were recruited from both Prolific (N = 517) and Lucid (N = 1,763) and had to speak the language as their mother tongue, be raised monolingually, hold nationality, had to be born in the target country and had to pass a vocabulary test [10] to make sure they were indeed speakers of the designated language. Participants were recruited from 22 languages (see Figure 9.5). Data of color-blind participants were excluded from the analysis based on the results of a color blindness test [1, 469]. The wage per hour was adapted to the local minimal wage.

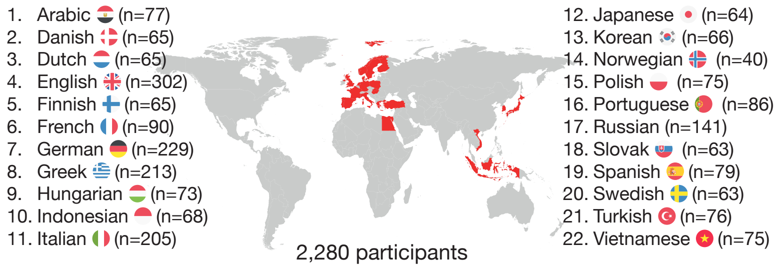
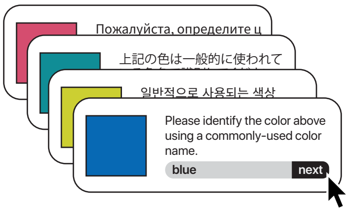


Figure 9.5: Participants Worldwide recruitment in 22 different languages.

9.2.2 Procedure

In accordance with the literature, the participants were presented with 330 Munsell color chips [470] through a web interface (see Figure 9.6). To prevent fatigue, each participant was only presented with a random subset of 50 color chips.



Пожалуйста, определите ц

上記の色は一般的に使われて

일반적으로 사용되는 색상

Please identify the color above using a commonly-used color name.

Figure 9.6: Task Participants were asked to name the color chip using a basic color term.

The following prompt was translated into all 22 languages using a professional translation service [471]:

> Please identify the color above using a commonly used color name. The color name should be the one you would normally use in everyday life to describe that color. Avoid using compound words. The color name should be a single-word

9.2.3 Language Models

All LLM experiments are conducted with OpenAI's GPT-4 [450] using the Microsoft Azure OpenAI API (version 0613 of the model) using the default temperature parameter (0.7). The following system prompt was used:

> Follow all instructions that users provide to you in their own language. Respond to users in their own language using only a single word and no other text. Do not use any compound words.

Followed by the translated user prompt:

> COLOR: `<hexcode>` Please identify the color above using a commonly-used color name. The color name should be the one you would normally use in everyday life to describe that color. Avoid using compound words. The color name should be a single word.

Since GPT-4 responses are stochastic at non-zero temperatures, 50 responses are sampled for every color for each language.

The experiments are repeated with the Vision Language Model (VLM): OpenAI's GPT-4V (version gpt-4-vision-preview on Microsoft Azure's OpenAI API). The model would receive the same prompt as GPT-4 but with an image of the color patch encoded in Base64. Due to rate limits at the time of the experiment, only a single response was obtained for each color for each language (at a temperature of 1).

9.2.4 Preprocessing

Since participants and LLMs provided free text, the responses had to be processed. Responses containing spaces, digits, or punctuation marks were removed. Furthermore, the word had to be written in the expected script (e.g., a Russian color term in Cyrillic). Lemmatization was performed to remove word variants [472] and replaced them with the most common variant. In the next steps, diacritics were removed (e.g., "rosá" to "rosa") and replaced characters with smaller units in Korean (Hangul) and Japanese (Katakana) to detect the same word written differently. For character-based languages, also the word "color" was removed since the word was often added to the color term. We replaced all variants under the same simplified form with the most common variant. To detect compound words, we identify the top color terms (occurring in $>$ 1% of all responses) and check all other color terms for whether they end with this term. If they do and also co-occur, the compound word is replaced with the top color term. For example, in Dutch, we would replace "donkerblauw" (dark blue) with "blauw" (blue). For all words, we look up their word frequency in a large text corpus [473]. Words that do not occur in the corpus are unlikely to be color terms normally used in everyday life and are likely to be typos. For each of the typos, we obtain a list of color terms it co-occurs with. For each of those terms, we compute the Levenshtein Distance [474] and merge if they match (score > 80%). We exclude all color terms that are used less than five times in total. For all minority colors (occurring less than 1% of all responses), we merge them with the majority color it co-occurs most with. Terms without co-occurrence are removed. Terms are only removed if this would not lead to removing all color terms in one chip (e.g. if a particular term is used only for one chip and never for the remaining 329 chips). Since we only have a single response per chip from GPT-4V, we did not apply the pipeline to the VLM.

9.3 Results

To show the differences in color patterns, we plot the majority color term for each chip in the Munsell space, by averaging the RGB values of the chips that were assigned to the same color term. All maps can be viewed interactively on: `https://global-colors.s3-eu-central-1.amazonaws.com/index.html`.

9.3.1 Consistent with previous data and across recruiters

Generally, we can see that the English color maps are consistent across the two recruiters (Prolific and Lucid) and are similar to previous data [92] (Figure 9.7). To measure the similarity between the color maps, we computed the Adjusted Rand Index [475], which measures how similar different clusterings are. We found a high ARI between [92] and Prolific and Lucid (.72 [.67, .77] and .67 [.62, .72], respectively, CIs via bootstrapping). We further found that for languages that were tested on both recruiters (English, Italian, and Greek), the ARIs were also high (.70 [.65, .74], .61 [.56, .66], and .63 [.57, .68], respectively). These findings suggest that the human maps were reliable, and our findings are consistent with prior literature for English.

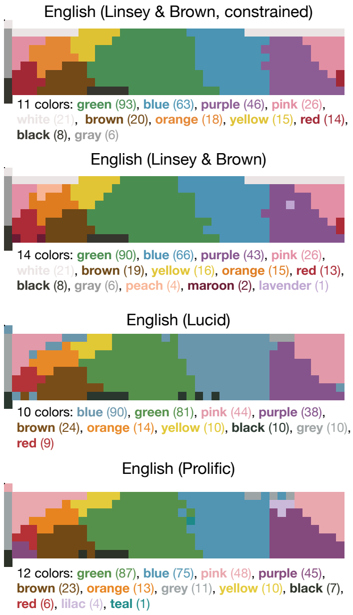


Figure 9.7: Consistency across data English color maps.

9.3.2 Variability across languages persists

However, when we look at the color maps for Dutch, Italian, Russian, and Japanese, we see that the variability across languages persists (Figure 9.8).

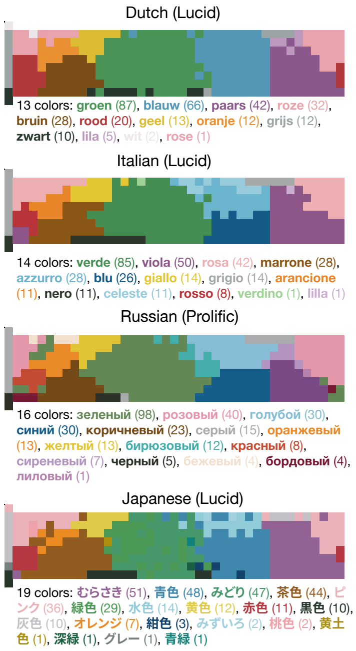


Figure 9.8: Variability across languages persists Collected maps for Dutch, Italian, Russian, and Japanese.

While Dutch still resembles the English color maps (.70 [.65, .76]), Italian (.59 [.55, .63]), Russian (.55 [.49, .60]), and Japanese (.42 [.38, .46]) show a more distinct pattern with more differentiation between light blue colors [90]. In Italian, for example, we see a distinction between three shades of blue: "blu", "azzurro", and "celeste".

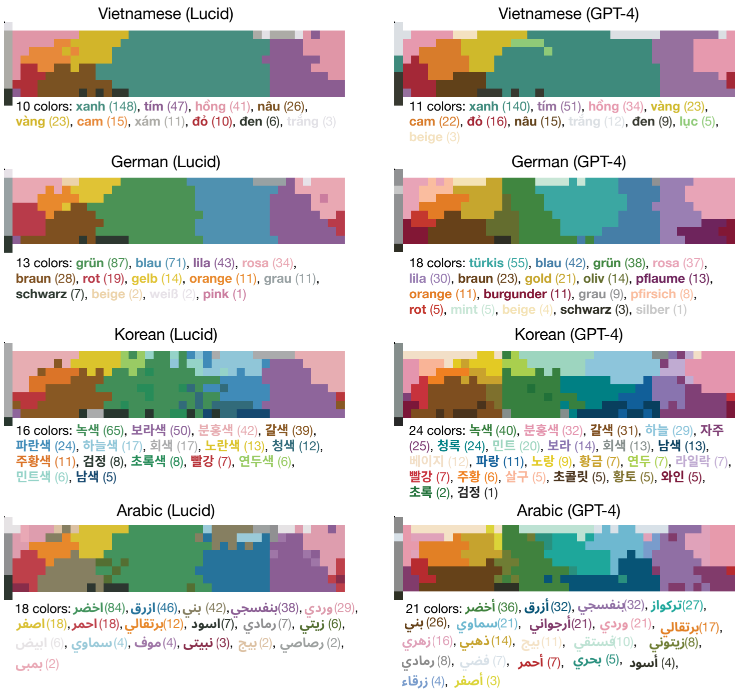


Figure 9.9: Human vs GPT Comparison of languages in humans and GPT-4.

9.3.3 LLMs use a larger number of color terms

Figure 9.9 shows examples of human and LLM color maps from the same language. While the maps look qualitatively similar and have similar color terms for the same language, GPT-4 maps exhibit a larger number of color terms.

To quantify this effect, we plotted in Figure 9.10 the number of color terms and the number of distinct responses per chip. We found that the newly collected color maps (red) significantly have fewer color terms compared to the GPT-4 color maps (blue; $p < 0.001$), but significantly more color terms compared to the WCS color maps (green; $p < 0.001$).

The vertical axis of Figure 9.10A represents the average consensus (number of distinct responses) within each chip. The WCS data showed the most diverse consensus, while GPT-4 provided far fewer distinct answers compared with our new human data. Due to the way the experiment was conducted, GPT-4V always provided a single answer (temperature 1).[1]

[1] Also GPT-4 with a temperature of 0.7 provided approximately a single answer per chip.

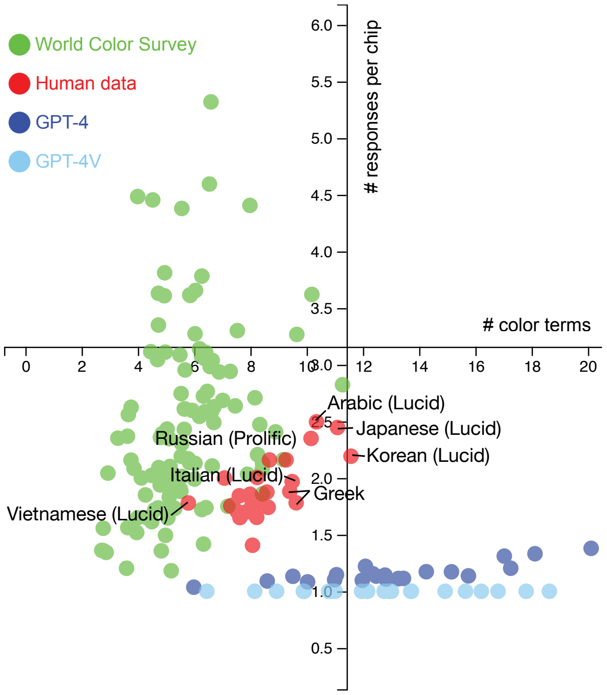


Figure 9.10: Color entropy space The x-axis is the number of color terms (exponent of the entropy of the color map), and the y-axis is the number of distinct responses per chip (exponent of the mean entropy per color chip).

9.4 Discussion

9.4.1 Summary

In this chapter,

- we collected color naming data for 22 languages using two online recruiting platforms.
- we found that the English color maps were consistent with the literature and across the two recruiting platforms.
- we showed that color naming patterns still vary across languages, even in globalized societies.
- we demonstrated that GPT-4 uses a significantly larger number of color terms than the online participants and uses significantly fewer different responses per chip.

9.4.2 Limitations

- **Free naming task**: The free naming task might have led to more diverse responses than a constrained task. It remains to be seen whether other paradigms such as color discrimination [90], serial reproduction [67], and similarity judgments [7] would lead to similar conclusions.
- **Unrepresentative sample**: Our sample does not have enough representation of the global South, in particular, South America and Africa [259, 464]. Future work should cover such locations.
- **Individual differences**: Our analysis was focused on data at the population level, but it is reasonable to assume that there is some degree of variation in color concepts across individuals from within the same culture [92]. Future work should probe individual-level variation in color naming.

- **Color calibration**: In the online experiments, participants did not conduct a color calibration of their screen. Future work should compare the online data with in-lab data.

Part III
Applications

In the last part of the thesis, we show that the HITL paradigms developed to solve core methodological problems in emotional prosody can also be used to solve more applied issues in the voice modality. In Chapter 10, we show that GSP (see Chapter 4) can be used for personalization by customizing voices for voice assistants and digital avatars. In Chapter 11, we show that GSP in combination with STEP can be used to align impressions across modalities (here, the auditory and visual modality).

Chapter 10

VoiceMe: Personalized voice generation in TTS

Based on

Pol van Rijn, Silvan Mertes, Dominik Schiller, Piotr Dura, Hubert Siuzdak, Peter M. C. Harrison, Elisabeth André, and Nori Jacoby. 2022. 'VoiceMe: Personalized Voice Generation in TTS'. *Interspeech*.

In this chapter, we demonstrate how a HITL approach can address a significant practical challenge in a related voice domain. We address a problem central for human-computer and human-robot interaction, specifically, creating a voice that fits an artificial agent. People make rich inferences about faces, including the emotion, personality, age, or background of the person [59, 476–479]. This implies that the voice should match the impression of the face. For example, if one sees a picture of a young girl, but one hears a deep, smoky voice this conflicts with our priors (children tend to have higher-pitched voices) and beliefs about the world (children don't smoke). This mismatch can lead to eerie feelings, as seen in the uncanny valley effect [25].

In Chapter 4, we showed how one can use GSP to characterize prototypes of emotions in speech. Here, we use GSP to a related domain where participants try to find voices that match the mental representation of faces. In contrast to the previous work on emotional prosody (see Section 4.3 and 4.4), participants do not optimize for a specific attribute (e.g., "happy"), but for a voice that matches the face. This tool can be used to create personalized voices for robots, speech assistants, or fictional characters, it can bring paintings to life, or help people with speech impairments find a voice that matches their mental representation.

10.1 Background

Recent multi-speaker text-to-speech (TTS) models can create entirely new high-quality voices [480, 481] that were not seen during training. Jia et al. [480] demonstrated that an independently trained speaker encoder network trained on a speaker verification task can produce useful conditioning for a multi-speaker text-to-speech model. By sampling random points from the obtained speaker embedding space, the authors generated fictitious voices that were not seen during the training. Another approach was proposed by Stanton et al. [481] that does not rely on transfer learning from the speaker verification task, but jointly learns a distribution over speaker embeddings, also allowing for sampling a novel voice. However, all of these papers focus on speaker generation, but not on speaker personalization, which we try to address here.

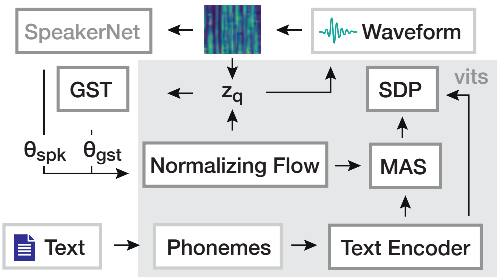


Figure 10.1: Modified VITS architecture The architecture used in this paper is a modified version of the VITS model using SpeakerNet (θ_{spk}) and GST embeddings (θ_{gst}). z_q is the posterior latent sequence for a speech sample.

10.2 Methods

10.2.1 VITS architecture

Here, we used the back then state-of-the-art TTS model VITS [315]. We chose this model over Tacotron 2 [317], which was used in the previous work on emotional prosody [3], because (i) it is not autoregressive, which makes it faster to train and sample from (the real-time factor is also relevant for the experiment) and (ii) VITS is trained in an end-to-end fashion, which leads to a higher quality of the generated speech than Tacotron 2, which first predicts mel spectrograms and then uses a vocoder to generate the audio. We chose VITS over other non-autoregressive models, because (i) it was frequently used [334], (ii) the implementation was available, and (iii) there is a pretrained model available that can be used for transfer learning.

VITS consists of the following components (depicted in Figure 10.1):

- **Text** frontend composed of text normalization followed by a grapheme-to-phoneme model producing **phonemes** [482].
- Transformer-based **Text Encoder** with a projection layer used to construct the prior distribution for speech generation.
- The model uses a latent variable z_q to represent the acoustic features of speech.
- A **Normalizing Flow** is used to map the latent variable z_q to the waveform space. Normalizing flows are invertible and allow for efficient sampling and density estimation.
- The Monotonic Alignment Search (**MAS**) ensures proper alignment between text input and audio output.
- The Stochastic Duration Predictor (**SDP**) models the timing and duration of speech sounds
- The spectrogram is converted to a waveform using the HiFi-GAN [483] vocoder architecture.

All components are trained jointly in an end-to-end fashion.

We make two extensions to the VITS model: SpeakerNet and GST.

SpeakerNet is a speaker verification network that was pretrained on a large dataset of speakers. Previous research has shown that the embeddings extracted from a speaker verification network can be used to generate new voices [480]. We, therefore, used a pretrained SpeakerNet-M [484] model to extract speaker embeddings.

We also used a bank of Global Style Tokens (GST) [316] to extract style embeddings from encoded spectrogram frames. We initialized 16 Global Style Tokens with 8 attention heads and the resulting embedding size was set to

256. For the GST encoder, an 8-layer convolutional network was used with the same architecture as the Posterior Encoder. During training, the speaker and style embedding were separately L2-normalized and concatenated. During the experiments, we use the same zero embedding to keep prosody approximately constant across samples.

To prevent the GSTs from learning speaker-dependent features in the presence of SpeakerNet embeddings, an additional adversarial loss is proposed as follows. Alongside discriminators, a separate shallow feed-forward neural network is trained to reconstruct the speaker embeddings from extracted style embeddings. During the discriminator step, this network minimizes a cosine distance between real and reconstructed speaker embeddings using cosine embedding loss: $(1 - \cos(x, \hat{x}))$. Conversely, during the generator step, the style extractor is penalized if this network succeeded in reconstructing speaker embeddings – the loss function is then: $(\max(0, \cos(x, \hat{x})))$.

We applied transfer learning from the publicly available VCTK checkpoint [315] and the training was continued using two NVIDIA V100 GPUs. For the first 400k iterations only the discriminators were allowed to train due to the lack of a published discriminator checkpoint, and then normal training continued for an additional 2M iterations with the learning rate lowered to 1e-4.

10.2.2 Faces

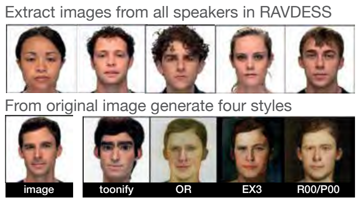

Extract images from all speakers in RAVDESS
From original image generate four styles
image
toonify
OR
EX3
R00/P00

Figure 10.2: Image materials Images are extracted for all speakers in the RAVDESS corpus (CC BY-NC-SA). We use deep-learning style transfer to convert the images to cartoons and paintings.

To create a dataset of faces, we extracted stills from the RAVDESS corpus [44], in which actors were video-recorded saying sentences with a neutral meaning for a variety of emotions and "neutral". For all 24 actors, we extracted the still from the same sentence that was intended to sound neutral.

To demonstrate our approach does not only work for real faces but also for fictional characters, we created for each original image four fictional characters based on style transfer. We used toonify [485] and three additional art portrait styles from Ai Gahaku [486]: `OR`, `EX3`, and `R00` or `P00`. We selected the images in the following way. We start by creating 12 art portraits and one toonified version and then select four styles with the highest perceptual similarity to the real photo [487] (see Figure 10.2). Thus, we select toonify, `OR`, and `EX3` styles, but in 22 of 24 cases, we select `R00` and in all other cases we select `P00`.

For all 24 speakers, we use the extracted images and four styles with the highest perceptual similarity totaling 120 chains. To each chain, we randomly assign one of the 720 phonetically balanced and semantically neutral Harvard sentences [310].

10.2.3 Parametrization

In the current experiment, participants change the first ten principal components of the SpeakerNet embeddings. Initial piloting suggested that these principal components had the desired property of intuitive interpretability (e.g., PC2 has a strong gender effect), and prior research with related models suggested that 10 principal components should be enough to achieve meaningful control over the stimuli [1]. The principal components were computed on SpeakerNet embeddings extracted on a single utterance of the 45,825 speakers present in the train, dev, and test partitions of the English CommonVoice dataset [488] and account for 25.4 % of the variance.

10.2.4 Interface

The participants are prompted to adjust a slider that corresponds to one principal component to make the voice maximally similar to a face (see Figure 10.2 for some of the faces). For practical reasons, every slider contains a finite resolution of 31 equally spaced slider positions. As opposed to using static images, we use Wav2Lip [489] to synchronize the lips to the voice so that the resulting stimulus looks more natural.

10.2.5 Participants

All participants were recruited from MTurk, were paid $9/hour, were at least 18 years old, had 99% or higher approval rate on at least 5,000 previous tasks, resided in the US, and were wearing headphones [312]. 180 participants contributed to the main experiment and 110 to the validation experiment.

10.3 Results

10.3.1 Main experiment

The experiment was terminated after 48 hours, after which 99 out of the 120 chains were full (22 iterations). In Figure 10.3A, we show that Euclidean distance between consecutive iterations within a chain decreases over the course of iterations, stabilizing toward the final 15 iterations. This means that participants move the sliders to a lesser extent at later iterations, suggesting convergence. Since the stills were extracted from audiovisual recordings, we could also compute the speaker embedding of the original reference. In Figure 10.3B, we show that the Euclidean distance to the original reference drops over the first ten dimensions and then mildly increases and decreases again.

10.3.2 Validation

In a separate validation experiment, participants rated how well the voice matches the moving face on a 5-point Mean Opinion Score (MOS): “Excellent”, “Good”, “Fair”, “Poor”, and “Bad match”. The validation included all stimuli generated in the first experiment (overall 2,409 stimuli). Participants performed 200 ratings per experiment and consequently on average every stimulus was rated 9.1 times. As depicted in Figure 10.3, the average Mean Opinion Score (MOS) increases over the course of iterations for all styles. However, the increase is largest for the original faces moving from a 2.7 MOS at iteration 0

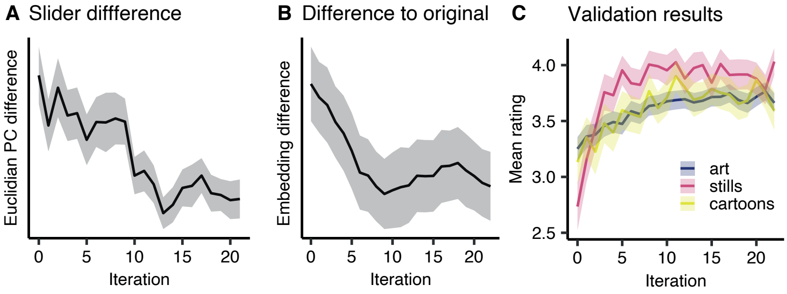


Figure 10.3: Validation **A** The Euclidean distance between consecutive iterations is larger for earlier iterations than later. **B** The Euclidean distance to the original reference drops over the first ten dimensions, increasing and decreasing slightly. **C** The mean opinion score increases throughout iterations.

to a MOS of 4.0 (“Good match”) in the later iterations (Wilcoxon rank sum test, $Z = .42$, $p < 0.001$, Bonferroni-adjusted). The trend is followed by the cartoons (Wilcoxon rank sum test, $Z = .18$, $p < 0.001$, Bonferroni-adjusted). For art portraits the improvement over iterations is smallest (Wilcoxon rank sum test, $Z = .16$, $p < 0.001$, Bonferroni-adjusted). One interpretation of this result is that people have less specific expectations about the voice of a cartoon or art portrait compared to a real face.

10.3.3 Toward personalized voice characteristics

To further understand what kind of voice features were selected by the personalization process, we visualize the speaker latent space using MDS on all voices created in the experiment. As shown in Figure 10.4, over the course of iterations, male and female faces occupy increasingly distinct areas in the voice latent space.

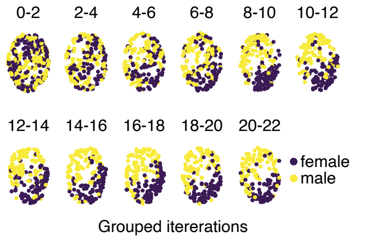


Figure 10.4: MDS on speaker embeddings Generated speaker embeddings in MDS space in which voices for male and female pictures occupy increasingly distinct areas in the voice latent space. Each row represents two iterations, read from left to right, top to bottom.

Furthermore, the average pitch starts at roughly the same point due to the random initialization of the voices and over the course of iterations is lowered for male and increased for female voices (Figure 10.5). Voices for males and females converge in a pitch range common for the sex (85–155 and 165–255 Hz respectively) as indicated by the shaded areas [490].

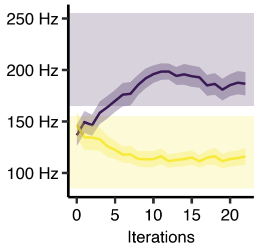


Figure 10.5: Pitch differences The pitch strongly increases for female faces and is lowered for male faces in the first ten iterations. The solid background reflects common pitch ranges for the sexes. Extracted f_0 is expressed in Hz.

Based on these results, we can state that the speaker’s gender apparent from the face is well-recovered in the voice. However, do people only focus on gender or also on other characteristics of the face? In order to address this question, we run another analysis.

Here, we compute the Euclidean difference between the voices created for different styles of the same speaker versus a random speaker of the same sex. Using bootstrapping ($n = 1{,}000$) we show that the voice differences within the same speaker are significantly smaller compared to the voice of a random speaker of the same sex (Figure 10.6), this indicates that the voice prototypes captures face-specific image features in addition to gender.

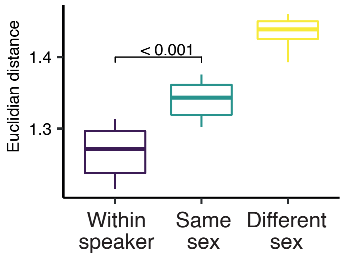


Figure 10.6: Sex differences In the final iterations, the difference in the generated voice is smaller within the different styles of the same face (purple), compared to a random same-sex (green) or different-sex face (yellow).

10.4 Discussion

10.4.1 Summary

In this chapter,

- We used GSP for generating speech personalized to a specific face.
- The generated speech matched the face better over the course of iterations.
- While we found that gender plays a dominant role in the face-voice match rating, the voice prototypes also capture face-specific features.

10.4.2 Limitations and Outlook

- **Room for improving the match**: The ratings plateau at a MOS of 4 (“good match”) indicating that there is still room for improvement. Future research can explore also doing GSP on the prosodic dimensions using the GST embeddings.

- **Vague priors**: Our results have shown that gender plays a dominant role in the face-voice match rating and while the created voices fit better to the face than a random voice, the improvement over random is rather modest. One potential reason for this is that people have rather vague priors about the voice of a face. Future research could try to include more diverse faces, such as faces from different cultures or with different facial expressions, and make the voice model more expressive to capture these differences (e.g. by using the GST embeddings as discussed above).
- **Cross-cultural extension**: This work was done on a dataset of English speakers, involving mainly white actors and involving US participants. Future research could extend this work, by including more diverse faces, and multilingual TTS and deploy the experiment in various cultures.

Chapter 11

RobotVoice: Giving Robots a Voice

Based on

Pol van Rijn, Silvan Mertes, Kathrin Janowski, Katharina Weitz, Nori Jacoby, and Elisabeth André. 2024. 'Giving Robots a Voice: Human-in-the-Loop Voice Creation and Open-Ended Labeling'. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

In this chapter, we explore another voice-related human-computer interaction domain. Again, we use two HITL approaches introduced in this thesis to study another practical problem: How to create a voice for a robot that matches its appearance and predict voices for unseen robots based on their appearance? Robots are used in a wide range of scenarios, and they vary in purpose and appearance [491]. The voice is an intuitive medium for humans to interact with robots, conveying not only spoken content but also intentions [313], personality [492], conversational goals [493], and emotions [119]. However, a discrepancy between what we see (the robot's appearance) and what we hear (its voice) can strongly hinder robots' usability. Previous research has stressed the importance of users' affective responses to robots in fulfilling their functions [494, 495]. However, a mismatched voice can result in a variety of aversive reactions, such as unsettling, eerie, uncanny, and repulsive responses [22–26]. The intensity of this dissonance can be influenced by factors like the user's age or the robot's realism [496, 497].

Robots are used in a wide range of scenarios for varying in purposes [491]. Thus, the appearance of robots has many degrees of freedom – most likely more than faces (Chapter 10). These rich impressions of robots, make it an interesting domain to study the alignment between the auditory (the voice) and visual modality (image of the robot) and thus demonstrate the versatility of the developed methods.

Aligning impressions across different modalities is an important problem to study [494–497] because misaligned impressions can lead to unsettling, eerie, uncanny, and repulsive responses [22–26].

Concretely, we implemented a three-step process to predict the new voices:

- Participants use GSP to create voices for robots by changing the voice of a TTS model and applying robotic effects.
- Participants use STEP to annotate both the impression of the robots in the visual and auditory modality. The top 40 most frequently used tags and those overlapping with literature are selected.
- Participants rate the voices and images along these 40 dimensions to obtain dense associations between the dimensions in both modalities. These associations are then used to predict a voice for a new robot. The voice prediction tool can be found here: `https://robotvoice.s3.amazonaws.com/predict.html`.

11.1 Background

Existing TTS models have frequently been used as voices for robots [498] and their quality has greatly improved in the last decade [334, 499], enabling them to produce speech that is nearly indistinguishable from human recordings [315]. Previous research has shown that humanlike voices are typically preferred over synthetic ones [500], which makes state-of-the-art TTS models an excellent voice creation tool. A recent switch from recurrent to non-autoregressive models [5, 325, 501] brought about major improvements in latency, allowing robots to produce voices faster than real-time. Modern TTS models have great factorization abilities [316, 326, 502], allowing users to independently change text, prosody, and speaker identity (i.e., *what* and *how* something is being said by *whom*). Harnessing such rich latent features [2] not only facilitates the crafting of new voice personae [480, 481] but also ensures that these synthesized voices encapsulate the nuances and diversity inherent to human speech.

A substantial body of extant work emphasizes the importance of synchronizing the robot's voice with its appearance [503, 504]. Simply adopting a TTS model that delivers humanlike speech might be incongruous for a robot that has a distinctly non-human appearance. For example, imagining *R2D2* from Star Wars speaking with a plain natural voice would be odd and likely uncanny [25, 26].

To align the voice with the impression of the robot, previous studies have investigated the correlation between appearance and voice along certain dimensions (e.g., gender or naturalness). For example, McGinn et al. [505] developed a voice association task (i.e., matching a picture of a robot to a voice) showing that gender and naturalness strongly affect the visual appearance people associate with a robot. Other studies have investigated the opposite relationship: How the voice influences the mental model that people have of a robot. For example, Powers et al. [506] showed that participants associate a male voice with a more knowledgeable person. This research also highlights the risks of reinforcing existing social biases when matching specific vocal characteristics, like a deep voice, with particular personality traits, such as being knowledgeable. In addition to aligning the voice and appearance of a robot, its behavior must also be synchronized. Torre et al. [507] show that while trust partly depends on the voice, the consistency of voice and behavior is more important [508].

This is in line with previous research showing that people prefer serious-sounding robots in work-related contexts [509] and empathetic voices for healthcare robots [510].

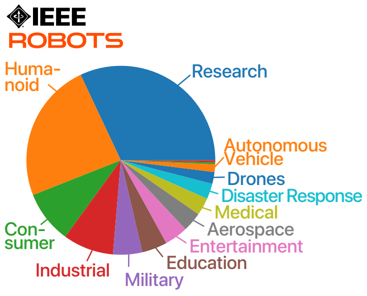


Figure 11.1: IEEE robots Pie chart of distribution of robot categories with an example from each category.

11.2 Methods

11.2.1 Images of Robots

As robots vary greatly in their appearance, the goal was to collect a variety of images that capture this variation. To simplify the complexity of possible presentation methods (such as images, videos, and 3D designs), we focus on static images. We used an existing dataset (IEEE Robots) and downloaded all robots from `https://robots.ieee.org/robots` (April 2022), removing robots without a frontal view and discarded devices such as exoskeletons or telepresence interfaces, which integrate a human user. For each robot, we selected the best image, ideally showing the entire robot in isolation. The selected images span diverse types of robots with 14 different categories, ranging from industrial to consumer robots and humanoids to drones (see Figure 11.1).

This list of 160 IEEE Robots was extended with 15 images that were collected from other sources, such as promotional pictures from manufacturer websites or photographs taken by ourselves. To avoid contextual cues, shades and backgrounds were replaced by a solid white background. This selection of 175 robots is notably larger than datasets in relevant previous literature with maximally eight different robots [503–506, 508, 510–514].

11.2.2 Voice Manipulation and Effects

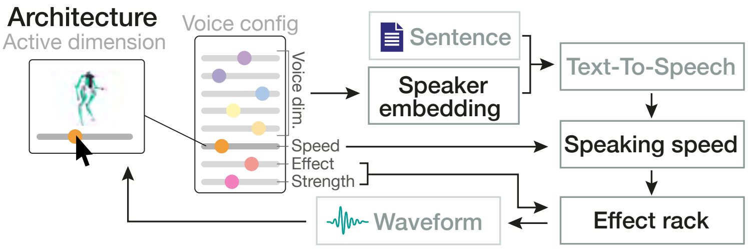


Figure 11.2: Architecture The voice of the robot is controlled via eight sliders. The first five sliders control the voice of the TTS model using the first five PCA dimensions on the speaker embeddings. The sixth slider controls the speed of the speech. The seventh slider selects one of the eight effects. The last slider determines the strength of the effect. When moving the slider, the voice configuration updates one parameter in the voice configuration (here: speed). This triggers the synthesis pipeline and the resulting audio is played back to the user. Robot: Digit by Agility Robotics (with permission).

To create a voice for a robot, one needs an expressive voice creation tool that is fully parametrizable. The proposed solution is depicted in Figure 11.2.

Overall, the architecture changes the voice of a TTS model, changes the speaking speed, and passes the resulting audio to a rack of effects. Participants use sliders to adjust the model parameters, thus changing the voice.

The first five sliders modify the voice of the speaker of a TTS model. We modified the state-of-the-art TTS model “VITS” [315] trained on the VCTK dataset [515] so that it can be used to directly modify the voice representation (speaker embedding). We performed a Principal Component Analysis (PCA) on all 110 speaker embeddings of the same dataset and used the first five PCA components, which seem to capture sufficient variation in the human voice (different speaking speed, voice timbre, and speaker sex). We perform reverse PCA to obtain a speaker embedding based on the PCA dimensions. For maximum expressivity and minimum distortion, the range is constrained to approximately four standard deviations in all dimensions.

Since the variability in speaking speed in natural human speech is rather limited and the PCA dimensions by themselves did not provide enough variability in terms of duration, we added a sixth slider that can parametrically change the speaking speed ranging from 46% to 153% of the original speed using Parselmouth [292], a Python wrapper for Praat [133].

The TTS system was trained on natural speech [515]. This means that the TTS model mainly produces naturalistic human voices and does not create robotic-like sounds. Therefore, we added sliders to apply robotic audio effects, by combining modern TTS with traditional signal processing techniques. The eight different effects implement commonly used to create robotic voices:

changing the pitch, decreasing synthesis quality, applying a timeshift, using a vocoder, or applying one of four different flanger configurations to the audio. The effects are applied in a sequential using an effect rack [516]. To avoid a strong mixture of voice effects, participants used a seventh slider to pick one of the eight effects and used an eighth slider to adjust the strength of the effect. The overall amplitude of the effects was manually normalized such that each effect would be approximately equally salient. The slider positions are linearly spaced (with a resolution of 16 positions) to make the synthesis computationally efficient. The following types of effects were implemented:

- **Pitch**: The signal was enhanced with two transposed audio tracks, where one was transposed five semitones up and the other transposed five semitones down. By doing so, the intonation pattern of the voice gets obscured, resulting in an unnatural voice. Further, both transposed signals are a minor seventh apart, which is generally considered a rather dissonant interval in Western music perception [517]. As such, additional tension in the voice is induced. The corresponding slider in our experiments allows us to control the ratio between the non-transposed and transposed signals.
- **Synthesis quality**: Older text-to-speech systems are poor at phase reconstruction, which results in audible artifacts that sound "robotic". To emulate this poor reconstruction, the signal is transformed into the frequency domain using a short-time Fourier Transform and then reconstructed using an inverse short-time Fourier Transform but with randomly initialized phase estimates. Our implementation utilized Librosa's Griffin-Lim algorithm [516] without executing phase approximation.
- **Timeshift**: To facilitate the creation of more "fuzzy" sounds, the original voice can be blended with slightly time-shifted version. By doing so, the warmth and resonance of a natural voice gets veiled. To obtain this effect, the original signal was delayed for a few milliseconds, and the time-shifted signal was combined with the original signal.
- **Vocoder**: Vocoder effects are frequently used to create robotic voices [518]. Here, the speech signal is used as a modulator for a carrier signal. By fixing that carrier signal to a certain frequency, the resulting voice sounds monotone and mechanical. The pipeline makes use of TAL Vocoder [519], a publicly available VST implementation, which is included into the codebase using Pedalboard [520].
- **Flanger**: Flanger is an audio effect that imparts a more synthetic quality to the sound. The flanger effect is achieved by combining a signal with a delayed version of itself where the delay time is modulated by a low-frequency oscillator. This addition offers an avenue to offset the voice's natural tone. Four different flanger configurations were implemented, each with a distinct modulation frequency and depth leading to unique auditory experiences.

Each robot was randomly assigned to a sentence from the 720 phonetically balanced and semantically neutral Harvard sentences [310].

11.2.3 Robot attributes proposed in the literature

To obtain a long list of attributes proposed for robots in literature, labels from the "Big Five" personality model [521–524], the Godspeed questionnaires [525] (focussing on social robots), AttrakDiff questionnaire [526] (focusing on user experience) and relevant dimensions for voice assistants [527] were included.

This list was extended by three adjectives signifying demographic features, namely “young”, “male” and “female” to specify age and gender and the word “animallike” because our collection of robots contains many artificial pets and animal-inspired robots. This yielded 260 unique attributes (see supplementary information of the paper for the full list [15]). While this list clearly does not capture all possible attributes ever mentioned for robots, it covers the most widely used attributes in the literature.

11.2.4 Participants and experiments

Overall, 2,505 participants took part in all experiments. Participants were recruited from Prolific, provided informed consent, were at least 18 years old, resided and were living in the UK, had to speak English as their first language, and had to have been raised monolingually. Participants earned at least 9 pounds per hour. If audio was played in the experiment, participants went through an automatic headphone check asserting they were wearing headphones [312]. To ensure that participants were highly proficient in English, they had to pass WikiVocab (Chapter 8).

11.3 Results

11.3.1 Voice Creation using GSP

803 participants engaged in the GSP experiment. In the experiment, participant control one of the eight dimensions of the voice of a robot and change it to best match the voice with the robot’s appearance. Each participant visits 20 different robots. The same slider is shown to five participants, and the median of their responses is carried forward to the subsequent iteration. Initially, all vocal parameters were uniformly randomized with the possible range values (see Methods, Section 11.2.2). The experiment concluded after 48 hours, during which time 70 images underwent 15 iterations, and 105 images experienced 16 iterations. Consequently, each of the eight dimensions was visited approximately twice.

Figure 11.3 shows that the standardized slider difference between consecutive iterations within a chain decreases over the course of iterations. This means that participants move the sliders to a lesser extent at later iterations, indicating convergence. In particular, there was a significant difference between the first and last iteration (Wilcoxon signed rank test: $V = 11277.0, n = 175, p < 0.001, r = .43$, this and all future tests are Bonferroni corrected for multiple comparisons) but we did not find a significant difference between the last iterations to the six iterations preceding it. The slider difference drops after all eight dimensions have been visited once, which is in line with previous studies [1, 3, 4] (see Chapter 4 and 10). The development over iterations can be listened to online: https://robotvoice.s3.amazonaws.com/iterations.html.

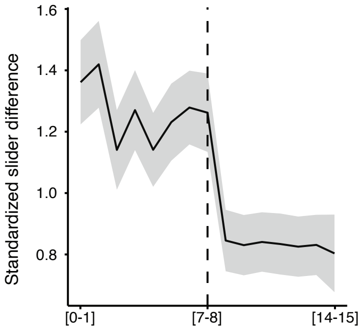


Figure 11.3: Successive slider difference Standardized difference between successive slider configurations. Shaded areas are confidence intervals.

To visualize the proximity of the matched robot voices to each other, we performed a PCA on the standardized slider positions of all stimuli in the experiment. Figure 11.4 depicts the first two principal components and shows the distribution of all slider configurations using a KDE (gray lines). The initial robot voice configurations are uniformly sampled from the sliders but occupy

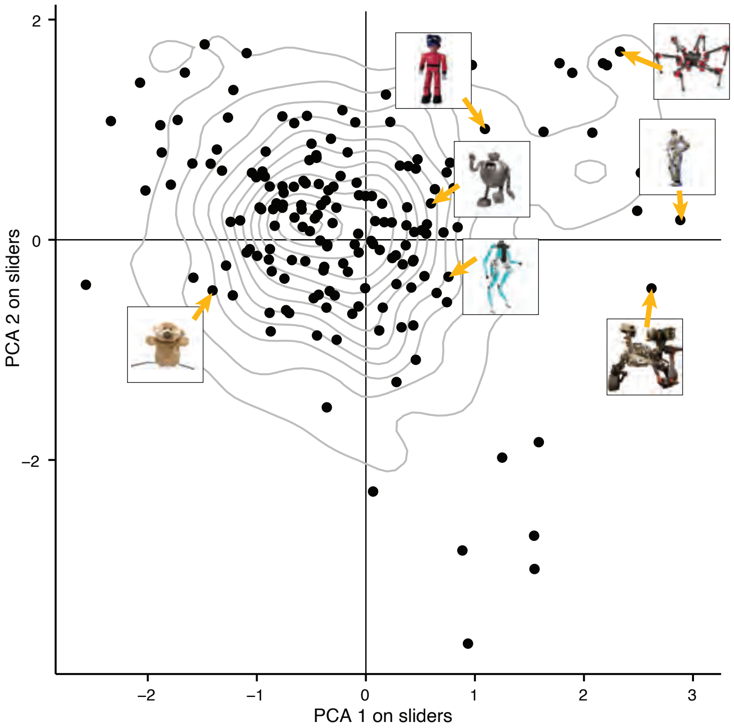


Figure 11.4: PCA on slider configurations PCA on all slider configurations from all iterations. The gray KDE indicates the distribution of all slider configurations in PCA space. The black points are the final slider configurations. Robot images are used with permission or are in the public domain.

distinct slider positions at the end of the experiment. For example, the spider-like robots in the upper right corner or the toy-like robots in the top left corner of the plot group together in slider space (i.e., they received similar voices in the final iteration). The final voices can be explored interactively using the online visualization: `https://robotvoice.s3.amazonaws.com/explore.html`.

In order to validate whether the voice and robot match improves over time, we recruited a separate group of participants (N = 142) that rated how well the voice matches the robot.

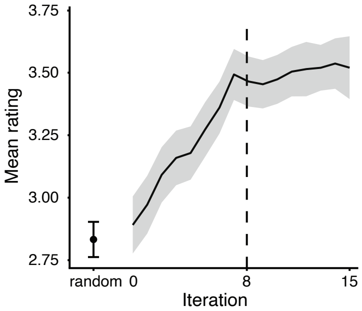


Figure 11.5: Results validation experiment Mean ratings as a function of the iterations and a random voice. Shaded areas are confidence intervals.

This experiment comprised 2,730 stimuli. All stimuli were generated in the GSP process with three additional random voices per robot. There were about 4.9 average ratings per stimulus. Totalling, 13,597 human judgments in this experiment. As depicted in Figure 11.5, the average match increases over the course of iterations. In particular, the average of the last three iterations was significantly larger than the first three iterations (Wilcoxon signed rank test: V = 1813.5, n = 175, $p < 0.001$, r = .64). In addition, the increase in rating over iterations reduces after each dimension is visited approximately once (again in line with previous GSP studies, see Chapter 4 and 10). For example, we did not find a significant difference between the average of iterations 8–10 and the average of iterations 13–15 (Wilcoxon signed rank test: V = 6321.5, n = 175, p = 0.018, r = .10).

11.3.2 Annotation using STEP

To obtain a list of attributes that are used to describe the robots in the auditory and visual modality, we used STEP. We recruited two new groups of participants to annotate the obtained final robot voices and the original images (N = 59 and N = 73 respectively). Each robot is sequentially annotated by 10 participants. To facilitate convergence and avoid spelling variants and duplicate tags, participants can see words that start with the same letters while typing and can select them if they find them appropriate. The proposed words are either tags provided by other participants or the 260 dimensions proposed in the aforementioned literature (see Methods, Section 11.2.3).

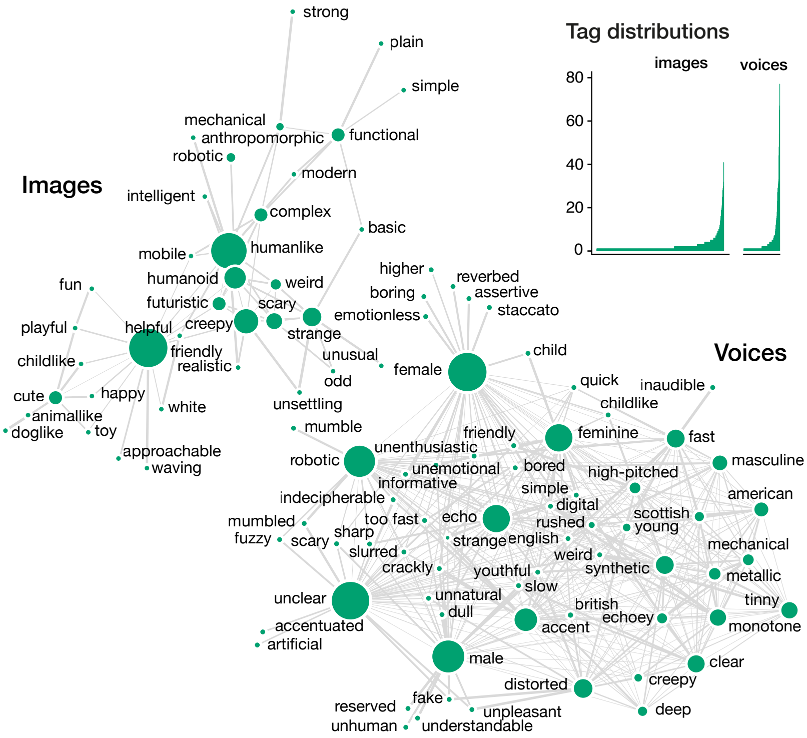


Figure 11.6: STEP results Co-occurrence networks between provided tags per modality. Tags with a co-occurrence below 4 are pruned to remove words that are rarely used. The size of the nodes indicates the degree. Networks are created using Gephi [528]. Tag distributions are the raw occurrences of single labels for the 175 images and 175 voices.

As depicted in the two histograms in Figure 11.6, the vocabulary used to describe the 175 images of robots is generally larger than those of 175 voices (765 and 217 unique tags for the image and voice modalities, respectively). Also, the same labels are used more frequently for the voice compared to the image modality (mean occurrence of 5.4 and 2.5 for the voice and image modalities, respectively).

To investigate which terms are particularly relevant, we visualize the co-occurrence network for each modality in Figure 11.6 using a network analysis [7]. The nodes are tags proposed in the STEP process and the edges indicate if they co-occur within the same robot with other tags. Those tags that have many connections to other tags – indicated by the larger dots – are likely to be relevant descriptors. In the co-occurrence network, terms that are semantically similar are often located near each other, such as “animallike” and “doglike” in the image modality. However, this isn’t always the case, as terms that are semantically related do not necessarily appear together if they are inapplicable to a significant number of robots.

Overall, we observed that tags in the voice modality are more interconnected (average degree: image = 3.8, voice = 11.3), suggesting that a relatively small number of recurring labels frequently appear together. This observation aligns with what is shown in the histograms in Figure 11.6. This pattern can be partially attributed to the challenge of identifying vocal properties compared to image attributes. Voice representations might be less easily described in words, or more ambiguous overall, leading to greater overlap in semantic labels.

Interestingly, while our approach is open-ended (e.g., we don’t use post-

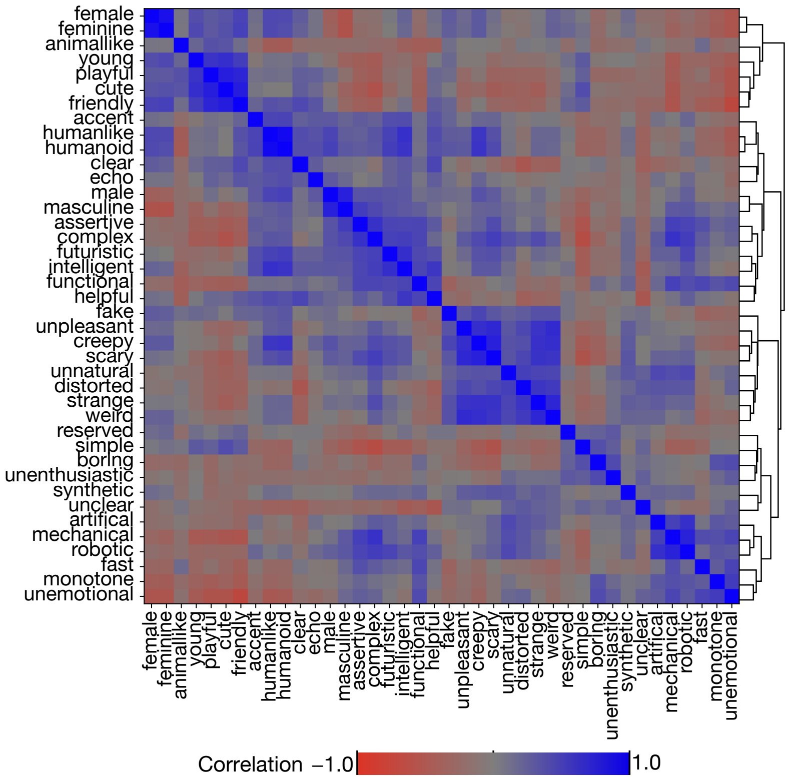


Figure 11.7: Correlations across dimensions for images Correlation across dimensions for images. Correlation matrices are sorted by the order in the dendrogram obtained via agglomerative clustering.

processing and involve lay participants), many central terms overlap with those commonly mentioned in literature such as "friendly", "humanlike" or "female" (see, for example, [529] or [506]). Figure 11.6 furthermore reveals that while some impressions are modality-specific (e.g., "high-pitched", "echo", "accent"), the majority of terms proposed by the participants reflect general impressions of the robot (e.g., "weird", "cute", "robotic" and "friendly") and are not modality-specific. However, other features differ across the two modalities. For example, the biological sex or the age of the speaker is an important category in voices, whereas the distinction between "animallike", "humanlike", and "robotic"/"mechanical" seems more important in the case of images. Furthermore, the participants came up with terms for the voices that refer to communication qualities, such as "inaudible" or "informative", and the communication style, such as "assertive" and "unenthusiastic". Obviously, the participants were able to produce voices that complemented the visual impression of the robots by assigning additional attributes to them via the voice modality. The observation that participants used terms related to communicating qualities and styles when judging voices, but not when viewing static images of robots, highlights the complementary of different sensory modalities, such as visual and auditory.

11.3.3 Dense Rating along Perceptual Dimensions

While STEP also involves rating the relevance of the tags, this rating data is fairly sparse: It is only available for the tags that were proposed by earlier participants and the later a tag was proposed, the fewer ratings it received. To get a denser representation, we compiled a list of 40 relevant dimensions that were rated by a new group of participants for the images and voices in separate experiments.[1]

[1] I limited the number of dimensions to 40 to ensure that each dimension was rated by enough participants.

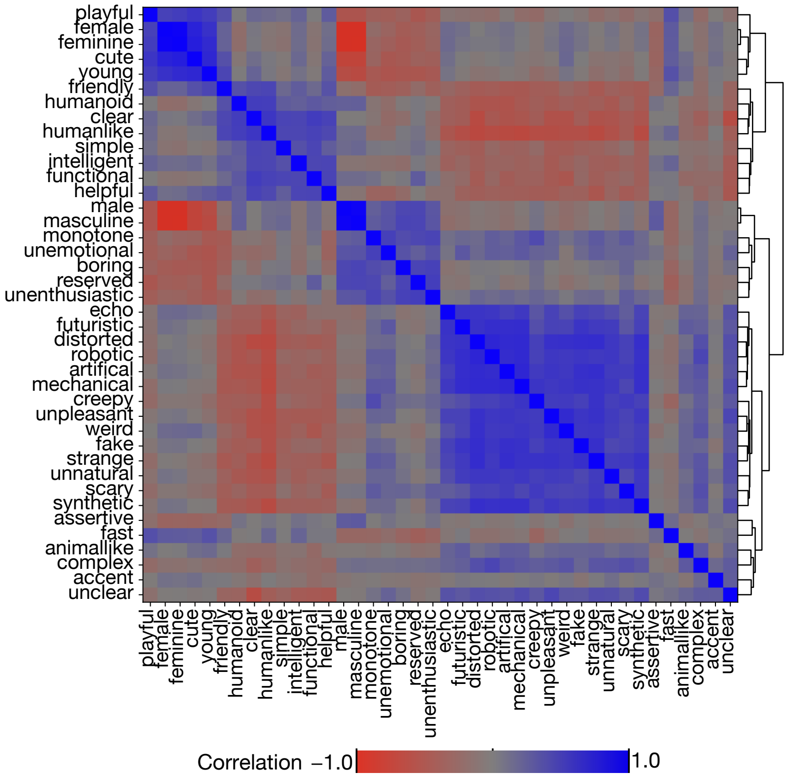


Figure 11.8: Correlations across dimensions for voices Same plot as in Figure 11.7 but for voices.

To compile the list of attributes, we selected the 26 dimensions that overlap between the list of 260 labels from previous literature. The remaining 14 dimensions are the 7 perceptually most salient features (based on STEP) in each of the modalities.

We recruited separate groups of participants to rate these 40 perceptual labels in the image and voice modalities ($N = 298$ and $N = 245$ respectively). Each participant rated the robot image or robot voice on 5 randomly selected dimensions using sliders that snap to 5 positions. On average, each stimulus and dimension was rated 7.5 times for the images and 6.1 times for the voices. Overall, the ratings were reliable for both experiments: the split-half reliability for images was $r = 0.65$ and $r = 0.61$ for the voices. To compare the consistency of the dense rating results with the STEP results in the previous experiments, we correlated STEP ratings with the dense ratings for the labels that occur in both datasets. We found that the mean rating was correlated with the mean number of stars a label received in STEP ($r = .31$ and $r = .24$ for images and voices respectively).

Figure 11.7 shows the correlations between the dimensions for the image modality (i.e., a correlation between average rating per stimulus between all dimensions). Generally, terms with similar meanings, such as "female" and "feminine", show strong positive correlations, while antonyms like "clear" and "unclear" display strong negative correlations. The matrix reveals an additional pattern: participants tend to associate female robots with labels like "young", "playful", "cute", and "friendly", while male robots are linked with traits such as "assertive", "functional", "complex", and "intelligent". These observations align with previous literature [506], which suggests that societal stereotypes influence how robots are perceived.

For the voice, the correlation matrix shows a more consistent structure (Figure 11.8): The largest cluster contains dimensions like "creepy", "unpleasant",

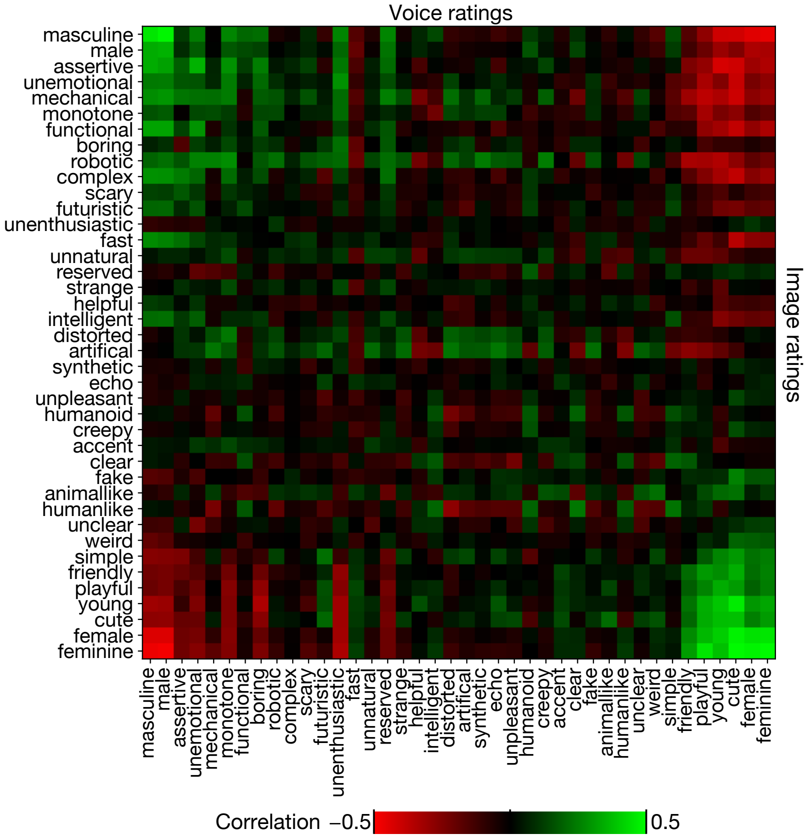


Figure 11.9: Correlations between dimensions across modalities Correlation across both modalities. The correlation matrix is sorted by mean correlation for the most consistently rated dimension "feminine".

"mechanical", and "robotic". Also, "female" is associated with a "young" and "cute" voice (consistent with previous literature) [530], but not with a "friendly" voice. Instead, a new cluster emerges for "friendly", "helpful", "clear", and "intelligent" voices. This suggests that voice modality presents a much harder challenge in terms of providing labels. Specifically, voices cluster to a smaller number of interconnected terms (consistent also with the usage of smaller vocabulary in Figure 11.6).

To investigate the robustness of our findings, we ran the dense rating experiment on 175 new images from the ABOT dataset [531] and on 175 randomly created voices using the voice tool. We found strong correlations between the two image (r = .85) and two voice datasets (r = .91). These findings indicate that the obtained correlations across the terms are robust across datasets.

We also investigated the correlations of the dimensions across the modalities. Generally, the correlations were lower, indicating that the association between the dimensions across the modalities is weaker (e.g., a masculine robot does not necessarily become a male-sounding voice, see legend of Figure 11.9). Furthermore, as depicted in Figure 11.9, the diagonals between the dimensions were much weaker or entirely vanished for certain dimensions for example for terms like "humanoid" or "unpleasant". This indicates that the same labels are not consistently used across modalities, for example, a "fast" voice does not mean that the image of the robot looks "fast" too. The dimensions that are best preserved across modalities are dimensions like "feminine", "young", and "cute" (Figure 11.9).

In Figure 11.9, one can see that there is a large overlap between associations from images to voices as well as from voices to images (e.g., male voices are associated with mechanical robots, and vice versa). However, this relationship is not always bidirectional; for example, assertive robots are associated with male voices ($r = 0.32$), but male robots are not really associated with assertive voices ($r = 0.09$). Further comparisons between the modalities can be made via the interactive visualization: https://robotvoice.s3.amazonaws.com/compare.html.

In the voice modality, the first principal component primarily captures the contrast between "humanlike" and "robotic", while the second dimension focuses on the male-female dichotomy. In the image modality, a similar contrasting pattern is observed between "humanlike" and "robotic" features, but here the emphasis is on terms related to automaticity, such as "fast," "monotone," and "unemotional," as opposed to terms like "playful," "friendly," and "cute". The gender dichotomy is somewhat less pronounced here.

11.3.4 Predict Voices based on Labels

To use the created voices in real applications, we test whether one can predict the voice of a robot based on the image ratings. So if we can predict a suitable voice for a new robot based on its appearance (e.g., find a voice for a "cute" and "female" looking robot)? To address this question, we recruited a new group of participants ($N = 94$) and presented them with different combinations of images and voices (Figure 11.10). For each robot image i, we provide five different combinations of an image and a voice: As ground truth, we included the original matched voice of robot i (*matched*). To see how well verbal descriptors of the image predict the voice, we searched for the robot i with perceptual image rating across the 40 dimensions and found the closest robot j (*closest*, i.e., with the highest cosine similarity, e.g., the Perseverance robot is closest to the Spirit & Opportunity robot). We then used the voice of the j in the final iteration of the GSP experiment. To test robustness, we also searched for robot j for the GSP slider configuration and selected the closest voice in slider space, which did not occur in any iteration for robot i and j (*selected*). As a negative reference, based on the slider configuration of the matched robot i we searched for the worst slider combination (*worst*, i.e., which is maximally dissimilar in cosine similarity). Finally, we also included a random voice configuration (*random*). The interface of the prediction experiment was identical to the GSP validation experiment. We had 875 stimuli and 7,444 human judgments overall, and each stimulus received an average of 8.5 ratings.

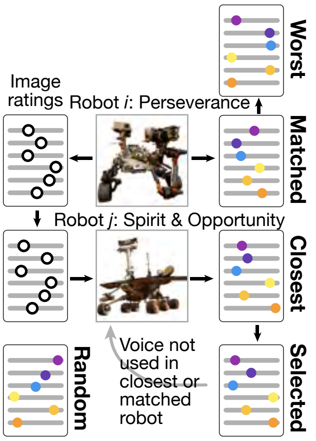

Worst
Image ratings
Robot i: Perseverance
Matched
Robot j: Spirit & Opportunity
Closest
Random
Voice not used in closest or matched robot
Selected

Figure 11.10: Prediction schematics Schematics of the procedure to select stimuli for the prediction validation experiment. Images are taken by JPL/NASA and are in the public domain.

Consistent with the validation of the GSP voices, the random voice received the lowest voice match score and the final voice the highest match score (Figure 11.11, left panel). While the closest and selected voices received a slightly lower match rating, we did not find a significant difference there (Wilcoxon signed rank test: $V = 6879.0$, $n = 175$, $p = 0.47$, $r = .07$). However, the matched, closest, and selected voices were all significantly better than the worst or random voices ($p < 0.001$ in all cases), which both have much lower ratings. Thus, this shows that while the predicted voices (closest and selected) were all better matches than a random voice, they were not significantly worse than the matched voice. This trend is not only visible when averaging over all participants, but also on a single-participant basis (see Supplementary Materials of the paper [15]).

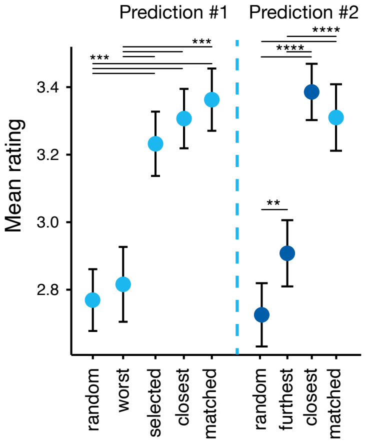


Figure 11.11: Prediction results Results of the prediction validation experiment. Matched receives the highest score followed by the closest and selected voice configuration. The worst and random voices receive the lowest scores and are significantly lower than the matched, closest, and selected voice configurations. *** indicates that the paired Wilcoxon signed rank test was significant ($p < 0.001$).

To assess if the findings also extrapolate to other datasets of robots, we run another prediction experiment ($N = 73$). We wonder if the annotated features of the new robot can be used to match the voice based on the old data set's annotated features. In a real-world scenario where an engineer might have a new, unseen robot image and want to use our results for voice matching, this validation is crucial as it should show that even when using voices tailored to the old dataset and a matching model trained solely on the old dataset, one can still achieve accurate predictions with a new set of independently annotated images. Thus, we looked up the closest robot in terms of its annotated features for each of the new 175 robots in the old set of matched robots (*closest*). As a reference, we also included the same matched voice and paired it with the directly matched old robot image (*matched*). As a negative reference, we looked up the perceptually furthest robot in the old dataset and selected its voice (*furthest*). We also add a random voice (*random*). As shown in Figure 11.11 (right panel), the closest and matched voices are all significantly better than the furthest and random voice ($p < 0.001$ in all cases). While the closest voice received a slightly higher rating than the matched voice, this difference was not significant (Wilcoxon signed rank test: $V = 8215.5$, $n = 175$, $p = 0.14$, $r = .11$). As in the previous prediction experiment, the furthest matched voice was slightly higher than random though both of them had low ratings overall. This is probably because random voices are uniformly sampled along the dimensions, leading in some cases to sample extreme values, which is not the case for the furthest or worst voices that were matched to a robot. This additional prediction experiment shows that the prediction also works for newly annotated robots from different datasets.

To try out the prediction tool, we provide an interactive voice prediction tool online: `https://robotvoice.s3.amazonaws.com/predict.html`.

11.4 Discussion

11.4.1 Summary

Concretely,

- We provided a voice creation tool that covers a wide range of robotic voices using both state-of-the-art TTS (extending upon previous work, Chapter 10) and robotic effects using signal processing.
- We used GSP to create a matched, synthetic voice for 175 robots.
- We employed STEP to identify labels that are relevant for the perception of robots, both in audition and vision, and compared them with attributes from the literature.
- We selected the most used attributes and participants annotated dataset of 175 robots along those attributes.
- We show how those perceptual ratings can be used to propose suitable voices for new robots.

11.4.2 Limitations and Outlook

- **Dynamic materials**: The study focused on the voice channel and used static images to control for voice manipulation. The way robots move can significantly affect human perception, and a wide range of literature illustrates how robots convey personality through body language, gestures, and facial expressions, as summarized in [495]. Future research can investigate how different use cases and scenarios of the same robot can affect the perceived appropriate voices.

- **Longer textual content**: The voices we used were matched with short, semantically neutral sentences, which might not generalize to longer textual content. Future research could investigate the impact of longer, semantically relevant spoken content on the perception of robots.
- **Contextual biases**: While we purposefully selected a neutral background for the robot to minimize contextual biases, it is essential to recognize that participants may have held varying perceptions of the robot's role, task, and target audience while adjusting voice dimensions. Empirical evidence indicates that factors beyond the robot's attributes, such as the task and user characteristics, significantly influence how it is perceived and how humans interact with it [513]. The transition from a toy-like robot to a robot serving as a speech assistant, as highlighted by Aylett [511] using the example of the Cosmo robot, can result in a mismatch between the robot's function and its voice. Tags used by participants to describe the robot, such as "functional" and "helpful", highlight the importance of its intended purposes and audience in addition to its audio-visual characteristics. Both the robot's visual appearance and the mental models it triggered in participants may have influenced voice modifications. Future research should gain further insights into these factors, by conducting additional experiments with systematic changes to the robot's visual context aligned with its intended functions.
- **Generalization**: We based the voice creation tools on an English dataset, which, while diverse in including multiple dialects, did not allow us to explore the intricate relationship between culture and robot perception. This limitation applies to the user's cultural background and the culture the robot is intended to portray. Prior research [512] demonstrated that a robot's social category membership, including culture, significantly influences how people perceive and interact with it. During the annotation process, the participants included tags related to English dialects like "Scottish" and "American", highlighting the relevance of group membership as a distinguishing characteristic. McGinn and Torre [512] manipulated the accent of a robot's voice to investigate its impact on the formation of stereotypes. However, due to the heterogeneous background of the participants, their findings on the effect of accent manipulation were not consistent. Further research should, therefore, focus on the alleged cultural background of robots portrayed by their accent. In a broader perspective, the study only involved monolingual English UK participants and future research should incorporate less "WEIRD" participants (Western, Educated, Industrial, Rich, Democratic) [246, 464] to uncover associations across perceptual dimensions in different cultures. Our approach is largely language-agnostic (it would solely require a TTS model trained on a different language) and thus can be applied to a variety of languages and cultures.
- **Expressiveness**: While the matched voices are significantly better than a random voice (Figure 11.11), the mean ratings for the matched voices (3.4) are not quite at ceiling performance (5.0). One explanation is that the voice model is not expressive enough yet. The voice dimensions mainly capture aspects of the voice such as gender or sex. Future research can improve the parametrization of the latent voice dimensions to capture more expressive features of the voice.
- **Individual differences**: The split-half reliabilities in the rating experi-

ments are high but there is still some disagreement across participants (dense: image *r* = .68 and .53, voice *r* = .65 and .48; prediction: *r* = .56 and .53). This indicates that future research should investigate individual differences in the perception of robots.

Chapter 12

General discussion

In this thesis, we have identified three core problems in constructing corpora for emotional prosody (Chapter 3) and have developed three Human-In-The-Loop (HITL) algorithms that provide solutions to them (Chapter 4–6), namely: the stimulus selection, taxonomy curation, and lost-in-translation problem.

12.1 Contributions

The first HITL algorithm – Gibbs Sampling with People (GSP) – is a domain-agnostic sampling technique involving humans (Chapter 4). GSP allows us to identify which parts of the stimulus space in a generative model (such as a TTS model changing prosody) are associated with particular semantic concepts (e.g., which prosodies are associated with “anger”). We have applied GSP to the domain of emotional prosody using two generative models: a parametric speech manipulation (e.g., changing the mean pitch of a voice recording) and a TTS model controlling prosody. In both projects, we have shown that GSP allows us to create prosodies that are reliably associated with a particular emotion. For the TTS model, we also show that emotional prosody optimized for a particular sentence can be transferred to other sentences without changing their emotional content. Since minimal constraints are posed on the stimulus space by the used generative models (e.g., the acoustic manipulations are not exclusively used in emotion communication or the TTS model is not trained on emotional speech), GSP allows to sample the full stimulus space in a representative manner, providing a solution to the stimulus selection problem.

We extend these ideas in two further projects also in the voice domain: VoiceMe and RobotVoice. In VoiceMe, participants develop a voice that matches their impression of a face by iteratively changing speaker embeddings in a TTS model (Chapter 10). The project shows that GSP is not limited to optimizing a stimulus for a particular label (such as “angry” or “beautiful”), but the paradigm can also be used to study associations across modalities. So, which voice matches the impression of a face. In RobotVoice, this idea is extended to creating voices for robots, in which participants change the voice, including robotic sound effects, to match it to an image of a robot (Chapter 11). The novelty of RobotVoice is that we use a human annotation pipeline (STEP; Chapter 6) to obtain dimensions along which the robot images and created robot voices differ and that we show that the perceptually meaningful dimensions obtained from this process (such as “cute” or “mechanic”) allow predicting matching voices for unseen robots without having to repeat the GSP process again.

The second HITL algorithm – Genetic Algorithm with People (GAP) – also provides a solution to the stimulus selection problem (Chapter 5). Most corpora of emotional prosody rely on emotion taxonomies to create the corpus (e.g., ask people to find speech recordings that sound “angry” or ask actors to say a particular sentence in an angry way). This is problematic because the stimulus selection is constrained by the selected emotions and thus leads to a constrained sample of the stimulus space. GAP overcomes this problem by letting creators imitate previous creations (without knowing the experiment is about emotions) and raters who have to select the most emotional creation. Over the course of iterations, this leads to increasingly emotional prosodies of the same sentence. Unlike existing corpora, GAP does not rely on any pre-assumed taxonomy and thus leads to a more representative sample of emotional prosody.

The third HITL algorithm – Sequential Transmission Evaluation Pipeline (STEP) – shows participants a stimulus and asks them to provide tags describing their impression and rate the tags of others (Chapter 6). Over the course of iterations, this leads to a weighted bag-of-words representation of all stimuli and can be used to distill a taxonomy for the stimulus set. Thus, STEP solves the taxonomy curation problem, which refers to the problem of superimposing an existing taxonomy on new data without assessing if the taxonomy fits the data.

STEP can also be used to solve the lost-in-translation problem, which means that the same objects are named differently across languages, which raises the question of how concepts in different languages relate to each other. For example, does the German word "Wut" mean the same thing as "anger" in English, or what about language-specific emotions like "Schadenfreude"? To make this possible, I have co-developed a Python package called PsyNet (Chapter 7) that is used to implement all three HITL algorithms and automates the process of running massive online experiments across the globe. To make sure our participants are fluent in the language they are tested in, I have developed a language proficiency test in Chapter 8. Finally, in Chapter 9, we test this global infrastructure by conducting a large-scale, cross-cultural color naming study, which is a well-studied instance of grounded semantics [84–98].

Based on this summary, the contributions of this thesis can be summarized as follows:

- We have identified three core methodological problems in emotional prosody (stimulus selection, taxonomy curation, and lost-in-translation problem).
- We have developed three HITL algorithms that provide solutions to those problems.
- We have established and validated an infrastructure to run massive online experiments across the globe to deploy HITL algorithms at a large scale.
- We have shown that while those HITL algorithms have been developed to solve core scientific problems, they can also be used in practical applications.

In the next section, we will show that while the HITL algorithms have been developed to solve problems in emotional prosody, they can also be applied to other domains and modalities.

12.2 Generalization of the problem

To generalize the problem to other domains and modalities, we formalize it as follows (see Figure 12.1): An emotional speech recording is a stimulus x in stimulus space $\mathbf{X}$, and the emotion associated with the recording is a verbal description y in semantic space $\mathbf{Y}$. The relationship between the two spaces can be studied in two directions. One can ask, for a given stimulus, which

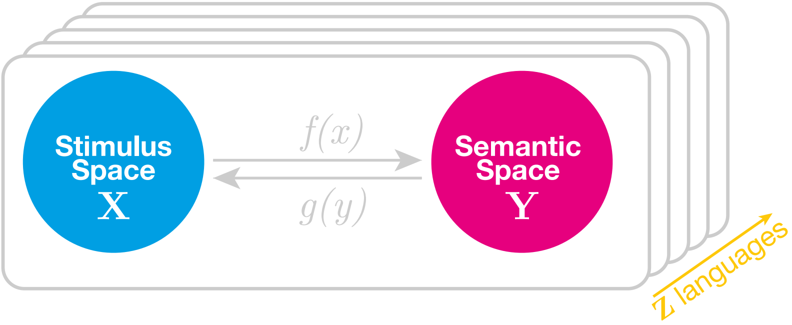


Figure 12.1: Formalization Assume the existence of a mapping f from semantic space $\mathbf{Y}$ to stimulus space $\mathbf{X}$ and vice versa (mapping g). The mappings can be studied across languages ($\mathbf{Z}$).

verbal descriptions are associated with it ($f : \mathbf{X} \rightarrow \mathbf{Y}$), or for a given verbal description, which stimuli are associated with it ($g : \mathbf{Y} \rightarrow \mathbf{X}$). The same stimuli ($\mathbf{X}$) can be presented to participants speaking different languages ($\mathbf{Z}$), and the obtained semantic spaces ($\mathbf{Y}$) can be compared.

This conceptualization allows us to think about the developed HITL algorithms in a more general framework: GAP allows us to create a stimulus space $\mathbf{X}$ without pre-assuming a particular taxonomy. STEP allows us to annotate the stimulus space and obtain a semantic space $\mathbf{Y}$, essentially implementing $f(x)$. Finally, GSP investigates the opposite mapping, $g(y)$, namely finding a stimulus x which best resembles a given verbal description y.

From this conceptualization, it becomes apparent that the three problems we identified provide contributions to different aspects of this formalization:

- Solving the **stimulus selection problem** provides a contribution with respect to the stimulus space $\mathbf{X}$.
- Solving the **taxonomy curation problem** provides a contribution to the semantic space $\mathbf{Y}$.
- Solving the **lost-in-translation problem** provides a contribution to studying the mapping across languages $\mathbf{Z}$.

This formalization also allows us to connect to a larger set of problems, both in computer and cognitive science.

In computer science, many corpora suffer from the same problems we identified for emotional prosody, including corpora for object [336], scene [340], sound [339], video [338], and facial expression recognition [532]. For example, the domain of object recognition suffers from the:

- **Stimulus selection problem**: Popular corpora for object recognition, like ImageNet [336], only contain a particular set of everyday objects, which might look different depending on the location in the world or might not even exist [533]. So, many object recognition corpora are an unrepresentative sample of the full stimulus space of all objects.
- **Taxonomy curation problem**: The objects in ImageNet are organized in a prespecified, particular taxonomy, which might not be an optimal fit to the data [329], so it suffers from the taxonomy curation problem.
- **Lost-in-translation problem**: The objects in ImageNet are annotated in English, but the same concepts of objects might not exist in other languages, are divided into different subcategories, or only exist as an abstract category [534]. The most famous example of this phenomenon is probably the larger number of words for “snow” in Inuit languages compared to Western languages [535, 536], showing that languages differ

in the granularity of how they describe the same concept. Thus, the lost-in-translation problem is also present in corpora of object recognition.

The formalization also fits naturally to describe a well-known phenomenon in cognitive science (called "grounded semantics"), describing that while all participants observe the same physical stimulus ($\mathbf{X}$, such as an emotional recording, a solid color, or a piece of music), speakers of different languages ($\mathbf{Z}$) categorize the space ($\mathbf{Y}$) differently [537]. The debate in cognitive science is about the question if those differences arise from cross-linguistic differences shaping perception [259, 372, 451] or if they can be explained by some universal principles [454–456]. The phenomenon was popularized by the "World Color Survey" [99], which showed that different languages have a different number of basic color terms. More recent work has shown that this phenomenon is not only limited to colors, but also to other modalities such as vision [99, 231], touch [240, 241], hearing [220, 233, 242], smell [62, 243], and taste [244]. Emotional prosody can be considered as another instance of grounded semantics [538], where participants all listen to the same emotional recordings, but annotate the space differently depending on the language they speak.

This shows that the three problems we identified for emotional prosody, deeply connect to a larger literature in cognitive science and to a related set of problems in computer science.

In the next sections, we will zoom out and discuss the over-arching limitations of the work presented in this thesis (Section 12.3), will show how this work fits in a broader research agenda (Section 12.4), and discuss the implications of the work to the field in general (Section 12.5).

12.3 Limitations

All three HITL algorithms incorporate human decisions in computer algorithms, by orchestrating human labor in a chain. The product of the chains can be prosodies associated with particular emotions, a representative sample of all emotional prosodies, or a weighted bag-of-words representation of all stimuli. In the next subsections, we will discuss the limitations and caveats of these approaches.

12.3.1 Individual differences

Previous research for GSP has shown that individual priors are similar to the priors of the group for certain domains, such as emotional prosody [1] and aesthetic appreciation [539], but this is not a priori the case for all domains, especially if priors are rather vague or subjective.

For example, in VoiceMe we showed that the face-voice match score mainly depended on aligning the assumed sex of the face and the voice. While we showed that voices for the same face were closer to each other than voices for the same sex (indicating that the matched voices were matched for characteristics beyond just sex), the effect was not as strong as the effect of assumed sex. This indicates that participants had vague priors about the relation between faces and voices, and this can be explained by the fact that one is rarely surprised about the voice of a person in daily life (except for some extreme cases, such as a deep voice for a woman, a smoky voice of a young person, or a pronounced accent where you would not expect it). Also, in RobotVoice, we anecdotally observed that people strongly disagreed on the voice of a robot, indicating that the priors of the participants for robot voices might be rather subjective.

The same problem exists for GAP and STEP, where raters in GAP might not agree on their representation of emotion (e.g., some raters would maybe first think about positive emotions and other raters first about negative emotions when asking about the concept of "emotion") and raters in STEP might not agree on the tags they provide (e.g., where some raters highly rate the relevance of the tag and other raters flag the tag for removal).

Chains are not guaranteed to converge if participants have different priors [69, 70]. This might partially explain that, across all validation experiments, we found that the ratings improved over iterations, but plateaued before reaching the end of the rating scale.

The difficulty here is that one a priori does not know if the priors of the participants are similar or not. Luckily, this can be quantified by comparing within- and across-participant chains in GSP. So, while earlier work has shown that group-level aggregation (e.g., taking the median response) can lead to faster convergence [1] and across-participant chains might be necessary in particular experiments[1], it might be harmful in domains where people disagree because they have different priors and the HITL algorithms will lead to suboptimal results. This needs to be considered when using these HITL algorithms.

12.3.2 Saliency of cues

Another difficulty is that many modalities, like the voice, consist of multifaceted signals and communicate multiple channels of information at the same time (e.g., assumed sex and age, emotions, and intentions), and some of these channels are more salient than others. For example, in the voice, the assumed sex of the speaker is more salient than voice timbre.

For STEP this implies that even when we ask people to describe their impression of an emotion in a fragment, some raters would still provide tags with respect to the assumed sex of the speaker because it is such a salient feature of the voice.

The same problem exists for GSP, where a few dominant channels can drive the percept. For example, in VoiceMe, we showed that the face-voice match score increased a lot once the assumed sex of the face was aligned with the voice. Interestingly, context can influence the sensitivity to those cues: In VoiceMe, we showed that fictional characters (such as cartoons or paintings) suffered less than photographs of faces from the assumed sex misalignment.

Also, we found that the same channel of information can differ in saliency across modalities. In RobotVoice, we showed again that the assumed sex was a dominant dimension for the voice, but for the image modality, the dichotomy between humanlike and robotic was particularly important. In this project, we also investigated cross-modal associations, where we showed that most tags were not used consistently across modalities (e.g., "animallike" robots do not have "animallike" voices), again indicating that the saliency of cues can differ across modalities.

[1] For example, a GSP experiment with 32 slider steps requires 32 stimuli at the same time. So, stimulus generation either needs to be very fast (32x real time) or massively parallelized. Since this is not always possible (e.g., for TTS models), across-participant chains (asynchronous design), as well as aggregation (reuse the same slider multiple times), can be used to relax these constraints.

These results highlight that not all channels of information in multifaceted stimuli are equally salient and that one piece of information can overshadow others, which should be taken into account when using these HITL algorithms.

12.3.3 Parametrization of the Stimulus Space

In GSP there is the additional question of how to parametrize the stimulus space. The results of a GSP experiment strongly depend on the parametrization of the stimulus space, and experimenters have to navigate the trade-off between the number of dimensions and expressiveness of the model.

For emotional prosody, we initially used parametric manipulations of the speech signal to change the prosody, however, such manipulations are problematic for more complex stimulus spaces—like prosody or music [540] – where changing a single stimulus dimension irrespective of the others can create artifacts and handcrafted manipulations can lead to a biased sample of the stimulus space.[2]

We, therefore, moved to the latent representation of the stimulus space, because if the model is trained on a diverse enough dataset, it (i) learns natural prosodic variations you can sample from using GSP (and you don't have to constrain the stimulus set by handpicked acoustic manipulations) and (ii) leading to fewer artifacts since you sample from natural variations the model learned during training.

However, latent representations – such as Global Style Tokens in Tacotron or a PCA on the speaker embeddings – have other challenges: the dimensions tend to be entangled, are often incomplete (e.g., the selected principal components only explain a portion of the variance), and it is difficult to balance naturalness and expressiveness of latent dimensions (especially if extreme values of the latent dimensions are not well-represented in the training data).

These limitations might have also contributed to the suboptimal results in the validation experiments (see Section 12.3.1).

It is to be noted that GSP in theory should be invariant to which generative model is used, as long as the model is expressive enough to cover the stimulus space and the slider is continuous in perceptual space (making it intuitive for the participants to use and easy to anticipate intermediate values). In various experiments, we have shown that using different generative models for the same stimulus space leads to consistent results. For example, faces optimized for a specific attribute using a StyleGAN trained on photos of faces leads to consistent results with a StyleGAN trained on art portraits [1] and also the prototypes of emotional prosody obtained using parametric manipulations of the speech signal are consistent with those obtained using a TTS model [3]. These results indicate that GSP is fairly invariant to the generative model used, as long as the generative model is expressive and continuous in perceptual space.

To summarize, the outcome of the GSP process can only be as good as the parametrization of the stimulus space. While there are stimulus spaces that can exhaustively be described with just a few parameters (such as color), this is not the case for more complex modalities. Therefore, researchers should carefully consider the parametrization of the stimulus space when using GSP.

[2] This also happened in the first GSP emotion experiment, where dimensions changing tremolo (as a proxy for shimmer) were inactive for all emotions.

12.3.4 Biases and stereotypes

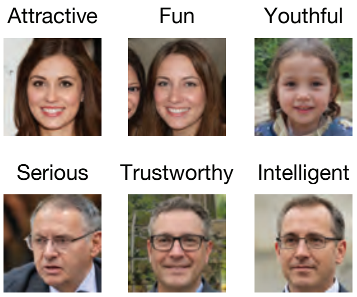


Figure 12.2: GSP faces Final face associations obtained using GSP and StyleGAN trained on Flickr-Faces-HQ [541] (cc-by-nc). Data from [1].

The three HITL algorithms allow us to describe mental representations and human biases. For example, in GSP we found that “attractive”, “fun”, and “youthful” were reliably associated with female faces and “serious”, “trustworthy”, and “intelligent” with male faces (Figure 12.2). Or with STEP we found that participants used tags like “smoky”, “sexy”, “submissive”, and “worried” for female voices and tags like “angry”, “loud”, “strong”, “bossy”, and “rich” for male voices (Figure 12.3). These findings reveal strong stereotypes about gender and raise the question of how these stereotypes can be mitigated and if these biases are inherent to the data or are introduced by the participants.

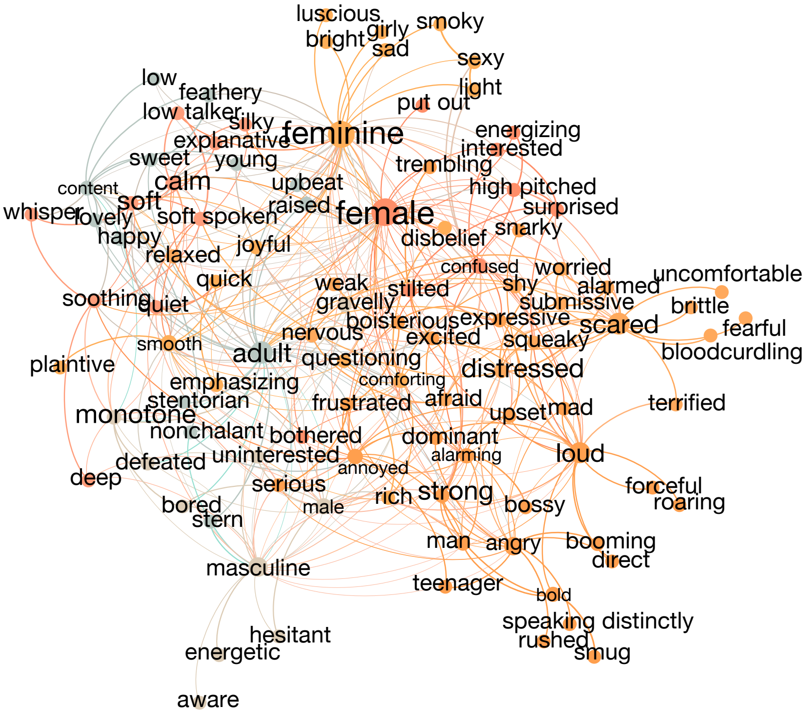


Figure 12.3: STEP describing speaker Unpublished data where participants were asked to describe the speaker fragments from the RAVDESS dataset [44].

One recurring concern has been that studying these biases and describing them has the potential to reinforce them. However, in order to mitigate biases it is important to first make them explicit. One particular problem is that readers of the work often conflate the subjective ratings with the objective truth. Just the fact that participants associate the term “attractive” with a female, does not mean that this is a causal relationship. It is a correlation, just like male hair loss correlates with income [542], but bald men don’t earn more money because they are bald, but because they are older and have more developed careers than young men. The same applies to the associations in the data: they are correlations, not causal relationships.

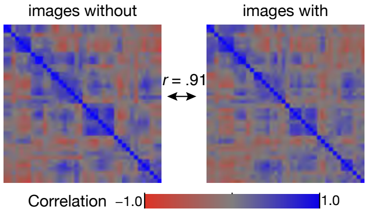


Figure 12.4: Effect of implicit bias training Correlations across terms with (left) and without (right) implicit bias training.

To actively mitigate these biases, we repeated the same experiments with and without implicit bias training in RobotVoice, where we found similar gender stereotypes for robots. Here, participants are first tested on their biases [543, 544], are confronted with the test outcome, then receive training on how to mitigate those biases, and we ascertain participants understood the training (text-comprehension questions). While participants clearly understood the training material (on average, 6/8 questions answered correctly), the training did not have a significant effect on the biases in the experiments. For the STEP experiment, describing the images of robots, the tags largely overlap, and the frequency of the shared tags is strongly correlated (*r* = .78). In the rating experiment, in which participants rated each of the dimensions for the images, the ratings with and without training were strongly correlated (*r* = .91; see Figure 12.4). This indicates that while participants were aware of their implicit biases (see comprehension questions), they did not substantially change their responses. This may be explained by the fact that the effects of the training are short-lived and hence barely change the implicit biases [545]. More broadly, this demonstrates that implicit biases are not becoming more pronounced as a result of our experimental procedure. In fact, it is quite challenging to influence these biases, even when employing what we consider the best practices in bias training. Instead, our method enables to exposure of these biases, representing a significant step toward mitigating their negative impact on societies.

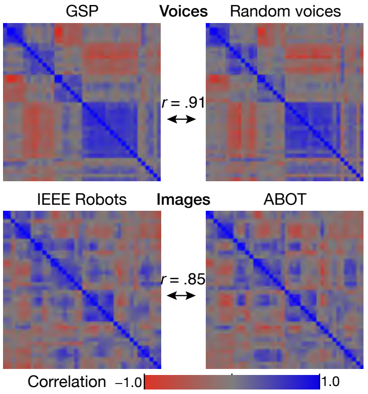


Figure 12.5: Replication in other datasets Correlations across terms with and without implicit bias training. Top row: Correlations between the association matrix in voices from GSP voices (left) and initial random voices (right). Bottom row: Correlations between association matrix in images from IEEE Robots (left) and ABOT (right) [531].

To address the question of whether the biases originate from participants or from the data (or the generative model which was trained on the data), we conducted two follow-up experiments in RobotVoice (Figure 12.5). First, we investigated if the GSP voice creation amplified the biases in the ratings of the voices. We repeated the voice rating experiment on random voices drawn from the GSP stimulus space. We found that the ratings on both datasets are highly correlated (*r* = .91), indicating that the biases in the voice ratings are not substantially amplified by the GSP process. In a second analysis, we re-ran the rating experiment on a new dataset of robot images [531]. Again, the ratings on both datasets are highly correlated (*r* = .85). While both image datasets might have similar data biases (for example, both datasets have an underrepresentation of perceived female robots, 13 % for IEEE Robots and 19 % for ABOT which is consistent with previous findings [546]), there was no overlap in the images between the two datasets and the images were drawn from different sources. This indicates that the biases in the image ratings are more likely to originate from the participants than from the data. Since GSP does not amplify the biases, future work can use GSP with a generative model to create images of robots, which would lead to the development of "customizable robots" as proposed by Schiebinger [547]. To summarize, this shows that the biases we discover are unlikely to be data biases, but reflect associations stored in the minds of participants. This is different from machine learning models trained on massive amounts of scraped data, which will inherit those biases even if they are not a representative sample of all data [548, 549].

One limitation of the implicit bias mitigation strategy might be that participants are never confronted with the associations in the data. For example, if a participant rates a cleaning robot as "submissive" and another participant as "female", this does not mean that the participant thinks that all females are submissive. One way to mitigate this is to confront participants with the associations and ask them if they agree with the association. We hypothesize that many participants would mark the association as incorrect because they are aware of the stereotype. This would allow us to systematically search for stimuli that have this association and prune them, thus reducing the bias in the data.

Another separate, but related concern is that the HITL algorithms may lead to stereotypical depictions of emotions, which are not representative of everyday emotional communication. For example, in GSP, participants will try to make the sentence the best representation of a particular emotion, which might lead to exaggerated or stereotypical depictions of the emotion. Also, in GAP, raters will select the most emotional creation, which over iterations might lead to selecting the most exaggerated prosody. However, both GSP and GAP can theoretically optimize for multiple objectives, for example, one part of the participants optimizes for naturalness and another part optimizes for emotionality. Furthermore, in GSP if a latent representation is used and the parametrization of the space samples from natural variations in the training data, the generated prosodies are less likely to be too extreme or exaggerated.

12.3.5 Cross-cultural validity

The goal of all three HITL algorithms has been to develop fairly general, language- and culture-agnostic cognitive pipelines that can potentially be employed in any culture. However, so far they have only been validated in a Western context. While GSP has a fairly simple interface (only a slider) and previous research has shown that sliders can be used cross-culturally [550, 551], the other two HITL algorithms might not be cross-culturally understandable. For example, in GAP asks creators to put themselves in the shoes of the previous creator, but this might not be socially acceptable in all cultures (e.g., it might be considered rude). Or raters in GAP might not be aligned in what is considered "emotional", for example, some cultures might not even have a term for "emotion" or the term might be associated with different kinds of emotions (e.g., while "disgust" is considered a basic emotion, it is maybe not the first emotion Westerners think of). The same limitations apply to STEP, where the text-based interface pre-assumes that participants are literate (although you can make an oral version) and participants are familiar with Western concepts learned in school (such as a Likert scale to rate the relevance of a tag).

These caveats should be considered when running these HITL algorithms in a cross-cultural context but can be mitigated by breaking down the task into subtasks and adapting the interface to the local context (e.g., providing an oral version of the task, providing a visual Likert scale, or providing a different interface), which requires careful piloting and validation in the respective culture. However, we empirically found that participants recruited from the internet across the globe tend to be fairly educated (thus are likely to be able to read and are familiar with school concepts like Likert scales, multiple- and forced-choice) and tend to be more globalized than participants from remote areas of the world, and thus more exposed to Western concepts, which might reduce the need for adaptation.

12.3.6 Human labour

All three HITL algorithms require human labor, which can be costly and time-consuming. A frequent concern is that human labor is not scalable and that the algorithms are not applicable to large datasets. One suggestion is to replace humans – at least in part – with DNNs. We have shown that LLMs can be integrated into the HITL algorithms and provide novel insights, for example, on the alignment of human and LLMs [12, 17]. This also has great potential for piloting the HITL algorithms and preparing a data analysis and visualization pipeline before the human data is collected [552, 553]. However, replacing humans entirely with DNNs comes at a risk of introducing biases and artifacts inherent to the model, which might not be representative of human data. So while the HITL algorithms are equally applicable to human and "machine participants", involving DNNs HITL algorithms is at most a proxy for human behavior and thus cannot be used instead of human data.

The HITL algorithms can also be used to create high-quality datasets, which, after pruning harmful biases and artifacts, can be used to train DNNs. Particularly, in the context of LLMs and foundation models, the HITL algorithms can be used to create a dataset that is representative of human data and can be used to fine-tune them to be more human-like [554].

Finally, the development of PsyNet and the international infrastructure we developed for running massive online experiments across the globe has shown that human labor can be efficiently orchestrated, the process can be automated, and the algorithms can be run on a large scale. Also, by running the experiments across the globe and by paying the participants according to the local wage, the costs can be reduced significantly, while enhancing the diversity of the participants and the generalizability of the results.

12.4 Future research

The methods developed in this thesis accommodate a wide range of research questions that can be applied to various domains and are both relevant to computer science and cognitive science.

The developed global infrastructure allows us to employ the HITL algorithms in a cross-cultural context and provide solutions to the lost-in-translation problem. In particular, one could train a multilingual, multispeaker model with a shared voice (i.e., speaker embeddings) and prosody latent space (e.g., Global Style Tokens). Since a priori, it is not clear if all languages have the same emotions and if they refer to the same concept, one can prompt users to match the prosody to emotionally evocative images (similar to VoiceMe and RobotVoice). The factorizability of the TTS model allows us to apply the same prosody to various languages, speakers, and genders, allowing us to shed light on how different perceptual categories (such as gender and culture) interact with the expression of emotions in speech.

Another project we are currently working on is to combine HITL algorithms to create a taxonomy of emotions in speech that is cross-culturally valid. In particular, we are using GAP to collect emotional prosodies from across the world and in a second step use STEP to distill a taxonomy of emotions. To break down the complexity of the project, we divide it into three stages. In the first stage, we want to collect GAP data from two languages I speak (e.g., English and German), validate that the emotionality of the prosodies increases over iterations, and run STEP in both languages on all stimuli. To make the representation denser, we will do a dense rating on the most-used tags per language. This pilot will help to: (i) validate the GAP task also works for non-English speakers, (ii) estimate the minimal number of shared stimuli across languages to align taxonomies in both languages,[3] (iii) verify the obtained taxonomy is of sufficient quality (e.g., "Wut" should be related to "anger", but not to "happiness"), (iv) estimate the need for dense rating, and if so, how many stimuli need to be rated to get a reliable estimate. In the second stage, we want to extend the pilot to nine languages (e.g., English, German, Tagalog-Filipino, Vietnamese, Russian, Greek, Hungarian, Mandarin, and Spanish) that span a large number of language families and are recruitable on Prolific[4]. Concretely, we will collect GAP for the seven new languages and run STEP on all stimuli from the own culture + the minimal number of shared stimuli per culture to align the taxonomies (which was determined in the previous pilot). This pilot will allow me to verify that the GAP task is also understandable for

[3] Alignment of taxonomies across languages is a non-trivial problem because the same emotion might be expressed differently in different languages. One way to solve this is to use a shared set of stimuli across languages and align the taxonomies based on the shared stimuli (use them as landmarks). So you could compute the co-occurrence of tags across languages and use this to align the taxonomies. Another approach is to do dimension reduction (e.g., MDS) on the landmarks and a separate MDS on all culture-specific stimuli. You can now predict the position of culture-specific stimuli in the landmark MDS space by minimizing the distance between the landmark coordinates in both MDS spaces. This analysis requires a minimal number of shared stimuli to be reliable, which is an empirical question.

[4] While we argued participants on Prolific are more WEIRD than on Lucid, they do on average give less noisy responses than Lucid participants, they know how to engage in audio experiments, and you can reach out to them if you have questions or they encounter problems. This makes Prolific a good platform for piloting cross-cultural studies.

languages very different from English and obtain an aligned taxonomy in all languages. The aligned taxonomy is then to be validated by native speakers of all nine languages. The results and analyses provide answers to the number of perceived emotions across languages, the proximity of the taxonomies (e.g., are linguistically close languages also close in the taxonomy), and the alignment sheds light on the universality of emotions.

In the last stage, we want to scale up the pipeline to a large number of languages and countries using Lucid. To do so, this requires two additional steps: (i) test the quality and the number of participants on Lucid who can engage in audio experiments and (ii) obtain a more principled way to obtain sentences that have a neutral meaning but can carry emotional prosody. To solve the latter, we consider running a sentence creation experiment in eight linguistically distinct languages. Participants should come up with sentences with a neutral meaning that can occur in their daily lives (blocking copy-pasting from the internet and prohibiting the use of LLMs). In the second step, all sentences are cross-translated, and all obtained sentences are rated by native speakers of all languages. The top-rated sentences across all languages are then used in the GAP experiments.

Once the pipeline is established, it will be fairly straightforward to investigate the relationship between the emotional space and stimulus spaces for other modalities. For example, one could also use GAP to create emotional vocalizations (like moaning or screaming) or use GSP within GAP to use sliders as proposed creations and use raters as some kind of "human aggregation" to overcome production constraints. Participants using generative models for stimulus creation will be allowed to extend the paradigm to a range of stimuli, such as music, images, or videos. STEP can also be used to annotate emotions in more principled sampled stimulus spaces. For example, we are currently working on a project in which music charts from the US, South Korea, and Brazil [63] are annotated with emotional tags across all three countries. Studying emotion taxonomies across modalities can provide insights into the alignment and divergence of emotion taxonomies (e.g., are certain emotions modality specific?) which goes far beyond existing work, who study the taxonomies for different modalities, but do not compare them [52, 53, 56, 208, 212–216].

Since we identified that emotional prosody is essentially a mapping problem between a stimulus and a semantic space (see Figure 12.1), one can also study this mapping in domains beyond emotion. For instance, the HITL algorithms can be used to study questions in grounded semantics, such as color naming, study mental representations beyond emotions (such as speaker intentions) or associations across domains [555, 556].

The developed HITL algorithms also contribute new tools to computer science. In particular, they can be used to create high-quality corpora and can provide solutions to three core problems inherent to most machine learning corpora. These corpora, as proposed earlier, can then be used to fine-tune LLMs and foundation models to improve the alignment between human and machine representations and obtain more human-like responses from the models. In particular, STEP can also be used to find more interpretable feature representations in trained models, which can then be used to obtain a better parametrization of the stimulus space in GSP. Finally, the HITL algorithms can be used to create benchmarks for evaluating the performance of state-of-the-art models.

12.5 Implications

The three HITL algorithms as well as the global infrastructure developed in this thesis have implications for both the study of emotional prosody and its applications in machine learning. GSP and STEP can help identify harmful biases and stereotypes in the data, which can be pruned to create high-quality datasets. GAP can be used to create a representative corpus of emotional prosody across a variety of languages, which can be used to train models that are more balanced and diverse. These corpora can also provide benchmarks for evaluating the performance of state-of-the-art models.

The work presented in this thesis also has theoretical implications for the study of emotional prosody. In particular, it advances the data-driven study of emotions by improved stimulus sampling (GAP), identifying emotional expressions in high-dimensional stimulus spaces (GSP), and by theory-free taxonomy curation (STEP). These tools can help answer fundamental questions about the universality of emotions, their origin, and the pattern of cross-cultural variation in emotional expressions.

Understanding how to create representative stimulus sets and effectively integrate human decisions to enhance the quality of training has become imperative. We believe the tools introduced in this dissertation can serve as a vital part of the evolving infrastructure that will shape the future of machine learning.

Bibliography

Here are the references in citation order.

[1] Peter M. C. Harrison, Raja Marjieh, Federico Adolfi, Pol van Rijn, Manuel Anglada-Tort, Ofer Tchernichovski, Pauline Larrouy-Maestri, and Nori Jacoby. 'Gibbs Sampling with People'. In: *Advances in Neural Information Processing systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. 2020 (cited on pages xi, 2, 4, 41–44, 104, 115, 123, 136–139).

[2] Dominik Schiller, Silvan Mertes, Pol van Rijn, and Elisabeth André. 'Analysis by Synthesis: Using an Expressive TTS Model as Feature Extractor for Paralinguistic Speech Classification'. In: *Interspeech*. 2021, pp. 486–490. DOI: 10.21437/Interspeech.2021-1587 (cited on pages xi, 4, 41, 120).

[3] Pol van Rijn, Silvan Mertes, Dominik Schiller, Peter M.C. Harrison, Pauline Larrouy-Maestri, Elisabeth André, and Nori Jacoby. 'Exploring Emotional Prototypes in a High Dimensional TTS Latent Space'. In: *Interspeech*. 2021, pp. 3870–3874. DOI: 10.21437/Interspeech.2021-1538 (cited on pages xi, 4, 41, 54, 114, 123, 138).

[4] Pol van Rijn, Silvan Mertes, Dominik Schiller, Piotr Dura, Hubert Siuzdak, Peter M. C. Harrison, Elisabeth André, and Nori Jacoby. 'VoiceMe: Personalized Voice Generation in TTS'. In: *Interspeech*. 2022, pp. 2588–2592. DOI: 10.21437/Interspeech.2022-10855 (cited on pages xi, 4, 113, 123).

[5] Hubert Siuzdak, Piotr Dura, Pol van Rijn, and Nori Jacoby. 'WavThruVec: Latent speech representation as intermediate features for neural speech synthesis'. In: *Interspeech 2022*. interspeech$_2$022. ISCA, Sept. 2022, pp. 833–837. DOI: 10.21437/interspeech.2022-10797 (cited on pages xi, 120).

[6] Pol van Rijn, Harin Lee, and Nori Jacoby. 'Bridging the Prosody GAP: Genetic Algorithm with People to Efficiently Sample Emotional Prosody'. In: *CogSci*. Proceedings of the Annual Meeting of the Cognitive Science Society. 2022 (cited on pages xi, 4, 51).

[7] Raja Marjieh, Pol van Rijn, Ilia Sucholutsky, Theodore Sumers, Harin Lee, Thomas L. Griffiths, and Nori Jacoby. 'Words Are All You Need? Language as an Approximation for Human Similarity Judgments'. In: *ICLR*. 2023 (cited on pages xi, 4, 42, 63, 104, 108, 125).

[8] Dominik Schiller, Silvan Mertes, Pol van Rijn, and Elisabeth André. 'Bridging the Gap: End-to-End Domain Adaptation for Emotional Vocalization Classification using Adversarial Learning'. In: *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*. MM '22. ACM, Oct. 2022, pp. 95–100. DOI: 10.1145/3551876.3554816 (cited on page xi).

[9] Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Tom Griffiths. 'What Language Reveals about Perception: Distilling Psychophysical Knowledge from Large Language Models'. In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 45.45 (2023) (cited on page xi).

[10] Pol van Rijn, Yue Sun, Harin Lee, Raja Marjieh, Ilia Sucholutsky, Francesca Lanzarini, Elisabeth André, and Nori Jacoby. 'Around the World in 60 Words: A Generative Vocabulary Test for Online Research'. In: *CogSci*. Proceedings of the Annual Meeting of the Cognitive Science Society. 2023 (cited on pages xi, 4, 80, 85, 104).

[11] Pol van Rijn and Pauline Larrouy-Maestri. 'Modelling Individual and Cross-Cultural Variation in the Mapping of Emotions to Speech Prosody'. In: *Nature Human Behaviour* 7.3 (2023), pp. 386–396. DOI: 10.1038/s41562-022-01505-5 (cited on pages xi, 4, 23, 29–31, 77).

[12] Jakob Niedermann, Ilia Sucholutsky, Raja Marjieh, Elif Celen, Thomas L Griffiths, Nori Jacoby, and Pol van Rijn. 'Studying the Effect of Globalization on Color Perception Using Multilingual Online Recruitment and Large Language Models'. In: *CogSci*. Proceedings of the Annual Meeting of the Cognitive Science Society. Vol. 46. 46. 2024 (cited on pages xii, 4, 85, 101, 141).

[13] Raja Marjieh, Pol van Rijn, Ilia Sucholutsky, Harin Lee, Tom Griffiths, and Nori Jacoby. 'A Rational Analysis of the Speech-to-Song Illusion'. In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 46.0 (2024) (cited on page xii).

[14] Raja Marjieh, Pol van Rijn, Ilia Sucholutsky, Harin Lee, Nori Jacoby, and Thomas Griffiths. 'Characterizing the Large-Scale Structure of Grounded Semantic Network'. In: (2024). DOI: 10.31234/osf.io/phte5. Pre-published (cited on pages xii, 4, 63).

[15] Pol van Rijn, Silvan Mertes, Kathrin Janowski, Katharina Weitz, Nori Jacoby, and Elisabeth André. 'Giving Robots a Voice: Human-in-the-Loop Voice Creation and Open-Ended Labeling'. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 1–34. DOI: 10.1145/3613904.3642038 (cited on pages xii, 4, 119, 123, 129).

[16] Harin Lee, Manuel Anglada-Tort, Oleg Sobchuk, Pol van Rijn, Marc Schönwiesner, Ofer Tchernichovski, Minsu Park, and Nori Jacoby. 'Global Music Discoveries Reveal Cultural Shifts during the War in Ukraine'. In: *10.31234/osf.io/7b98u* (2024). DOI: 10.31234/osf.io/7b98u. Pre-published (cited on page xii).

[17] Dun-Ming Huang, Pol van Rijn, Ilia Sucholutsky, Raja Marjieh, and Nori Jacoby. 'Characterizing Similarities and Divergences in Conversational Tones in Humans and LLMs by Sampling with People'. In: *ACL*. ACL 2024. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 10486–10512. DOI: 10.18653/v1/2024.acl-long.565 (cited on pages xii, 141).

[18] Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L. Griffiths. 'Large Language Models Predict Human Sensory Judgments across Six Modalities'. In: *Scientific Reports* 14.1 (2024), p. 21445. DOI: 10.1038/s41598-024-72071-1 (cited on pages xii, 4, 63, 104).

[19] Elif Celen, Pol van Rijn, Harin Lee, and Nori Jacoby. 'Are Expressions for Music Emotions the Same Across Cultures?' In: *arXiv:2502.08744* (2025) (cited on page xii).

[20] Harin Lee, Eline Van Geert, Elif Celen, Raja Marjieh, Pol van Rijn, Minsu Park, and Nori Jacoby. 'Visual and Auditory Aesthetic Preferences Across Cultures'. In: *arXiv:2502.14439* (2025) (cited on page xii).

[21] Elisabeth André, Laila Dybkjær, Wolfgang Minker, and Paul Heisterkamp, eds. *Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004. Proceedings*. Springer Berlin Heidelberg, 2004 (cited on page 1).

[22] Wade J Mitchell, Kevin A Szerszen, Amy Shirong Lu, Paul W Schermerhorn, Matthias Scheutz, and Karl F MacDorman. 'A Mismatch in the Human Realism of Face and Voice Produces an Uncanny Valley'. In: *i-Perception* 2.1 (2011), pp. 10–12. DOI: 10.1068/i0415 (cited on pages 1, 119).

[23] Lianne F. S. Meah and Roger K. Moore. 'The Uncanny Valley: A Focus on Misaligned Cues'. In: *Lecture Notes in Computer Science*. Springer International Publishing, 2014, pp. 256–265. DOI: 10.1007/978-3-319-11973-1_26 (cited on pages 1, 119).

[24] Angela Tinwell, Mark Grimshaw, and Deborah Abdel Nabi. 'The Effect of Onset Asynchrony in Audio-Visual Speech and the Uncanny Valley in Virtual Characters'. In: *International Journal of Mechanisms and Robotic Systems* 2.2 (2015), p. 97. DOI: 10.1504/ijmrs.2015.068991 (cited on pages 1, 119).

[25] Masahiro Mori, Karl MacDorman, and Norri Kageki. 'The Uncanny Valley [from the Field]'. In: *IEEE Robotics & Automation Magazine* 19.2 (2012), pp. 98–100. DOI: 10.1109/mra.2012.2192811 (cited on pages 1, 113, 119, 120).

[26] Valentin Schwind. *Implications of the Uncanny Valley of Avatars and Virtual Characters for Human-Computer Interaction*. Universität Stuttgart, 2018. DOI: 10.18419/OPUS-9936 (cited on pages 1, 119, 120).

[27] Kathrin Janowski, Hannes Ritschel, Birgit Lugrin, and Elisabeth André. 'Sozial Interagierende Roboter in Der Pflege'. In: *Pflegeroboter*. Springer Fachmedien Wiesbaden, 2018, pp. 63–87. DOI: 10.1007/978-3-658-22698-5_4 (cited on page 1).

[28] Natalie Vannini, Sibylle Enz, Maria Sapouna, Dieter Wolke, Scott Watson, Sarah Woods, Kerstin Dautenhahn, Lynne Hall, Ana Paiva, Elizabeth André, Ruth Aylett, and Wolfgang Schneider. '"FearNot!": A Computer-Based Anti-Bullying-Programme Designed to Foster Peer Intervention'. In: *European Journal of Psychology of Education* 26.1 (2010), pp. 21–44. DOI: 10.1007/s10212-010-0035-4 (cited on page 1).

[29] Maria Sapouna, Dieter Wolke, Natalie Vannini, Scott Watson, Sarah Woods, Wolfgang Schneider, Sibylle Enz, Lynne Hall, Ana Paiva, Elisabeth André, Kerstin Dautenhahn, and Ruth Aylett. 'Virtual Learning Intervention to Reduce Bullying Victimization in Primary School: A Controlled Trial'. In: *Journal of Child Psychology and Psychiatry* 51.1 (2009), pp. 104–112. DOI: 10.1111/j.1469-7610.2009.02137.x (cited on page 1).

[30] Scott E.J. Watson, Natalie Vannini, Sarah Woods, Kerstin Dautenhahn, Maria Sapouna, Sibylle Enz, Wolfgang Schneider, Dieter Wolke, Lynne Hall, Ana Paiva, Elizabeth André, and Ruth Aylett. 'Inter-Cultural Differences in Response to a Computer-Based Anti-Bullying Intervention'. In: *Educational Research* 52.1 (2010), pp. 61–80. DOI: 10.1080/00131881003588261 (cited on page 1).

[31] Tobias Baur, Ionut Damian, Patrick Gebhard, Kaska Porayska-Pomsta, and Elisabeth André. 'A Job Interview Simulation: Social Cue-Based Interaction with a Virtual Character'. In: *2013 International Conference on Social Computing*. IEEE, 2013, pp. 220–227. DOI: 10.1109/socialcom.2013.39 (cited on page 1).

[32] Markus Langer, Cornelius J. König, Patrick Gebhard, and Elisabeth André. 'Dear Computer, Teach Me Manners: Testing Virtual Employment Interview Training'. In: *International Journal of Selection and Assessment* 24.4 (2016), pp. 312–323. DOI: 10.1111/ijsa.12150 (cited on page 1).

[33] Keith Anderson, Elisabeth André, T. Baur, Sara Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, H. Jones, M. Ochs, C. Pelachaud, Kaśka Porayska-Pomsta, P. Rizzo, and Nicolas Sabouret. 'The TARDIS Framework: Intelligent Virtual Agents for Social Coaching in Job Interviews'. In: *Advances in Computer Entertainment*. Springer International Publishing, 2013, pp. 476–491. DOI: 10.1007/978-3-319-03161-3_35 (cited on page 1).

[34] Patrick Gebhard, Tanja Schneeberger, Elisabeth André, Tobias Baur, Ionut Damian, Gregor Mehlmann, Cornelius Konig, and Markus Langer. 'Serious Games for Training Social Skills in Job Interviews'. In: *IEEE Transactions on Games* 11.4 (2019), pp. 340–351. DOI: 10.1109/tg.2018.2808525 (cited on page 1).

[35] Patrick Gebhard, Tobias Baur, Ionut Damian, Gregor Mehlmann, Johannes Wagner, and Elisabeth André. 'Exploring Interaction Strategies for Virtual Characters to Induce Stress in Simulated Job Interviews'. In: *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*. Aamas '14. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 661–668 (cited on page 1).

[36] Ionut Damian, Tobias Baur, Birgit Lugrin, Patrick Gebhard, Gregor Mehlmann, and Elisabeth André. 'Games Are Better than Books: In-situ Comparison of an Interactive Job Interview Game with Conventional Training'. In: *Artificial Intelligence in Education*. Springer International Publishing, 2015, pp. 84–94. DOI: 10.1007/978-3-319-19773-9_9 (cited on page 1).

[37] Matteo Spezialetti, Giuseppe Placidi, and Silvia Rossi. 'Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives'. In: *Frontiers in Robotics and AI* 7 (2020). DOI: 10.3389/frobt.2020.532279 (cited on page 1).

[38] Joe Crumpton and Cindy L. Bethel. 'A Survey of Using Vocal Prosody to Convey Emotion in Robot Speech'. In: *International Journal of Social Robotics* 8.2 (2016), pp. 271–285. DOI: 10.1007/s12369-015-0329-4 (cited on page 1).

[39] Klaus R. Scherer. 'Acoustic Patterning of Emotion Vocalizations'. In: *The Oxford Handbook of Voice Perception*. Ed. by Sascha Frühholz and Pascal Belin. Oxford University Press, 2018, pp. 60–92. DOI: 10.1093/oxfordhb/9780198743187.013.4 (cited on pages 1, 5–7, 29, 33, 41).

[40] Patrik N. Juslin and Petri Laukka. 'Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code?' In: *Psychological Bulletin* 129.5 (2003), pp. 770–814. DOI: 10/ff6wbc (cited on pages 1, 7, 8, 29, 30, 33, 39, 41, 63).

[41] M. Kathleen Pichora-Fuller and Kate Dupuis. *Toronto Emotional Speech Set (TESS)*. Scholars Portal Dataverse, 2020. DOI: 10.5683/SP2/E8H2MF (cited on pages 1, 29, 30).

[42] Petri Laukka, Hillary Anger Elfenbein, Nutankumar S. Thingujam, Thomas Rockstuhl, Frederick K. Iraki, Wanda Chui, and Jean Althoff. 'The Expression and Recognition of Emotions in the Voice across Five Nations: A Lens Model Analysis Based on Acoustic Features.' In: *Journal of Personality and Social Psychology* 111.5 (2016), pp. 686–705. DOI: 10/f3tfdg (cited on pages 1, 29, 30, 41, 45, 51, 56, 63).

[43] S. Haq and P.J.B. Jackson. 'Machine Audition: Principles, Algorithms and Systems'. In: ed. by W. Wang. Hershey PA: IGI Global, 2010, pp. 398–423 (cited on pages 1, 29, 30).

[44] Steven R. Livingstone and Frank A. Russo. 'The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English'. In: *PLOS ONE* 13.5 (2018). Ed. by Joseph Najbauer, e0196391. DOI: 10/gd8gt8 (cited on pages 1, 29, 30, 52, 67–69, 115, 139).

[45] Tanja Bänziger, Marcello Mortillaro, and Klaus R. Scherer. 'Introducing the Geneva Multimodal Expression Corpus for Experimental Research on Emotion Perception.' In: *Emotion* 12.5 (2012), pp. 1161–1179. DOI: 10/bkgz8f (cited on pages 1, 29, 30).

[46] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. 'CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset'. In: *IEEE Transactions on Affective Computing* 5.4 (2014), pp. 377–390. DOI: 10/ggjdbh (cited on pages 1, 2, 29, 30, 51, 52, 56).

[47] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret McRorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, Noam Amir, and Kostas Karpouzis. 'The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data'. In: *Affective Computing and Intelligent Interaction*. Ed. by Ana C. R. Paiva, Rui Prada, and Rosalind W. Picard. Berlin, Heidelberg: Springer, 2007, pp. 488–500. DOI: 10.1007/978-3-540-74889-2_43 (cited on page 1).

[48] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 'The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent'. In: *IEEE Transactions on Affective Computing* 3.1 (2012), pp. 5–17. DOI: 10.1109/T-AFFC.2011.20 (cited on page 1).

[49] Angeliki Metallinou, Zhaojun Yang, Chi-chun Lee, Carlos Busso, Sharon Carnicke, and Shrikanth Narayanan. 'The USC CreativeIT Database of Multimodal Dyadic Interactions: From Speech and Full Body Motion Capture to Continuous Emotional Annotations'. In: *Language Resources and Evaluation* 50.3 (2015), pp. 497–521. DOI: 10.1007/s10579-015-9300-0 (cited on pages 1, 53).

[50] Olga Perepelkina, Evdokia Kazimirova, and Maria Konstantinova. 'RAMAS: Russian Multimodal Corpus of Dyadic Interaction for Affective Computing'. In: *Speech and Computer*. Springer International Publishing, 2018, pp. 501–510. DOI: 10.1007/978-3-319-99579-3_52 (cited on pages 1, 53).

[51] Reza Lotfian and Carlos Busso. 'Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings'. In: *IEEE Transactions on Affective Computing* 10.4 (2019), pp. 471–483. DOI: 10.1109/TAFFC.2017.2736999 (cited on page 1).

[52] Alan S. Cowen, Xia Fang, Disa Sauter, and Dacher Keltner. 'What Music Makes Us Feel: At Least 13 Dimensions Organize Subjective Experiences Associated with Music across Different Cultures'. In: *Proceedings of the National Academy of Sciences* 117.4 (2020), pp. 1924–1934. DOI: 10/gghmj2 (cited on pages 1, 2, 15, 16, 63, 143).

[53] Alan S. Cowen and Dacher Keltner. 'Self-Report Captures 27 Distinct Categories of Emotion Bridged by Continuous Gradients'. In: *Proceedings of the National Academy of Sciences* 114.38 (2017), E7900–E7909. DOI: 10/gfhzd2 (cited on pages 1, 15, 16, 22, 143).

[54] Amandine Lassalle, Delia Pigat, Helen O'Reilly, Steve Berggen, Shimrit Fridenson-Hayo, Shahar Tal, Sigrid Elfström, Anna Råde, Ofer Golan, Sven Bölte, Simon Baron-Cohen, and Daniel Lundqvist. 'The EU-emotion Voice Database'. In: *Behavior Research Methods* 51.2 (2018), pp. 493–506. DOI: 10.3758/s13428-018-1048-1 (cited on page 1).

[55] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Björn Schuller, Kam Star, Elnar Hajiyev, and Maja Pantic. 'SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.3 (2021), pp. 1022–1040. DOI: 10.1109/TPAMI.2019.2944808 (cited on page 1).

[56] Alan S. Cowen, Jeffrey A. Brooks, Gautam Prasad, Misato Tanaka, Yukiyasu Kamitani, Vladimir Kirilyuk, Krishna Somandepalli, Brendan Jou, Florian Schroff, Hartwig Adam, Disa Sauter, Xia Fang, Kunalan Manokara, Panagiotis Tzirakis, Moses Oh, and Dacher Keltner. 'How Emotion Is Experienced and Expressed in Multiple Cultures: A Large-Scale Experiment across North America, Europe, and Japan'. In: *Frontiers in Psychology* 15 (2024). DOI: 10.3389/fpsyg.2024.1350631 (cited on pages 1, 2, 15, 16, 22, 143).

[57] Daniel T. Cordaro, Rui Sun, Shanmukh Kamble, Niranjan Hodder, Maria Monroy, Alan Cowen, Yang Bai, and Dacher Keltner. 'The Recognition of 18 Facial-Bodily Expressions across Nine Cultures.' In: *Emotion* (2019). DOI: 10/ggfjht (cited on page 2).

[58] Cristina Soriano et al. 'Cross-Cultural Data Collection with the GRID Instrument,' in: *Components of Emotional Meaning: A Sourcebook*. Ed. by Johnny J. R. Fontaine, Klaus R. Scherer, and Cristina Soriano. Oxford University Press, 2013. DOI: 10.1093/acprof:oso/9780199592746.003.0007 (cited on page 2).

[59] Paul Ekman. 'An Argument for Basic Emotions'. In: *Cognition and Emotion* 6.3-4 (1992), pp. 169–200. DOI: 10/bh2cq3 (cited on pages 2, 10, 11, 30, 38, 113).

[60] Joshua Conrad Jackson, Joseph Watts, Teague R. Henry, Johann-Mattis List, Robert Forkel, Peter J. Mucha, Simon J. Greenhill, Russell D. Gray, and Kristen A. Lindquist. 'Emotion Semantics Show Both Cultural Variation and Universal Structure'. In: *Science* 366.6472 (2019), pp. 1517–1522. DOI: 10.1126/science.aaw8160 (cited on pages 2, 22, 24, 27, 39, 104).

[61] Bill Thompson, Seán G. Roberts, and Gary Lupyan. 'Cultural Influences on Word Meanings Revealed through Large-Scale Semantic Alignment'. In: *Nature Human Behaviour* 4.10 (2020), pp. 1029–1038. DOI: 10/gg7v9b (cited on pages 2, 24, 39).

[62] Asifa Majid et al. 'Differential Coding of Perception in the World's Languages'. In: *Proceedings of the National Academy of Sciences* 115.45 (2018), pp. 11369–11376. DOI: 10.1073/pnas.1720419115 (cited on pages 2, 23, 25, 136).

[63] Harin Lee, Frank Höger, Marc Schönwiesner, Minsu Park, and Nori Jacoby. 'Cross-Cultural Mood Perception in Pop Songs and Its Alignment with Mood Detection Algorithms'. In: *The 22nd International Society for Music Information Retrieval Conference (ISMIR)*. Online, 2021, pp. 366–373. DOI: 10.5281/zenodo.5625680 (cited on pages 2, 143).

[64] Adam Sanborn and Thomas Griffiths. 'Markov Chain Monte Carlo with People'. In: *Advances in Neural Information Processing Systems*. Ed. by J. Platt, D. Koller, Y. Singer, and S. Roweis. Vol. 20. Curran Associates, Inc., 2008 (cited on pages 2, 21, 54).

[65] Raja Marjieh, Ilia Sucholutsky, Thomas A. Langlois, Nori Jacoby, and Thomas L. Griffiths. 'Analyzing Diffusion as Serial Reproduction'. In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. ICML'23. Honolulu, Hawaii, USA: JMLR.org, 2023, pp. 24005–24019 (cited on pages 2, 18).

[66] Simon Kirby, Hannah Cornish, and Kenny Smith. 'Cumulative Cultural Evolution in the Laboratory: An Experimental Approach to the Origins of Structure in Human Language'. In: *Proceedings of the National Academy of Sciences* 105.31 (2008), pp. 10681–10686 (cited on pages 2, 18, 21, 66).

[67] Jing Xu, Mike Dowman, and Thomas L. Griffiths. 'Cultural Transmission Results in Convergence towards Colour Term Universals'. In: *Proceedings of the Royal Society B: Biological Sciences* 280.1758 (2013), p. 20123073. DOI: 10.1098/rspb.2012.3073 (cited on pages 2, 21, 103, 108).

[68] T. L. Griffiths, Adam N. Sanborn, R. Marjieh, T. Langlois, J. Xu, and N. Jacoby. 'Estimating Subjective Probability Distributions'. In: ed. by T. L. Griffiths, Nick Chater, and Joshua B. Tenenbaum. Boston, MA: MIT Press, 2024 (cited on pages 2, 20).

[69] Thomas L. Griffiths and Michael L. Kalish. 'Language Evolution by Iterated Learning With Bayesian Agents'. In: *Cognitive Science* 31.3 (2007), pp. 441–480. DOI: 10.1080/15326900701326576 (cited on pages 2, 18, 21, 137).

[70] Thomas L Griffiths and Michael L Kalish. 'A Bayesian View of Language Evolution by Iterated Learning'. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 27. 27. 2005 (cited on pages 2, 21, 66, 137).

[71] Frederic Charles Bartlett. *Remembering: A Study in Experimental and Social Psychology*. Remembering: A Study in Experimental and Social Psychology. New York, NY, US: Cambridge University Press, 1932, pp. xix, 317. xix, 317 (cited on pages 2, 18).

[72] Thomas A. Langlois, Nori Jacoby, Jordan W. Suchow, and Thomas L. Griffiths. 'Serial Reproduction Reveals the Geometry of Visuospatial Representations'. In: *Proceedings of the National Academy of Sciences* 118.13 (2021). DOI: 10.1073/pnas.2012938118 (cited on pages 2, 18, 19).

[73] Thomas Langlois, Nori Jacoby, Jordan W. Suchow, and Tom Griffiths. 'Orthogonal Multi-View Three-Dimensional Object Representations in Memory Revealed by Serial Reproduction'. In: *CogSci*. Proceedings of the Annual Meeting of the Cognitive Science Society. Ed. by Ashok K. Goel, Colleen M. Seifert, and Christian Freksa. cognitivesciencesociety.org, 2019, pp. 2078–2083 (cited on page 2).

[74] Nori Jacoby and Josh H. McDermott. 'Integer Ratio Priors on Musical Rhythm Revealed Cross-culturally by Iterated Reproduction'. In: *Current Biology* 27.3 (2017), pp. 359–370. DOI: 10/f9rjct (cited on pages 2, 18–20).

[75] Nori Jacoby et al. 'Commonality and Variation in Mental Representations of Music Revealed by a Cross-Cultural Comparison of Rhythm Priors in 15 Countries'. In: *Nature Human Behaviour* (2024). DOI: 10.1038/s41562-023-01800-9 (cited on pages 2, 18, 19, 85).

[76] Manuel Anglada-Tort, Peter M. C. Harrison, Harin Lee, and Nori Jacoby. 'Large-Scale Iterated Singing Experiments Reveal Oral Transmission Mechanisms Underlying Music Evolution'. In: *Current Biology* 33.8 (2023), 1472–1486.e12. DOI: 10.1016/j.cub.2023.02.070 (cited on pages 2, 18, 20).

[77] Manuel Anglada-Tort, Peter M. C. Harrison, and Nori Jacoby. 'Studying the Effect of Oral Transmission on Melodic Structure Using Online Iterated Singing Experiments'. In: (2022). DOI: 10.1101/2022.05.10.491366 (cited on pages 2, 18).

[78] Adam N. Sanborn, Thomas L. Griffiths, and Richard M. Shiffrin. 'Uncovering Mental Representations with Markov Chain Monte Carlo'. In: *Cognitive Psychology* 60.2 (2010), pp. 63–106. DOI: 10.1016/j.cogpsych.2009.07.001 (cited on pages 2, 21).

[79] Charles Blundell, Adam Sanborn, and Tom Griffiths. 'Look-Ahead Monte Carlo with People'. In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 34.34 (2012) (cited on page 2).

[80] Anne S. Hsu, Jay B. Martin, Adam N. Sanborn, and Thomas L. Griffiths. 'Identifying Category Representations for Complex Stimuli Using Discrete Markov Chain Monte Carlo with People'. In: *Behavior Research Methods* 51.4 (2019), pp. 1706–1716. DOI: 10.3758/s13428-019-01201-9 (cited on page 2).

[81] Christina Katsimerou, Joris Albeda, Alina Huldtgren, Ingrid Heynderickx, and Judith A. Redi. 'Crowdsourcing Empathetic Intelligence: The Case of the Annotation of EMMA Database for Emotion and Mood Recognition'. In: *ACM Trans. Intell. Syst. Technol.* 7.4 (2016), 51:1–51:27. DOI: 10.1145/2897369 (cited on page 2).

[82] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 'GoEmotions: A Dataset of Fine-Grained Emotions'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, 2020, pp. 4040–4054. DOI: 10.18653/v1/2020.acl-main.372 (cited on page 2).

[83] Jennifer J. Sun, Ting Liu, Alan S. Cowen, Florian Schroff, Hartwig Adam, and Gautam Prasad. 'EEV Dataset: Predicting Expressions Evoked by Diverse Videos'. 2020 (cited on page 2).

[84] Letizia Cerqueglini. 'Changes in Mutallat Arabic Color Language and Cognition Induced by Contact with Modern Hebrew'. In: *Studies in the Linguistic Sciences: Illinois Working Papers* (2021) (cited on pages 3, 101, 134).

[85] Abdulrahman S. Al-rasheed. 'Further Evidence for Arabic Basic Colour Categories'. In: *Psychology (Savannah, Ga.)* 5.15 (2014), pp. 1714–1729. DOI: 10.4236/psych.2014.515179 (cited on pages 3, 101, 134).

[86] Isabel Forbes. 'Structural Semantics with Particular Reference to the Vocabulary of Colour in Modern Standard French'. PhD thesis. The University of Edinburgh, 1976 (cited on pages 3, 101, 134).

[87] Heinrich Zollinger. 'Why Just Turquoise? Remarks on the Evolution of Color Terms'. In: *Psychological Research* 46.4 (1984), pp. 403–409. DOI: 10.1007/BF00309072 (cited on pages 3, 101, 103, 134).

[88] Ian Davies and Greville Corbett. 'The Basic Color Terms of Russian'. In: *Linguistics* 32 (1994), pp. 65–90. DOI: 10.1515/ling.1994.32.1.65 (cited on pages 3, 101, 103, 134).

[89] Emre Ozgen and Ian Davies. 'Turkish Color Terms: Tests of Berlin and Kay's Theory of Color Universels and Linguistic Relativity'. In: *Linguistics* 36 (1998), pp. 919–956. DOI: 10.1515/ling.1998.36.5.919 (cited on pages 3, 101, 103, 134).

[90] Jonathan Winawer, Nathan Witthoft, Michael C. Frank, Lisa Wu, Alex R. Wade, and Lera Boroditsky. 'Russian Blues Reveal Effects of Language on Color Discrimination'. In: *Proceedings of the National Academy of Sciences* 104.19 (2007), pp. 7780–7785. DOI: 10.1073/pnas.0701644104 (cited on pages 3, 23, 24, 101, 103, 106, 108, 134).

[91] Guillaume Thierry, Panos Athanasopoulos, Alison Wiggett, Benjamin Dering, and Jan-Rouke Kuipers. 'Unconscious Effects of Language-Specific Terminology on Preattentive Color Perception'. In: *PNAS Proceedings of the National Academy of Sciences of the United States of America* 106.11 (2009), pp. 4567–4570. DOI: 10.1073/pnas.0811155106 (cited on pages 3, 101, 103, 134).

[92] D. T. Lindsey and A. M. Brown. 'The Color Lexicon of American English'. In: *Journal of Vision* 14.2 (2014), pp. 17–17. DOI: 10.1167/14.2.17 (cited on pages 3, 101, 104, 106, 108, 134).

[93] Ichiro Kuriki, Ryan Lange, Yumiko Muto, Angela M. Brown, Kazuho Fukuda, Rumi Tokunaga, Delwin T. Lindsey, Keiji Uchikawa, and Satoshi Shioiri. 'The Modern Japanese Color Lexicon'. In: *Journal of Vision* 17.3 (2017), p. 1. DOI: 10.1167/17.3.1 (cited on pages 3, 101, 104, 134).

[94] Mathilde Josserand, Serge Caparos, François Pellegrino, and Dan Dediu. 'The Colour Lexicon Is Shaped by Environment and Biology: Comparing Himba and French Colour Perception'. In: *Joint Conference on Language Evolution*. Ed. by Andrea Ravignani, Rie Asano, Olga Vasileva, Slawomir Wacewicz, Daria Valente, Francesco Ferretti, Stefan Hartmann, Misato Hayashi, Yannick Jadoul, Mauricio Martins, Yohei Oseki, and Evelina Daniela Rodrigues. The Evolution of Language Proceedings of the Joint Conference on Language Evolution (JCoLE). Kanazawa, Japan: Joint Conference on Language Evolution (JCoLE), 2022, pp. 368–370 (cited on pages 3, 101, 134).

[95] Mingshan Xu, Jingtao Zhu, and Antonio Benítez-Burraco. 'A Comparison of Basic Color Terms in Mandarin and Spanish'. In: *Color Research & Application* 48.6 (2023), pp. 709–720. DOI: 10.1002/col.22863 (cited on pages 3, 101, 134).

[96] Giulia Paggetti, Gloria Menegaz, and Galina V. Paramei. 'Color Naming in Italian Language'. In: *Color Research and Application* 41 (2016), pp. 402–415 (cited on pages 3, 101, 134).

[97] Delwin T. Lindsey and Angela M. Brown. 'World Color Survey Color Naming Reveals Universal Motifs and Their Within-Language Diversity'. In: *Proceedings of the National Academy of Sciences* 106.47 (2009), pp. 19785–19790. DOI: 10.1073/pnas.0910981106 (cited on pages 3, 101, 134).

[98] Terry Regier and Paul Kay. 'Language, Thought, and Color: Whorf Was Half Right'. In: *Trends in Cognitive Sciences* 13.10 (2009), pp. 439–446. DOI: 10.1016/j.tics.2009.07.001 (cited on pages 3, 101, 134).

[99] Paul Kay, Brent Berlin, Luisa Maffi, William R Merrifield, and Richard Cook. *The World Color Survey*. Citeseer, 2009 (cited on pages 3, 23, 104, 136).

[100] Florian Eyben, Klaus R. Scherer, Bjorn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 'The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing'. In: *IEEE Transactions on Affective Computing* 7.2 (2016), pp. 190–202. DOI: 10/f3sfxq (cited on pages 4, 8, 31).

[101] Alan Cruttenden. *Intonation*. 2nd ed. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press, 1997 (cited on page 5).

[102] Delphine Dahan. 'Prosody and Language Comprehension'. In: *WIREs Cognitive Science* 6.5 (2015), pp. 441–452. DOI: 10.1002/wcs.1355 (cited on page 5).

[103] Roger K. Moore and B. Mitchinson. 'Creating a Voice for MiRo, the World's First Commercial Biomimetic Robot'. In: INTERSPEECH. Stockholm, Sweden, 2017 (cited on page 5).

[104] Attila Andics and Tamás Faragó. 'Voice Perception Across Species'. In: *The Oxford Handbook of Voice Perception*. Ed. by Sascha Frühholz and Pascal Belin. Oxford University Press, 2018, pp. 362–392. DOI: 10.1093/oxfordhb/9780198743187.013.16 (cited on pages 5, 6, 8, 9).

[105] Melissa A. Redford. *The Handbook of Speech Production*. 1st ed. Blackwell Handbooks in Linguistics. Malden, MA: Wiley-Blackwell, 2015 (cited on page 5).

[106] Ingo R. Titze. *Principles of Voice Production*. Prentice Hall, 1994. 392 pp. (cited on pages 5, 6).

[107] Gunnar Fant. *Acoustic Theory of Speech Production, With Calculations based on X-Ray Studies of Russian Articulations*. Reprint 2012. Berlin, Boston: De Gruyter Mouton, 2012 (cited on pages 5, 6).

[108] Theodore Dimon and G. David Brown. *Anatomy of the Voice: An Illustrated Guide for Singers, Vocal Coaches, and Speech Therapists*. Berkeley, California: North Atlantic Books, 2018. 102 pp. (cited on page 5).

[109] Paul Boersma. 'Chapter 17: Acoustic Analysis'. In: *Research Methods in Linguistics.* Ed. by Robert J. Podesva and Devyani Sharma. Cambridge: Cambridge University Press, 2014, pp. 375–396 (cited on page 5).

[110] A. J. Oxenham. 'Pitch Perception'. In: *Journal of Neuroscience* 32.39 (2012), pp. 13335–13338. DOI: 10/ggcgq8 (cited on page 5).

[111] Jeannette D Hoit and Gary Weismer. *Foundations of Speech and Hearing: Anatomy and Physiology*. Plural Publishing, 2018 (cited on page 5).

[112] Carlos Gussenhoven. *The Phonology of Tone and Intonation (Research Surveys in Linguistics)*. Cambridge University Press, 2004 (cited on pages 6, 7).

[113] Arthur H. Benade. *Fundamentals of Musical Acoustics*. 2., rev. ed. Dover Books on Music, Music History. New York, NY: Dover Publ, 1990. 596 pp. (cited on page 6).

[114] Johan Sundberg. 'The Acoustics of the Singing Voice'. In: *Scientific American* 236.3 (1977), pp. 82–91. DOI: 10/d67wt2 (cited on page 6).

[115] Randolph E. Deal and Floyd W. Emanuel. 'Some Waveform and Spectral Features of Vowel Roughness'. In: *Journal of Speech and Hearing Research* 21.2 (1978), pp. 250–264. DOI: 10.1044/jshr.2102.250 (cited on pages 7–9).

[116] C. Rose Rabinov, Jody Kreiman, Bruce R. Gerratt, and Steven Bielamowicz. 'Comparing Reliability of Perceptual Ratings of Roughness and Acoustic Measures of Jitter'. In: *Journal of Speech, Language, and Hearing Research* 38.1 (1995), pp. 26–32. DOI: 10/ghtztc (cited on pages 7–9, 45).

[117] Paul de Lacy, ed. *The Cambridge Handbook of Phonology*. Cambridge Handbooks in Language and Linguistics. Cambridge: Cambridge University Press, 2007 (cited on page 7).

[118] Peter Ladefoged and Keith Johnson. *A Course in Phonetics*. Cengage Learning, 2014. 356 pp. (cited on page 7).

[119] Rainer Banse and Klaus R. Scherer. 'Acoustic Profiles in Vocal Emotion Expression.' In: *Journal of Personality and Social Psychology* 70.3 (1996), pp. 614–636. DOI: 10.1037/0022-3514.70.3.614 (cited on pages 7, 8, 29, 41, 52, 119).

[120] Martijn Goudbeek and Klaus Scherer. 'Beyond Arousal: Valence and Potency/Control Cues in the Vocal Expression of Emotion'. In: *The Journal of the Acoustical Society of America* 128.3 (2010), p. 1322. DOI: 10.1121/1.3466853 (cited on pages 7, 63).

[121] Hiroya Fujisaki. 'Prosody, Models, and Spontaneous Speech'. In: *Computing Prosody: Computational Models for Processing Spontaneous Speech*. Ed. by Yoshinori Sagisaka, Nick Campbell, and Norio Higuchi. New York, NY: Springer US, 1997, pp. 27–42. DOI: 10.1007/978-1-4612-2258-3_3 (cited on page 7).

[122] Julia Hirschberg. 'Pragmatics and Prosody'. In: *The Oxford Handbook of Pragmatics*. Ed. by Yan Huang. Oxford University Press, 2017. DOI: 10.1093/oxfordhb/9780199697960.013.28 (cited on page 7).

[123] Yi Xu. 'Prosody, Tone, and Intonation'. In: *The Routledge Handbook of Phonetics*. Routledge, 2019 (cited on pages 7, 101).

[124] George Saintsbury. *A History of English Prosody from the Twelfth Century to the Present Day*. Macmillan and Company, limited, 1906. 458 pp. (cited on page 7).

[125] Eric Weiskott. 'Before Prosody: Early English Poetics in Practice and Theory'. In: *Modern Language Quarterly* 77.4 (2016), pp. 473–498. DOI: 10.1215/00267929-3648701 (cited on page 7).

[126] Nimal P. Parawahera. 'The Emergence of Prosody in Linguistic Theory'. In: (1999) (cited on page 7).

[127] D. Robert Ladd. *Intonational Phonology*. 2. Cambridge Studies in Linguistics. Cambridge u.a.: Cambridge Univ. Press, 2008. XIX, 349 S. (Cited on page 7).

[128] Daniel Hirst. 'The Analysis by Synthesis of Speech Melody: From Data to Models'. In: *Journal of Speech Sciences* 1.1 (2011), pp. 55–83 (cited on page 7).

[129] Björn Schuller, Gerhard Rigoll, and Manfred Lang. 'Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine-Belief Network Architecture'. In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. IEEE, 2004, pp. I–577 (cited on pages 7, 10).

[130] B. Schuller, A. Batliner, Christian Bergler, Eva-Maria Messner, A. Hamilton, Shahin Amiriparian, A. Baird, Georgios Rizos, M. Schmitt, Lukas Stappen, H. Baumeister, Alexis Deighton MacIntyre, and Simone Hantke. 'The INTERSPEECH 2020 Computational Paralinguistics Challenge. Elderly Emotion, Breathing & Masks'. In: *INTERSPEECH*. 2020 (cited on page 7).

[131] Charles Darwin. *The Expression of the Emotions in Man and Animals*. The Expression of the Emotions in Man and Animals. London, England: John Murray, 1872, pp. vi, 374. vi, 374 (cited on pages 7, 10, 11).

[132] Janet Breckenridge Pierrehumbert. 'The Phonology and Phonetics of English Intonation'. Thesis. Massachusetts Institute of Technology, 1980 (cited on page 7).

[133] Paul Boersma and David Weenink. *Praat: Doing Phonetics by Computer [Computer Program]*. 2018. URL: http://www.praat.org/ (cited on pages 7, 8, 31, 44, 121).

[134] Florian Eyben, Martin Wöllmer, and Björn Schuller. 'Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor'. In: *Proceedings of the 18th ACM International Conference on Multimedia*. MM '10: ACM Multimedia Conference. Firenze Italy: ACM, 2010, pp. 1459–1462. DOI: 10.1145/1873951.1874246 (cited on pages 7, 8).

[135] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 'Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor'. In: *Proceedings of the 21st ACM International Conference on Multimedia*. MM '13: ACM Multimedia Conference. Barcelona Spain: ACM, 2013, pp. 835–838. DOI: 10.1145/2502081.2502224 (cited on pages 7, 8).

[136] Megan K. MacPherson, Defne Abur, and Cara E. Stepp. 'Acoustic Measures of Voice and Physiologic Measures of Autonomic Arousal during Speech as a Function of Cognitive Load'. In: *Journal of Voice* 31.4 (2017), 504.e1–504.e9. DOI: 10.1016/j.jvoice.2016.10.021 (cited on page 8).

[137] Keith W. Godin and John H. L. Hansen. 'Physical Task Stress and Speaker Variability in Voice Quality'. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2015.1 (2015), p. 29. DOI: 10.1186/s13636-015-0072-7 (cited on page 8).

[138] Erika H. Siegel and Jeanine K. Stefanucci. 'A Little Bit Louder Now: Negative Affect Increases Perceived Loudness'. In: *Emotion* 11.4 (2011), pp. 1006–1011. DOI: 10.1037/a0024590 (cited on page 8).

[139] Michael J Owren and Jo-Anne Bachorowski. 'Measuring Emotion-Related Vocal Acoustics'. In: *Handbook of emotion elicitation and assessment* (2007), pp. 239–266 (cited on page 8).

[140] Timo Leino. 'Long-Term Average Spectrum in Screening of Voice Quality in Speech: Untrained Male University Students'. In: *Journal of Voice* 23.6 (2009), pp. 671–676. DOI: 10/cvnt93 (cited on pages 8, 9).

[141] Lucas Tamarit, Martijn Goudbeek, and Klaus Scherer. 'Spectral Slope Measurements in Emotionally Expressive Speech'. In: *Speech Analysis and Processing for Knowledge Discovery*. ISCA ITRW. 2008, p. 4 (cited on pages 8, 9).

[142] Ralph O. Coleman. 'Male and Female Voice Quality and Its Relationship to Vowel Formant Frequencies'. In: *Journal of Speech and Hearing Research* 14.3 (1971), pp. 565–577. DOI: 10.1044/jshr.1403.565 (cited on pages 8, 9).

[143] Agaath M. C. Sluijter and Vincent J. van Heuven. 'Spectral Balance as an Acoustic Correlate of Linguistic Stress'. In: *The Journal of the Acoustical Society of America* 100.4 (1996), pp. 2471–2485. DOI: 10/c8rrnq (cited on pages 8, 9).

[144] Christer Gobl and Ailbhe Ní Chasaide. 'The Role of Voice Quality in Communicating Emotion, Mood and Attitude'. In: *Speech Communication* 40.1-2 (2003), pp. 189–212. DOI: 10.1016/S0167-6393(02)00082-1 (cited on page 8).

[145] K.V. Krishna Kishore and P. Krishna Satish. 'Emotion Recognition in Speech Using MFCC and Wavelet Features'. In: *2013 3rd IEEE International Advance Computing Conference (IACC)*. 2013 3rd IEEE International Advance Computing Conference (IACC). 2013, pp. 842–847. DOI: 10.1109/IAdCC.2013.6514336 (cited on page 8).

[146] S. Lalitha, D. Geyasruti, R. Narayanan, and Shravani M. 'Emotion Detection Using MFCC and Cepstrum Features'. In: *Procedia Computer Science*. Proceedings of the 4th International Conference on Eco-friendly Computing and Communication Systems 70 (2015), pp. 29–35. DOI: 10.1016/j.procs.2015.10.020 (cited on page 8).

[147] M. S. Likitha, Sri Raksha R. Gupta, K. Hasitha, and A. Upendra Raju. 'Speech Based Human Emotion Recognition Using MFCC'. In: *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). 2017, pp. 2257–2260. DOI: 10.1109/WiSPNET.2017.8300161 (cited on page 8).

[148] Kurt Hammerschmidt and Uwe Jürgens. 'Acoustical Correlates of Affective Prosody'. In: *Journal of Voice* 21.5 (2007), pp. 531–540. DOI: 10/c8d66q (cited on page 8).

[149] Ratree Wayland, Scott Gargash, and Allard Longman. 'Acoustic and Perceptual Investigation of Breathy Voice'. In: *The Journal of the Acoustical Society of America* 97.5 (1995), pp. 3364–3364. DOI: 10/fwpgb3 (cited on page 9).

[150] Birgitta Burger, Marc R. Thompson, Geoff Luck, Suvi Saarikallio, and Petri Toiviainen. 'Influences of Rhythm- and Timbre-Related Musical Features on Characteristics of Music-Induced Movement'. In: *Frontiers in Psychology* 4 (2013). DOI: 10/gbfpv2 (cited on page 9).

[151] Francesc Alías, Joan Socoró, and Xavier Sevillano. 'A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds'. In: *Applied Sciences* 6.5 (2016), p. 143. DOI: 10/ghvnff (cited on page 9).

[152] Henrik Nordström, Petri Laukka, Marc Pell, Stockholms universitet, and Samhällsvetenskapliga fakulteten. *Emotional Communication in the Human Voice*. 2019 (cited on page 9).

[153] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 'Adieu Features? End-to-End Speech Emotion Recognition Using a Deep Convolutional Recurrent Network'. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5200–5204 (cited on pages 9, 77).

[154] Michael Freitag, Shahin Amiriparian, Sergey Pugachevskiy, Nicholas Cummins, and Bj\ "{o}rn Schuller. 'auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks'. In: *Journal of Machine Learning Research* 18.173 (2018), pp. 1–5 (cited on pages 9, 77).

[155] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller. 'Snore Sound Classification Using Image-Based Deep Spectrum Features'. In: Proc. Interspeech 2017. 2017, pp. 3512–3516. DOI: 10.21437/Interspeech.2017-434 (cited on page 9).

[156] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 'Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations'. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020, pp. 12449–12460 (cited on page 9).

[157] HsuWei-Ning, BolteBenjamin, TsaiYao-Hung Hubert, LakhotiaKushal, SalakhutdinovRuslan, and Mohamed-Abdelrahman. 'HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units'. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2021). DOI: 10.1109/TASLP.2021.3122291 (cited on page 9).

[158] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 'WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing'. In: *IEEE Journal of Selected Topics in Signal Processing* 16.6 (2022), pp. 1505–1518. DOI: 10.1109/JSTSP.2022.3188113 (cited on page 9).

[159] Mirco Ravanelli et al. *SpeechBrain: A General-Purpose Speech Toolkit*. 2021 (cited on page 9).

[160] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. 'Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.9 (Sept. 2023), pp. 10745–10759. DOI: 10.1109/tpami.2023.3263585 (cited on pages 9, 77).

[161] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 'SUPERB: Speech Processing Universal Performance Benchmark'. In: *Proc. Interspeech 2021*. 2021, pp. 1194–1198. DOI: 10.21437/Interspeech.2021-1775 (cited on page 9).

[162] Johannes Wagner, Dominik Schiller, Andreas Seiderer, and Elisabeth André. 'Deep Learning in Paralinguistic Recognition Tasks: Are Hand-crafted Features Still Relevant?' In: Proc. Interspeech 2018. 2018, pp. 147–151. DOI: 10.21437/Interspeech.2018-1238 (cited on page 9).

[163] Björn W. Schuller, Anton Batliner, Shahin Amiriparian, Alexander Barnhill, Maurice Gerczuk, Andreas Triantafyllopoulos, Alice E. Baird, Panagiotis Tzirakis, Chris Gagne, Alan S. Cowen, Nikola Lackovic, Marie-José Caraty, and Claude Montacié. 'The ACM Multimedia 2023 Computational Paralinguistics Challenge: Emotion Share & Requests'. In: *Proceedings of the 31st ACM International Conference on Multimedia*. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 9635–9639. DOI: 10.1145/3581783.3612835 (cited on page 9).

[164] James A. Russell. 'A Circumplex Model of Affect'. In: *Journal of Personality and Social Psychology* 39.6 (1980), pp. 1161–1178. DOI: 10.1037/h0077714 (cited on pages 10, 14, 67, 70).

[165] Björn Schuller, Stefan Steidl, and Anton Batliner. 'The Interspeech 2009 Emotion Challenge'. In: (2009) (cited on page 10).

[166] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. 'AVEC 2011–The First International Audio/Visual Emotion Challenge'. In: *Affective Computing and Intelligent Interaction*. Ed. by Sidney D'Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin. Berlin, Heidelberg: Springer, 2011, pp. 415–424. DOI: 10.1007/978-3-642-24571-8_53 (cited on page 10).

[167] Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. 'AVEC 2012: The Continuous Audio/Visual Emotion Challenge'. In: *Proceedings of the 14th ACM International Conference on Multimodal Interaction*. ICMI '12. New York, NY, USA: Association for Computing Machinery, 2012, pp. 449–456. DOI: 10.1145/2388676.2388776 (cited on page 10).

[168] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. 'Paralinguistics in Speech and Language—State-of-the-art and the Challenge'. In: *Computer Speech & Language*. Special Issue on Paralinguistics in Naturalistic Speech and Language 27.1 (2013), pp. 4–39. DOI: 10.1016/j.csl.2012.02.005 (cited on page 10).

[169] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 'AVEC 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge'. In: *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*. AVEC '13. New York, NY, USA: Association for Computing Machinery, 2013, pp. 3–10. DOI: 10.1145/2512530.2512533 (cited on page 10).

[170] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 'AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge'. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. AVEC '14. New York, NY, USA: Association for Computing Machinery, 2014, pp. 3–10. DOI: 10.1145/2661806.2661807 (cited on page 10).

[171] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 'AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge'. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. AVEC '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 3–10. DOI: 10.1145/2988257.2988258 (cited on page 10).

[172] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, Heysem Kaya, Maximilian Schmitt, Shahin Amiriparian, Nicholas Cummins, Denis Lalanne, Adrien Michaud, Elvan Ciftçi, Hüseyin Güleç, Albert Ali Salah, and Maja Pantic. 'AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition'. In: *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. AVEC'18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 3–13. DOI: 10.1145/3266302.3266316 (cited on page 10).

[173] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic. 'AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition'. In: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. AVEC '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 3–12. DOI: 10.1145/3347320.3357688 (cited on page 10).

[174] Gemini Team et al. *Gemini: A Family of Highly Capable Multimodal Models*. 2024. DOI: 10.48550/arXiv.2312.11805. Pre-published (cited on page 10).

[175] Yichong Leng, Zhifang Guo, Kai Shen, Xu Tan, Zeqian Ju, Yanqing Liu, Yufei Liu, Dongchao Yang, Leying Zhang, Kaitao Song, Lei He, Xiang-Yang Li, Sheng Zhao, Tao Qin, and Jiang Bian. *PromptTTS 2: Describing and Generating Voices with Text Prompt*. 2023. DOI: 10.48550/arXiv.2309.02285. Pre-published (cited on page 10).

[176] Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Haohan Guo, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Zhou Zhao, Xixin Wu, and Helen M. Meng. 'UniAudio: Towards Universal Audio Generation with Large Language Models'. In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp. Vol. 235. Proceedings of Machine Learning Research. PMLR, 2024, pp. 56422–56447 (cited on page 10).

[177] Björn Schuller, Adria Mallol-Ragolta, Alejandro Peña Almansa, Iosif Tsangko, Mostafa M. Amin, Anastasia Semertzidou, Lukas Christ, and Shahin Amiriparian. *Affective Computing Has Changed: The Foundation Model Disruption*. 2024. DOI: 10.48550/arXiv.2409.08907. Pre-published (cited on page 10).

[178] Aristotle. *Art of Rhetoric*. Cambridge, MA: Harvard University Press, 1926 (cited on page 10).

[179] William James. 'What Is an Emotion?' In: *Mind* 9.34 (1884), pp. 188–205 (cited on page 10).

[180] W. B. Cannon. 'The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory'. In: *The American Journal of Psychology* 39 (1927), pp. 106–124. DOI: 10.2307/1415404 (cited on page 11).

[181] Stanley Schachter and Jerome Singer. 'Cognitive, Social, and Physiological Determinants of Emotional State'. In: *Psychological Review* 69.5 (1962), pp. 379–399. DOI: 10.1037/h0046234 (cited on page 11).

[182] Richard S. Lazarus PhD and Susan Folkman PhD. *Stress, Appraisal, and Coping*. Springer Publishing Company, 1984. 460 pp. (cited on page 11).

[183] Richard S. Lazarus. *Emotion and Adaptation*. Emotion and Adaptation. New York, NY, US: Oxford University Press, 1991, pp. xiii, 557. xiii, 557 (cited on page 11).

[184] Klaus R. Scherer. 'Vocal Affect Expression: A Review and a Model for Future Research.' In: *Psychological Bulletin* 99.2 (1986), pp. 143–165. DOI: 10/cwm6bn (cited on pages 11, 12, 30).

[185] James A. Russell, Jo-Anne Bachorowski, and José-Miguel Fernández-Dols. 'Facial and Vocal Expressions of Emotion'. In: *Annual Review of Psychology* 54.1 (2003), pp. 329–349. DOI: 10/ftfxdn (cited on pages 11, 13, 30).

[186] A. Moors. 'Comparison of Four Families of Psychological Emotion Theories'. In: *The Routlegde Handbook of Emotion Theory*. Ed. by A. Scarantino. New York: Taylor & Francis/Routledge, 2020 (cited on page 11).

[187] Paul Ekman and Wallace V. Friesen. 'Constants across Cultures in the Face and Emotion'. In: *Journal of Personality and Social Psychology* 17.2 (1971), pp. 124–129. DOI: 10.1037/h0030377 (cited on page 11).

[188] Hillary Anger Elfenbein and Nalini Ambady. 'Is There an In-Group Advantage in Emotion Recognition?' In: *Psychological Bulletin* 128.2 (2002), pp. 243–249. DOI: 10/dcmqvp (cited on pages 12, 14, 35).

[189] Hillary Anger Elfenbein, Martin Beaupré, Manon Lévesque, and Ursula Hess. 'Toward a Dialect Theory: Cultural Differences in the Expression and Recognition of Posed Facial Expressions.' In: *Emotion* 7.1 (2007), pp. 131–146. DOI: 10/dfpzdq (cited on pages 12, 34, 77).

[190] Petri Laukka, Daniel Neiberg, and Hillary Anger Elfenbein. 'Evidence for Cultural Dialects in Vocal Emotion Expression: Acoustic Classification within and across Five Nations.' In: *Emotion* 14.3 (2014), pp. 445–449. DOI: 10.1037/a0036048 (cited on pages 12, 29).

[191] Petri Laukka and Hillary Anger Elfenbein. 'Cross-Cultural Emotion Recognition and In-Group Advantage in Vocal Expression: A Meta-Analysis'. In: *Emotion Review* (2020), p. 175407391989729. DOI: 10/ggk846 (cited on pages 12, 29, 30, 63).

[192] David Matsumoto. 'Cultural Similarities and Differences in Display Rules'. In: *Motivation and Emotion* 14.3 (1990), pp. 195–214. DOI: 10/b5hctg (cited on pages 12, 14).

[193] Christina M. Moran, James M. Diefendorff, and Gary J. Greguras. 'Understanding Emotional Display Rules at Work and Outside of Work: The Effects of Country and Gender'. In: *Motivation and Emotion* 37.2 (2013), pp. 323–334. DOI: 10.1007/s11031-012-9301-x (cited on page 12).

[194] Dolichan Kollareth, Jose-Miguel Fernandez-Dols, and James Russell. 'Shame as a Culture-Specific Emotion Concept'. In: *Journal of Cognition and Culture* 18 (Aug. 2018), pp. 274–292. DOI: 10.1163/15685373-12340031 (cited on page 12).

[195] Agnes Moors, Phoebe C. Ellsworth, Klaus R. Scherer, and Nico H. Frijda. 'Appraisal Theories of Emotion: State of the Art and Future Development'. In: *Emotion Review* 5.2 (2013), pp. 119–124. DOI: 10.1177/1754073912468165 (cited on page 12).

[196] Klaus R. Scherer. 'The Dynamic Architecture of Emotion: Evidence for the Component Process Model'. In: *Cognition and Emotion* 23.7 (2009), pp. 1307–1351. DOI: 10.1080/02699930902928969 (cited on page 12).

[197] Klaus R Scherer. 'Appraisal Considered as a Process of Multilevel Sequential Checking'. In: *Appraisal Processes in Emotion: Theory, Methods, Research*. Ed. by Klaus R Scherer, Angela Schorr, and Tom Johnstone. Oxford University Press, 2001. DOI: 10.1093/oso/9780195130072.003.0005 (cited on page 12).

[198] Andrew Ortony, Gerald L. Clore, and Allan Collins. *The Cognitive Structure of Emotions*. Cambridge, UK: Cambridge University Press, 1988 (cited on page 13).

[199] Klaus R. Scherer. 'The Role of Culture in Emotion-Antecedent Appraisal'. In: *Journal of Personality and Social Psychology* 73.5 (1997), pp. 902–922. DOI: 10.1037/0022-3514.73.5.902 (cited on pages 13, 14).

[200] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 'Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements'. In: *Psychological Science in the Public Interest* 20.1 (2019), pp. 1–68. DOI: 10/gf4772 (cited on pages 13, 38, 50).

[201] Wilhelm Wundt. *Grundzüge der physiologischen Psychologie*. Leipzig: Wilhelm Engelmann, 1874 (cited on page 13).

[202] Albert Mehrabian and James A. Russell. *An Approach to Environmental Psychology*. Cambridge, MA: MIT Press, 1974 (cited on page 14).

[203] Albert Mehrabian. 'Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament'. In: *Current Psychology* 14.4 (1996), pp. 261–292. DOI: 10.1007/BF02686918 (cited on page 14).

[204] Andy Clark. 'Whatever next? Predictive Brains, Situated Agents, and the Future of Cognitive Science'. In: *The Behavioral and Brain Sciences* 36.3 (2013), pp. 181–204. DOI: 10.1017/S0140525X12000477 (cited on page 14).

[205] Lisa Feldman Barrett. 'The Theory of Constructed Emotion: An Active Inference Account of Interoception and Categorization'. In: *Social Cognitive and Affective Neuroscience* 12.1 (2017), pp. 1–23. DOI: 10.1093/scan/nsw154 (cited on page 14).

[206] Lisa Feldman Barrett and W. Kyle Simmons. 'Interoceptive Predictions in the Brain'. In: *Nature Reviews Neuroscience* 16.7 (2015), pp. 419–429. DOI: 10.1038/nrn3950 (cited on page 14).

[207] Anil K. Seth and Karl J. Friston. 'Active Interoceptive Inference and the Emotional Brain'. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 371.1708 (2016), p. 20160007. DOI: 10.1098/rstb.2016.0007 (cited on page 14).

[208] Alan S. Cowen and Dacher Keltner. 'What the Face Displays: Mapping 28 Emotions Conveyed by Naturalistic Expression'. In: *The American psychologist* 75.3 (2020), pp. 349–364. DOI: 10.1037/amp0000488 (cited on pages 14, 15, 22, 143).

[209] Guanxiong Pei, Haiying Li, Yandi Lu, Yanlei Wang, Shizhen Hua, and Taihao Li. ‘Affective Computing: Recent Advances, Challenges, and Future Trends’. In: *Intelligent Computing* 3 (2024), p. 0076. DOI: 10.34133/icomputing.0076 (cited on page 15).

[210] Alan S. Cowen and Dacher Keltner. ‘Semantic Space Theory: A Computational Approach to Emotion’. In: *Trends in Cognitive Sciences* (2020). DOI: 10/ghqg72 (cited on page 15).

[211] Alan Cowen, Disa Sauter, Jessica L. Tracy, and Dacher Keltner. ‘Mapping the Passions: Toward a High-Dimensional Taxonomy of Emotional Experience and Expression’. In: *Psychological Science in the Public Interest* 20.1 (2019), pp. 69–90. DOI: 10/gf479w (cited on page 15).

[212] Alan S. Cowen, Petri Laukka, Hillary Anger Elfenbein, Runjing Liu, and Dacher Keltner. ‘The Primacy of Categories in the Recognition of 12 Emotions in Speech Prosody across Two Cultures’. In: *Nature Human Behaviour* 3.4 (2019), p. 369. DOI: 10/gfw2hr (cited on pages 15, 16, 30, 52, 63, 143).

[213] Alan S. Cowen, Hillary Anger Elfenbein, Petri Laukka, and Dacher Keltner. ‘Mapping 24 Emotions Conveyed by Brief Human Vocalization’. In: *The American Psychologist* 74.6 (2019), pp. 698–712. DOI: 10.1037/amp0000399 (cited on pages 15, 16, 143).

[214] Jeffrey A. Brooks, Panagiotis Tzirakis, Alice Baird, Lauren Kim, Michael Opara, Xia Fang, Dacher Keltner, Maria Monroy, Rebecca Corona, Jacob Metrick, and Alan S. Cowen. ‘Deep Learning Reveals What Vocal Bursts Express in Different Cultures’. In: *Nature Human Behaviour* 7.2 (2023), pp. 240–250. DOI: 10.1038/s41562-022-01489-2 (cited on pages 15, 16, 143).

[215] Alan S. Cowen, Dacher Keltner, Florian Schroff, Brendan Jou, Hartwig Adam, and Gautam Prasad. ‘Sixteen Facial Expressions Occur in Similar Contexts Worldwide’. In: *Nature* (2020), pp. 1–7. DOI: 10/ghp7fg (cited on pages 15, 143).

[216] Eftychia Stamkou, Dacher Keltner, Rebecca Corona, Eda Aksoy, and Alan S. Cowen. ‘Emotional Palette: A Computational Mapping of Aesthetic Experiences Evoked by Visual Art’. In: *Scientific Reports* 14.1 (2024), p. 19932. DOI: 10.1038/s41598-024-69686-9 (cited on pages 15, 143).

[217] L.J.P. van der Maaten and G.E. Hinton. ‘Visualizing High-Dimensional Data Using t-SNE’. In: *Journal of Machine Learning Research* 9 (nov 2008), pp. 2579–2605 (cited on page 16).

[218] Lisa Feldman Barrett, Zulqarnain Khan, Jennifer Dy, and Dana Brooks. ‘Nature of Emotion Categories: Comment on Cowen and Keltner’. In: *Trends in cognitive sciences* 22.2 (2018), pp. 97–99. DOI: 10.1016/j.tics.2017.12.004 (cited on page 16).

[219] Alan S. Cowen and Dacher Keltner. ‘Clarifying the Conceptualization, Dimensionality, and Structure of Emotion: Response to Barrett and Colleagues’. In: *Trends in Cognitive Sciences* 22.4 (2018), pp. 274–276. DOI: 10/gfj8v5 (cited on page 16).

[220] Nori Jacoby et al. ‘Commonality and Variation in Mental Representations of Music Revealed by a Cross-Cultural Comparison of Rhythm Priors in 15 Countries’. In: *Nature Human Behaviour* 8.5 (2024), pp. 846–877. DOI: 10.1038/s41562-023-01800-9 (cited on pages 18, 23, 136).

[221] Akiyoshi Kitaoka. *Grey Strawberries*. The Illusions Index. 2024. URL: https://www.illusionsindex.org/i/grey-strawberries (cited on page 18).

[222] Jing Xu and Thomas L. Griffiths. ‘A Rational Analysis of the Effects of Memory Biases on Serial Reproduction’. In: *Cognitive Psychology* 60.2 (2010), pp. 107–126. DOI: 10.1016/j.cogpsych.2009.09.002 (cited on page 18).

[223] G. Kochanski. ‘Using Mimicry to Learn about Phonology’. In: 2008 (cited on page 20).

[224] Alberto Acerbi and Joseph M. Stubbersfield. ‘Large Language Models Show Human-like Content Biases in Transmission Chain Experiments’. In: *Proceedings of the National Academy of Sciences* 120.44 (2023). DOI: 10.1073/pnas.2313790120 (cited on page 20).

[225] Simon Kirby, Tom Griffiths, and Kenny Smith. ‘Iterated Learning and the Evolution of Language’. In: *Current Opinion in Neurobiology*. SI: Communication and Language 28 (2014), pp. 108–114. DOI: 10.1016/j.conb.2014.07.014 (cited on page 21).

[226] Paul Kay and Richard S. Cook. 'World Color Survey'. In: *Encyclopedia of Color Science and Technology*. Ed. by Ming Ronnier Luo. New York, NY: Springer New York, 2016, pp. 1265–1271. DOI: 10.1007/978-1-4419-8071-7_113 (cited on pages 21, 102).

[227] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 'Equation of State Calculations by Fast Computing Machines'. In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092. DOI: 10.1063/1.1699114 (cited on page 21).

[228] W. Keith Hastings. 'Monte Carlo Sampling Methods Using Markov Chains and Their Applications'. In: *Biometrika* 57.1 (1970), pp. 97–109. DOI: 10.1093/biomet/57.1.97 (cited on page 21).

[229] Jingjun Liang, Shizhe Chen, Jinming Zhao, Qin Jin, Haibo Liu, and Li Lu. 'Cross-culture Multimodal Emotion Recognition with Adversarial Learning'. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2019, pp. 4000–4004. DOI: 10.1109/icassp.2019.8683725 (cited on page 22).

[230] Silvan Mertes, Dominik Schiller, Florian Lingenfelser, Thomas Kiderle, Valentin Kroner, Lama Diab, and Elisabeth André. 'Intercategorical Label Interpolation for Emotional Face Generation with Conditional Generative Adversarial Networks'. In: *Deep Learning Theory and Applications*. Springer Nature Switzerland, 2023, pp. 67–87. DOI: 10.1007/978-3-031-37320-6_4 (cited on page 22).

[231] Juan I. Durán, Rainer Reisenzein, and José-Miguel Fernández-Dols. *Coherence Between Emotions and Facial Expressions*. Vol. 1. Oxford University Press, 2017 (cited on pages 22–24, 27, 136).

[232] Erika H. Siegel, Molly K. Sands, Wim Van den Noortgate, Paul Condon, Yale Chang, Jennifer Dy, Karen S. Quigley, and Lisa Feldman Barrett. 'Emotion Fingerprints or Emotion Populations? A Meta-Analytic Investigation of Autonomic Features of Emotion Categories.' In: *Psychological Bulletin* 144.4 (2018), pp. 343–393. DOI: 10/gdbbpd (cited on pages 22, 24, 27, 38, 39).

[233] D. A. Sauter, F. Eisner, P. Ekman, and S. K. Scott. 'Cross-Cultural Recognition of Basic Emotions through Nonverbal Emotional Vocalizations'. In: *Proceedings of the National Academy of Sciences* 107.6 (2010), pp. 2408–2412. DOI: 10/dcr38b (cited on pages 22–24, 27, 63, 136).

[234] Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 'Different Languages, Similar Encoding Efficiency: Comparable Information Rates across the Human Communicative Niche'. In: *Science Advances* 5.9 (2019). DOI: 10.1126/sciadv.aaw2594 (cited on page 23).

[235] Tor Norretranders. *The User Illusion*. Penguin Press Science S. Harlow, England: Penguin Books, 1999 (cited on page 23).

[236] Benjamin Lee Whorf. *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. 1956 (cited on page 23).

[237] Asifa Majid, Melissa Bowerman, Sotaro Kita, Daniel B.M. Haun, and Stephen C. Levinson. 'Can Language Restructure Cognition? The Case for Space'. In: *Trends in Cognitive Sciences* 8.3 (2004), pp. 108–114. DOI: 10.1016/j.tics.2004.01.003 (cited on pages 23–25, 104).

[238] Mark Lowry and Judith Bryant. 'Blue Is in the Eye of the Beholder: A Cross-Linguistic Study on Color Perception and Memory'. In: *Journal of Psycholinguistic Research* 48.1 (2019), pp. 163–179. DOI: 10.1007/s10936-018-9597-0 (cited on pages 23, 24, 103).

[239] Jules Davidoff, Ian Davies, and Debi Roberson. 'Colour Categories in a Stone-Age Tribe'. In: *Nature* 398.6724 (1999), pp. 203–204. DOI: 10.1038/18335 (cited on pages 23, 24, 103).

[240] Laura J. Speed and Asifa Majid. 'Grounding Language in the Neglected Senses of Touch, Taste, and Smell'. In: *Cognitive Neuropsychology* 37.5–6 (2019), pp. 363–392. DOI: 10.1080/02643294.2019.1623188 (cited on pages 23, 136).

[241] Anthony Diller. 'Cross-Cultural Pain Semantics'. In: *Pain* 9.1 (1980), pp. 9–26. DOI: 10.1016/0304-3959(80)90025-1 (cited on pages 23, 136).

[242] Damián E. Blasi, Søren Wichmann, Harald Hammarström, Peter F. Stadler, and Morten H. Christiansen. 'Sound–Meaning Association Biases Evidenced across Thousands of Languages'. In: *Proceedings of the National Academy of Sciences* 113.39 (2016), pp. 10818–10823. DOI: 10.1073/pnas.1605782113 (cited on pages 23, 136).

[243] Asifa Majid and Niclas Burenhult. 'Odors Are Expressible in Language, as Long as You Speak the Right Language'. In: *Cognition* 130.2 (2014), pp. 266–270. DOI: 10.1016/j.cognition.2013.11.004 (cited on pages 23, 25, 104, 136).

[244] K. Ikeda. 'New Seasonings'. In: *Chemical Senses* 27.9 (2002), pp. 847–849. DOI: 10.1093/chemse/27.9.847 (cited on pages 23, 136).

[245] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 'The Weirdest People in the World?' In: *Behavioral and Brain Sciences* 33.2-3 (2010), pp. 61–83. DOI: 10.1017/S0140525X0999152X (cited on page 23).

[246] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 'Most People Are Not WEIRD'. In: *Nature* 466.7302 (2010), pp. 29–29. DOI: 10.1038/466029a (cited on pages 24, 50, 63, 77, 85, 131).

[247] Joseph Henrich. 'WEIRD'. In: *Open Encyclopedia of Cognitive Science*. MIT Press, 2024. DOI: 10.21428/e2759450.8e9a83b0 (cited on pages 24, 77).

[248] Melissa A. Ilardo, Ida Moltke, Thorfinn S. Korneliussen, Jade Cheng, Aaron J. Stern, Fernando Racimo, Peter de Barros Damgaard, Martin Sikora, Andaine Seguin-Orlando, Simon Rasmussen, Inge C. L. van den Munckhof, Rob ter Horst, Leo A. B. Joosten, Mihai G. Netea, Suhartini Salingkat, Rasmus Nielsen, and Eske Willerslev. 'Physiological and Genetic Adaptations to Diving in Sea Nomads'. In: *Cell* 173.3 (2018), 569–580.e15. DOI: 10.1016/j.cell.2018.03.054 (cited on page 24).

[249] Asifa Majid. 'Human Olfaction at the Intersection of Language, Culture, and Biology'. In: *Trends in Cognitive Sciences* 25.2 (2021), pp. 111–123. DOI: 10.1016/j.tics.2020.11.005 (cited on page 24).

[250] Helen E. Davis and Elizabeth Cashdan. 'Spatial Cognition, Navigation, and Mobility among Children in a Forager-Horticulturalist Population, the Tsimané of Bolivia'. In: *Cognitive Development* 52 (2019), p. 100800. DOI: 10.1016/j.cogdev.2019.100800 (cited on page 24).

[251] Armin Falk, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 'Global Evidence on Economic Preferences*'. In: *The Quarterly Journal of Economics* 133.4 (2018), pp. 1645–1692. DOI: 10.1093/qje/qjy013 (cited on page 24).

[252] L. R. Goldberg. 'The Structure of Phenotypic Personality Traits'. In: *The American Psychologist* 48.1 (1993), pp. 26–34. DOI: 10.1037//0003-066x.48.1.26 (cited on page 24).

[253] Michael Gurven, Christopher von Rueden, Maxim Massenkoff, Hillard Kaplan, and Marino Lero Vie. 'How Universal Is the Big Five? Testing the Five-Factor Model of Personality Variation among Forager–Farmers in the Bolivian Amazon.' In: *Journal of Personality and Social Psychology* 104.2 (2013), pp. 354–370. DOI: 10.1037/a0030841 (cited on page 24).

[254] Jonathan Haidt and Craig Joseph. 'Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues'. In: *Daedalus* 133.4 (2004), pp. 55–66. DOI: 10.1162/0011526042365555 (cited on page 24).

[255] Mohammad Atari, Jesse Graham, and Morteza Dehghani. 'Foundations of Morality in Iran'. In: *Evolution and Human Behavior*. Beyond Weird 41.5 (2020), pp. 367–384. DOI: 10.1016/j.evolhumbehav.2020.07.014 (cited on page 24).

[256] H. Clark Barrett, Alexander Bolyanatz, Alyssa N. Crittenden, Daniel M. T. Fessler, Simon Fitzpatrick, Michael Gurven, Joseph Henrich, Martin Kanovsky, Geoff Kushnick, Anne Pisor, Brooke A. Scelza, Stephen Stich, Chris von Rueden, Wanying Zhao, and Stephen Laurence. 'Small-Scale Societies Exhibit Fundamental Variation in the Role of Intentions in Moral Judgment'. In: *Proceedings of the National Academy of Sciences* 113.17 (2016), pp. 4688–4693. DOI: 10.1073/pnas.1522070113 (cited on page 24).

[257] Michael D. Gurven and Daniel E. Lieberman. 'WEIRD Bodies: Mismatch, Medicine and Missing Diversity'. In: *Evolution and Human Behavior*. Beyond Weird 41.5 (2020), pp. 330–340. DOI: 10.1016/j.evolhumbehav.2020.04.001 (cited on page 24).

[258] Mohammad Atari, Mona J. Xue, Peter S. Park, Damián Blasi, and Joseph Henrich. *Which Humans?* 2024. URL: https://osf.io/5b26t. Pre-published (cited on page 24).

[259] Damián E. Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. 'Over-Reliance on English Hinders Cognitive Science'. In: *Trends in Cognitive Sciences* 26.12 (2022), pp. 1153–1170. DOI: 10.1016/j.tics.2022.09.015 (cited on pages 24, 77, 104, 108, 136).

[260] Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. *Glottolog/Glottolog-Cldf: Glottolog Database 4.8 as CLDF*. Version v4.8. Zenodo, 2023. DOI: 10.5281/zenodo.8131091 (cited on pages 24, 77).

[261] Hedvig Skirgård et al. 'Grambank Reveals the Importance of Genealogical Constraints on Linguistic Diversity and Highlights the Impact of Language Loss'. In: *Science Advances* 9.16 (2023), eadg6175. DOI: 10.1126/sciadv.adg6175 (cited on pages 24, 77, 96).

[262] Barbara Tversky. 'Visualizing Thought'. In: *Handbook of Human Centric Visualization*. Ed. by Weidong Huang. New York, NY: Springer, 2014, pp. 3–40. DOI: 10.1007/978-1-4614-7485-2_1 (cited on pages 24, 25).

[263] Qi Wang. 'The Cultural Foundation of Human Memory'. In: *Annual Review of Psychology* 72.1 (2021), pp. 151–179. DOI: 10.1146/annurev-psych-070920-023638 (cited on pages 24, 25, 104).

[264] Federica Amici, Alex Sánchez-Amaro, Carla Sebastián-Enesco, Trix Cacchione, Matthias Allritz, Juan Salazar-Bonet, and Federico Rossano. 'The Word Order of Languages Predicts Native Speakers' Working Memory'. In: *Scientific Reports* 9.1 (2019), p. 1124. DOI: 10.1038/s41598-018-37654-9 (cited on page 25).

[265] Daniel B. M. Haun, Christian J. Rapold, Josep Call, Gabriele Janzen, and Stephen C. Levinson. 'Cognitive Cladistics and Cultural Override in Hominid Spatial Cognition'. In: *Proceedings of the National Academy of Sciences* 103.46 (2006), pp. 17568–17573. DOI: 10.1073/pnas.0607999103 (cited on page 25).

[266] Felix K. Ameka and Marina Terkourafi. 'What If…? Imagining Non-Western Perspectives on Pragmatic Theory and Practice'. In: *Journal of Pragmatics*. Quo Vadis, Pragmatics? 145 (2019), pp. 72–82. DOI: 10.1016/j.pragma.2019.04.001 (cited on page 25).

[267] Jin Li. *Cultural Foundations of Learning: East and West*. Cambridge: Cambridge University Press, 2012 (cited on page 25).

[268] Simeon Floyd. 'Conversation and Culture'. In: *Annual Review of Anthropology* 50 (Volume 50, 2021 2021), pp. 219–240. DOI: 10.1146/annurev-anthro-101819-110158 (cited on page 25).

[269] We Are Social & Meltwater. *Digital: Global Overview Report 2024* (cited on pages 25, 77).

[270] Asifa Majid. 'Establishing Psychological Universals'. In: *Nature Reviews Psychology* 2.4 (2023), pp. 199–200. DOI: 10.1038/s44159-023-00169-w (cited on page 25).

[271] Silke Paulmann and Marc D. Pell. 'Is There an Advantage for Recognizing Multi-Modal Emotional Stimuli?' In: *Motivation and Emotion* 35.2 (2011), pp. 192–201. DOI: 10/c3btn2 (cited on pages 29, 53, 63).

[272] Silke Paulmann and Ayse K. Uskul. 'Cross-Cultural Emotional Prosody Recognition: Evidence from Chinese and British Listeners'. In: *Cognition and Emotion* 28.2 (2013), pp. 230–244. DOI: 10/ggcgqr (cited on pages 29, 63).

[273] Timo Kevin Koch, Gabriella M. Harari, Ramona Schoedel, Samuel D. Gosling, Zachariah Marrero, Florian Bemmann, Markus Bühner, and Clemens Stachl. 'Semantic Content Outperforms Speech Prosody in Predicting Affective Experience in Naturalistic Settings'. In: (2024). DOI: 10.31234/osf.io/n48uz (cited on pages 29, 53).

[274] Eva Navas, Inmaculada Hernáez, Amaia Castelruiz, and Iker Luengo. 'Obtaining and Evaluating an Emotional Database for Prosody Modelling in Standard Basque'. In: *Text, Speech and Dialogue*. Ed. by Petr Sojka, Ivan Kopeček, and Karel Pala. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2004, pp. 393–400. DOI: 10/bzdvmt (cited on pages 29, 30).

[275] Ibon Saratxaga, Eva Navas, Inmaculada Hernáez, and Iker Luengo. 'Designing and Recording an Emotional Speech Database for Corpus Based Synthesis in Basque'. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). European Language Resources Association (ELRA), 2006, p. 4 (cited on pages 29, 30).

[276] Philippe Gournay, Olivier Lahaie, and Roch Lefebvre. 'A Canadian French Emotional Speech Dataset'. In: *Proceedings of the 9th ACM Multimedia Systems Conference*. MMSys '18: 9th ACM Multimedia Systems Conference. Amsterdam Netherlands: ACM, 2018, pp. 399–402. DOI: 10/gj4hht (cited on pages 29, 30).

[277] Alberto Battocchi, Fabio Pianesi, and Dina Goren-Bar. 'DaFEx: Database of Facial Expressions'. In: *Intelligent Technologies for Interactive Entertainment*. Ed. by Mark Maybury, Oliviero Stock, and Wolfgang Wahlster. Red. by David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, and Gerhard Weikum. Vol. 3814. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 303–306. DOI: 10.1007/11590323_39 (cited on pages 29, 30).

[278] Imene Hadjadji, Leila Falek, Lyes Demri, and Hocine Teffahi. 'Emotion Recognition in Arabic Speech'. In: *2019 International Conference on Advanced Electrical Engineering (ICAEE)*. 2019 International Conference on Advanced Electrical Engineering (ICAEE). 2019, pp. 1–5. DOI: 10.1109/ICAEE47123.2019.9014809 (cited on pages 29, 30).

[279] Felix Burkhardt, A Paeschke, M Rolfes, W Sendlmeier, and B Weiss. 'A Database of German Emotional Speech'. In: INTERSPEECH. Lisbon, Portugal, 2005, p. 4 (cited on pages 29, 30).

[280] Rene Altrov and Hille Pajupuu. 'Estonian Emotional Speech Corpus: Theoretical Base and Implementation'. In: Conference: 4th International Workshop on Corpora for Research on EMOTION SENTIMENT & SOCIAL SIGNALS ESÉ Istanbul, Turkey, 2012 (cited on pages 29, 30).

[281] Leanne Nagels, Etienne Gaudrain, Debi Vickers, Marta Matos Lopes, Petra Hendriks, and Deniz BaÉkent. *Vocal Emotion Recognition in School-Age Children: Normative Data for the EmoHI Test*. preprint. PeerJ Preprints, 2019. DOI: 10.7287/peerj.preprints.27921v1 (cited on pages 29, 30).

[282] Sungbok Lee, Serdar Yildirim, Abe Kazemzadeh, and Shrikanth Narayanan. 'An Articulatory Study of Emotional Speech Production'. In: INTERSPEECH. Lisbon, Portugal, 2005, p. 4 (cited on pages 29, 30).

[283] O. Martin, I. Kotsia, B. Macq, and I. Pitas. 'The eNTERFACE&#146;05 Audio-Visual Emotion Database'. In: *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. 22nd International Conference on Data Engineering Workshops (ICDEW'06). Atlanta, GA, USA: IEEE, 2006, pp. 8–8. DOI: 10/cckzj4 (cited on pages 29, 30).

[284] Fay. Ykhlef, A. Derbal, W. Benzaba, R. Boutaleb, D. Bouchaffra, H. Meraoubi, and Far. Ykhlef. 'Towards Building an Emotional Speech Corpus of Algerian Dialect: Criteria and Preliminary Assessment Results'. In: *2019 International Conference on Advanced Electrical Engineering (ICAEE)*. 2019 International Conference on Advanced Electrical Engineering (ICAEE). 2019, pp. 1–6. DOI: 10/ghjxg4 (cited on pages 29, 30).

[285] Skyler T. Hawk, Gerben A. van Kleef, Agneta H. Fischer, and Job van der Schalk. '"Worth a Thousand Words": Absolute and Relative Decoding of Nonlinguistic Affect Vocalizations.' In: *Emotion* 9.3 (2009), pp. 293–305. DOI: 10/d8k9ps (cited on pages 29, 30).

[286] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 'MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception'. In: *IEEE Transactions on Affective Computing* 8.1 (2017), pp. 67–80. DOI: 10/f9t4sr (cited on pages 29, 30).

[287] Marc D. Pell, Silke Paulmann, Chinar Dara, Areej Alasseri, and Sonja A. Kotz. 'Factors in the Recognition of Vocally Expressed Emotions: A Comparison of Four Languages'. In: *Journal of Phonetics* 37.4 (2009), pp. 417–435. DOI: 10/bwzhnd (cited on pages 29, 30, 52, 63).

[288] Shashidhar G. Koolagudi, Sudhamay Maity, Vuppala Anil Kumar, Saswat Chakrabarti, and K. Sreenivasa Rao. 'IITKGP-SESC: Speech Database for Emotion Analysis'. In: *Contemporary Computing*. Ed. by Sanjay Ranka, Srinivas Aluru, Rajkumar Buyya, Yeh-Ching Chung, Sumeet Dua, Ananth Grama, Sandeep K. S. Gupta, Rajeev Kumar, and Vir V. Phoha. Communications in Computer and Information Science. Berlin, Heidelberg: Springer, 2009, pp. 485–492. DOI: 10/ckb3nt (cited on pages 29–31, 53).

[289] Shashidhar G. Koolagudi, Ramu Reddy, Jainath Yadav, and K. Sreenivasa Rao. 'IITKGP-SEHSC : Hindi Speech Corpus for Emotion Analysis'. In: *2011 International Conference on Devices and Communications (ICDeCom)*. 2011 International Conference on Devices and Communications (ICDeCom). 2011, pp. 1–5. DOI: 10/c9xq5m (cited on pages 29, 30, 53).

[290] Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. 'The Emotional Voices Database: Towards Controlling the Emotion Dimension in Voice Generation Systems'. 2018 (cited on pages 29, 30).

[291] Petri Laukka and Hillary Anger Elfenbein. 'Emotion Appraisal Dimensions Can Be Inferred From Vocal Expressions'. In: *Social Psychological and Personality Science* 3.5 (2011), pp. 529–536. DOI: 10/bp2sdh (cited on page 29).

[292] Yannick Jadoul, Bill Thompson, and Bart de Boer. 'Introducing Parselmouth: A Python Interface to Praat'. In: *Journal of Phonetics* 71 (2018), pp. 1–15. DOI: 10/ggc7w8 (cited on pages 31, 44, 121).

[293] William Revelle. *Psych: Procedures for Psychological, Psychometric, and Personality Research*. manual. Evanston, Illinois: Northwestern University, 2022 (cited on page 31).

[294] Paul-Christian Bürkner. 'Advanced Bayesian Multilevel Modeling with the R Package Brms'. In: *The R Journal* 10.1 (2018), p. 395. DOI: 10/gfxzpn (cited on page 32).

[295] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. '*Stan* : A Probabilistic Programming Language'. In: *Journal of Statistical Software* 76.1 (2017). DOI: 10/b2pm (cited on page 32).

[296] James M. Diefendorff and Erin M. Richard. 'Antecedents and Consequences of Emotional Display Rule Perceptions.' In: *Journal of Applied Psychology* 88.2 (2003), pp. 284–294. DOI: 10.1037/0021-9010.88.2.284 (cited on page 38).

[297] Annett Schirmer and Ralph Adolphs. 'Emotion Perception from Face, Voice, and Touch: Comparisons and Convergence'. In: *Trends in Cognitive Sciences* 21.3 (2017), pp. 216–228. DOI: 10/f9xkqq (cited on page 38).

[298] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. 'How to Find Trouble in Communication'. In: *Speech Communication* 40.1 (2003), pp. 117–143. DOI: 10/cm69cn (cited on page 38).

[299] Andrey Anikin and César F. Lima. 'Perceptual and Acoustic Differences between Authentic and Acted Nonverbal Emotional Vocalizations'. In: *Quarterly Journal of Experimental Psychology* (2017), p. 17470218.2016.1. DOI: 10.1080/17470218.2016.1270976 (cited on pages 38, 41, 50).

[300] Doron Atias and Hillel Aviezer. 'Real-Life and Posed Vocalizations to Lottery Wins Differ Fundamentally in Their Perceived Valence'. In: *Emotion* (2020), No Pagination Specified–No Pagination Specified. DOI: 10/ghm4df (cited on page 38).

[301] T. Vogt and E. André. 'Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition'. In: *2005 IEEE International Conference on Multimedia and Expo*. 2005 IEEE International Conference on Multimedia and Expo. 2005, pp. 474–477. DOI: 10.1109/ICME.2005.1521463 (cited on page 41).

[302] Roger N Shepard. 'Multidimensional Scaling, Tree-Fitting, and Clustering'. In: *Science* 210.4468 (1980), pp. 390–398 (cited on page 41).

[303] Roger N Shepard. 'Toward a Universal Law of Generalization for Psychological Science'. In: *Science* 237.4820 (1987), pp. 1317–1323 (cited on page 41).

[304] Joshua B Tenenbaum and Thomas L Griffiths. 'Generalization, Similarity, and Bayesian Inference'. In: *Behavioral and brain sciences* 24.4 (2001), pp. 629–640 (cited on page 41).

[305] Amos Tversky. 'Features of Similarity.' In: *Psychological review* 84.4 (1977), p. 327 (cited on page 41).

[306] Ron Dotsch and Alexander Todorov. 'Reverse Correlating Social Face Perception'. In: *Social Psychological and Personality Science* 3.5 (2011), pp. 562–571. DOI: 10/bx4tz6 (cited on page 41).

[307] Michael C. Mangini and Irving Biederman. 'Making the Ineffable Explicit: Estimating the Information Employed for Face Classifications'. In: *Cognitive Science* 28.2 (2004), pp. 209–226. DOI: 10/fcb673 (cited on pages 41, 42).

[308] Alan E. Gelfand and Adrian F. M. Smith. 'Sampling-Based Approaches to Calculating Marginal Densities'. In: *Journal of the American Statistical Association* 85.410 (1990), pp. 398–409 (cited on page 42).

[309] George H. Joblove and Donald Greenberg. 'Color Spaces for Computer Graphics'. In: *Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques*. 1978, pp. 20–25 (cited on page 43).

[310] *IEEE Recommended Practice for Speech Quality Measurements*. IEEE. DOI: 10.1109/IEEESTD.1969.7405210 (cited on pages 44, 47, 56, 115, 122).

[311] Philippa Demonte. 'HARVARD Corpus Speech Shaped Noise and Speech Modulated Noise for SIN Test'. In: (2019). DOI: 10/ggx9sp (cited on page 44).

[312] Kevin J. P. Woods, Max H. Siegel, James Traer, and Josh H. McDermott. 'Headphone Screening to Facilitate Web-Based Auditory Experiments'. In: *Attention, Perception, & Psychophysics* 79.7 (2017), pp. 2064–2072. DOI: 10/gbzdxf (cited on pages 45, 56, 80, 116, 123).

[313] Emmanuel Ponsot, Juan José Burred, Pascal Belin, and Jean-Julien Aucouturier. 'Cracking the Social Code of Speech Prosody Using Reverse Correlation'. In: *Proceedings of the National Academy of Sciences* 115.15 (2018), pp. 3972–3977. DOI: 10.1073/pnas.1716090115 (cited on pages 46, 119).

[314] Petri Laukka, Patrik Juslin, and Roberto Bresin. 'A Dimensional Approach to Vocal Expression of Emotion'. In: *Cognition &amp; Emotion* 19.5 (2005), pp. 633–653. DOI: 10.1080/02699930441000445 (cited on page 46).

[315] Jaehyeon Kim, Jungil Kong, and Juhee Son. *VITS Implementation*. 2022. URL: https://github.com/jaywalnut310/vits (cited on pages 46, 114, 115, 120, 121).

[316] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. 'Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis'. In: *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189 (cited on pages 46, 53, 114, 120).

[317] Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 'Tacotron: Towards End-to-End Speech Synthesis'. In: *Interspeech 2017*. 2017, pp. 4006–4010. DOI: 10.21437/Interspeech.2017-1452 (cited on pages 46, 114).

[318] Simon King and Vasilis Karaiskos. 'The Blizzard Challenge 2013'. In: *The Blizzard Challenge 2013*. 2013, pp. 1–12. DOI: 10.21437/Blizzard.2013-1 (cited on page 47).

[319] Patrik N. Juslin and Petri Laukka. 'Impact of Intended Emotion Intensity on Cue Utilization and Decoding Accuracy in Vocal Expression of Emotion.' In: *Emotion* 1.4 (2001), pp. 381–412. DOI: 10/dzptgs (cited on page 51).

[320] Omid Mohamad Nezami, Paria Jamshid Lou, and Mansoureh Karami. 'ShEMO: A Large-Scale Validated Database for Persian Speech Emotion Detection'. In: *Language Resources and Evaluation* 53.1 (2019), pp. 1–16. DOI: 10.1007/s10579-018-9427-x (cited on page 53).

[321] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 'IEMOCAP: Interactive Emotional Dyadic Motion Capture Database'. In: *Language Resources and Evaluation* 42.4 (2008), pp. 335–359. DOI: 10/bcbjzg (cited on page 53).

[322] Anton Batliner, Stefan Steidl, and Elmar Nöth. 'Releasing a Thoroughly Annotated and Processed Spontaneous Emotional Database: The FAU Aibo Emotion Corpus'. In: *Proc. of a Satellite Workshop of LREC 2008 on Corpora for Research on Emotion and Affect* (Marrakesh). Ed. by Laurence Devillers, Jean-Claude Martin, Roddy Cowie, Ellen Douglas-Cowie, and Batliner Anton. Marrakesh: LREC, 2008, pp. 28–31 (cited on page 53).

[323] Lukas Stappen, Alice Baird, Lea Schumann, and Björn Schuller. 'The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements'. In: *IEEE Trans. Affect. Comput.* 14.2 (2023), pp. 1334–1350. DOI: 10.1109/TAFFC.2021.3097002 (cited on page 53).

[324] Thurid Vogt, Elisabeth André, and Johannes Wagner. 'Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation'. In: *Affect and Emotion in Human-Computer Interaction: From Theory to Applications*. Ed. by Christian Peter and Russell Beale. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2008, pp. 75–91. DOI: 10.1007/978-3-540-85099-1_7 (cited on page 53).

[325] Rafael Valle, Kevin J. Shih, Ryan Prenger, and Bryan Catanzaro. 'Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis'. In: *International Conference on Learning Representations*. 2021 (cited on pages 53, 56, 120).

[326] Rafael Valle, Jason Li, Ryan Prenger, and Bryan Catanzaro. 'Mellotron: Multispeaker Expressive Voice Synthesis by Conditioning on Rhythm, Pitch and Global Style Tokens'. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020, pp. 6189–6193. DOI: 10/gg3jcz (cited on pages 53, 120).

[327] Andreas Triantafyllopoulos, Björn W. Schuller, Gökçe İymen, Metin Sezgin, Xiangheng He, Zijiang Yang, Panagiotis Tzirakis, Shuo Liu, Silvan Mertes, Elisabeth André, Ruibo Fu, and Jianhua Tao. 'An Overview of Affective Speech Synthesis and Conversion in the Deep Learning Era'. In: *Proceedings of the IEEE* 111.10 (2023), pp. 1355–1381. DOI: 10.1109/JPROC.2023.3250266 (cited on page 53).

[328] Silvan Mertes, Daksitha Withanage Don, Otto Grothe, Johanna Kuch, Ruben Schlagowski, and Elisabeth André. *VoiceX: A Text-To-Speech Framework for Custom Voices*. Version 1. 2024. DOI: 10.48550/arXiv.2408.12170. (Visited on 02/28/2025). Pre-published (cited on page 54).

[329] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 'Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations'. In: *International Journal of Computer Vision* 123.1 (2017), pp. 32–73 (cited on pages 55, 64, 135).

[330] Robert M. MacCallum, Matthias Mauch, Austin Burt, and Armand M. Leroi. 'Evolution of Music by Public Choice'. In: *Proceedings of the National Academy of Sciences* 109.30 (2012), pp. 12081–12086. DOI: 10.1073/pnas.1203182109 (cited on page 55).

[331] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Irene Kotsia, Alice Baird, Chris Gagne, Chunchang Shao, and Guanyu Hu. 'The 6th Affective Behavior Analysis In-the-Wild (ABAW) Competition'. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024, pp. 4587–4598. DOI: 10.1109/CVPRW63382.2024.00461 (cited on page 55).

[332] Kristin Lemhöfer and Mirjam Broersma. 'Introducing LexTALE: A Quick and Valid Lexical Test for Advanced Learners of English'. In: *Behavior Research Methods* 44.2 (2012), pp. 325–343. DOI: 10.3758/s13428-011-0146-0 (cited on pages 56, 70, 86, 91, 93).

[333] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 'Robust Speech Recognition via Large-Scale Weak Supervision'. In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. ICML'23. Honolulu, Hawaii, USA: JMLR.org, 2023, pp. 28492–28518 (cited on page 61).

[334] Seamless Communication et al. *SeamlessM4T: Massively Multilingual & Multimodal Machine Translation*. 2023 (cited on pages 61, 114, 120).

[335] Federica Biassoni, Stefania Balzarotti, Micaela Giamporcaro, and Rita Ciceri. 'Hot or Cold Anger? Verbal and Vocal Expression of Anger While Driving in a Simulated Anger-Provoking Scenario'. In: *SAGE Open* 6.3 (2016), p. 215824401665808. DOI: 10/ggcgqh (cited on page 63).

[336] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 'Imagenet: A Large-Scale Hierarchical Image Database'. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255 (cited on pages 63, 69, 135).

[337] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 'Microsoft COCO: Common Objects in Context'. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755 (cited on pages 63, 64, 69).

[338] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. *The Kinetics Human Action Video Dataset*. 2017 (cited on pages 63, 135).

[339] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 'Audio Set: An Ontology and Human-Labeled Dataset for Audio Events'. In: *Proc. IEEE ICASSP 2017*. New Orleans, LA, 2017 (cited on pages 63, 135).

[340] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 'Places: A 10 Million Image Database for Scene Recognition'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017) (cited on pages 63, 135).

[341] James A. Green, Pamela G. Whitney, and Gwen E. Gustafson. 'Vocal Expressions of Anger'. In: *International Handbook of Anger: Constituent and Concomitant Biological, Psychological, and Social Processes*. Ed. by Michael Potegal, Gerhard Stemmler, and Charles Spielberger. New York, NY: Springer, 2010, pp. 139–156. DOI: 10.1007/978-0-387-89676-2_9 (cited on page 63).

[342] Luis Von Ahn and Laura Dabbish. 'Designing Games with a Purpose'. In: *Communications of the ACM* 51.8 (2008), pp. 58–67 (cited on pages 63, 64).

[343] Luis Von Ahn and Laura Dabbish. 'Labeling Images with a Computer Game'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2004, pp. 319–326 (cited on page 64).

[344] Edith LM Law, Luis Von Ahn, Roger B Dannenberg, and Mike Crawford. 'TagATune: A Game for Music and Sound Annotation'. In: *ISMIR*. Vol. 3. 2007, p. 2 (cited on page 64).

[345] Luis von Ahn, Ruoran Liu, and Manuel Blum. 'Peekaboom: A Game for Locating Objects in Images'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '06. New York, NY, USA: Association for Computing Machinery, 2006, pp. 55–64. DOI: 10.1145/1124772.1124782 (cited on page 64).

[346] Luis von Ahn, Shiry Ginosar, Mihir Kedia, and Manuel Blum. 'Improving Image Search with PHETCH'. In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07. Vol. 4. 2007, pp. IV-1209-IV–1212. DOI: 10.1109/ICASSP.2007.367293 (cited on page 64).

[347] Luis von Ahn, Manuel Blum, and John Langford. 'Telling Humans and Computers Apart Automatically'. In: *Communications of the ACM* 47.2 (2004), pp. 56–60. DOI: 10.1145/966389.966390 (cited on pages 64, 99).

[348] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 'Cascade: Crowdsourcing Taxonomy Creation'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13. New York, NY, USA: Association for Computing Machinery, 2013, pp. 1999–2008. DOI: 10.1145/2470654.2466265 (cited on page 64).

[349] Jonathan Bragg, - Mausam, and Daniel Weld. 'Crowdsourcing Multi-Label Classification for Taxonomy Creation'. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 1 (2013), pp. 25–33. DOI: 10.1609/hcomp.v1i1.13091 (cited on page 64).

[350] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 'YFCC100M: The New Data in Multimedia Research'. In: *Commun. ACM* 59.2 (2016), pp. 64–73. DOI: 10.1145/2812802 (cited on page 64).

[351] Ranjay A. Krishna, Kenji Hata, Stephanie Chen, Joshua Kravitz, David A. Shamma, Li Fei-Fei, and Michael S. Bernstein. 'Embracing Error to Enable Rapid Crowdsourcing'. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 3167–3179. DOI: 10.1145/2858036.2858115 (cited on page 65).

[352] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 'Cheap and Fast—but Is It Good? Evaluating Non-Expert Annotations for Natural Language Tasks'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '08. USA: Association for Computational Linguistics, 2008, pp. 254–263 (cited on page 65).

[353] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 'Fast Unfolding of Communities in Large Networks'. In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008 (cited on page 67).

[354] Robyn Speer, Joshua Chin, and Catherine Havasi. 'Conceptnet 5.5: An Open Multilingual Graph of General Knowledge'. In: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017 (cited on page 68).

[355] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. In: *Proceedings of the 2019 Conference of the North*. Proceedings of the 2019 Conference of the North. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423 (cited on page 68).

[356] Lin CY Rouge. 'A Package for Automatic Evaluation of Summaries'. In: *Proceedings of Workshop on Text Summarization of ACL, Spain*. 2004 (cited on page 68).

[357] Joshua C Peterson, Joshua T Abbott, and Thomas L Griffiths. 'Evaluating (and Improving) the Correspondence between Deep Neural Networks and Human Representations'. In: *Cognitive Science* 42.8 (2018), pp. 2648–2669 (cited on page 68).

[358] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 'Rethinking Spatiotemporal Feature Learning: Speed-accuracy Trade-Offs in Video Classification'. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 305–321 (cited on pages 68, 69, 71).

[359] Duncan J. Watts. *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton University Press, 2004. 282 pp. (cited on page 69).

[360] Duncan J. Watts and Steven H. Strogatz. 'Collective Dynamics of 'Small-World' Networks'. In: *Nature* 393.6684 (1998), pp. 440–442. DOI: 10.1038/30918 (cited on page 69).

[361] Albert-László Barabási and Réka Albert. 'Emergence of Scaling in Random Networks'. In: *Science* 286.5439 (1999), pp. 509–512. DOI: 10.1126/science.286.5439.509 (cited on page 69).

[362] M. E. J. Newman. 'The Structure of Scientific Collaboration Networks'. In: *Proceedings of the National Academy of Sciences* 98.2 (2001), pp. 404–409. DOI: 10.1073/pnas.98.2.404 (cited on page 69).

[363] Mark Steyvers and Joshua B. Tenenbaum. 'The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth'. In: *Cognitive Science* 29.1 (2005), pp. 41–78. DOI: 10.1207/s15516709cog2901_3 (cited on page 69).

[364] George A. Miller. 'WordNet: A Lexical Database for English'. In: *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 8-11, 1994*. HLT 1994. 1994 (cited on page 69).

[365] DL Nelson. 'The University of South Florida Word Association Norms'. In: *http://w3.usf.edu/FreeAssociation* (1999) (cited on page 69).

[366] Peter Mark Roget. *Roget's Thesaurus of English Words and Phrases.* TY Crowell Company, 1911 (cited on page 69).

[367] Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. 'BOLD5000, a Public fMRI Dataset While Viewing 5000 Visual Images'. In: *Scientific data* 6.1 (2019), p. 49 (cited on pages 69, 70).

[368] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 'Sun Database: Large-scale Scene Recognition from Abbey to Zoo'. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3485–3492 (cited on page 69).

[369] Saif M. Mohammad and Svetlana Kiritchenko. 'An Annotated Dataset of Emotions Evoked by Art'. In: *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*. Miyazaki, Japan, 2018 (cited on pages 69, 72).

[370] Hillary Anger Elfenbein, Petri Laukka, Jean Althoff, Wanda Chui, Frederick K. Iraki, Thomas Rockstuhl, and Nutankumar S. Thingujam. 'What Do We Hear in the Voice? An Open-Ended Judgment Study of Emotional Speech Prosody'. In: *Personality and Social Psychology Bulletin* (2021), p. 014616722110297. DOI: 10/gmxp9c (cited on page 77).

[371] Matias Fernandez-Duque, Sayuri Hayakawa, and Viorica Marian. 'Speakers of Different Languages Remember Visual Scenes Differently'. In: *Science Advances* 9.33 (2023). DOI: 10.1126/sciadv.adh0064 (cited on page 77).

[372] Claire Kramsch. 'Language and Culture'. In: *AILA review* 27.1 (2014), pp. 30–55 (cited on pages 77, 136).

[373] Thomas N. Wisdom, Xianfeng Song, and Robert L. Goldstone. 'Social Learning Strategies in Networked Groups'. In: *Cognitive Science* 37.8 (2013), pp. 1383–1425. DOI: 10.1111/cogs.12052 (cited on page 78).

[374] Hirokazu Shirado and Nicholas A. Christakis. 'Locally Noisy Autonomous Agents Improve Global Human Coordination in Network Experiments'. In: *Nature* 545.7654 (2017), pp. 370–374. DOI: 10.1038/nature22332 (cited on page 78).

[375] Bertrand Jayles, Hye-rin Kim, Ramón Escobedo, Stéphane Cezera, Adrien Blanchet, Tatsuya Kameda, Clément Sire, and Guy Theraulaz. 'How Social Information Can Improve Estimation Accuracy in Human Groups'. In: *Proceedings of the National Academy of Sciences* 114.47 (2017), pp. 12620–12625. DOI: 10.1073/pnas.1703695114 (cited on page 78).

[376] Damon Centola and Andrea Baronchelli. 'The Spontaneous Emergence of Conventions: An Experimental Study of Cultural Evolution'. In: *Proceedings of the National Academy of Sciences* 112.7 (2015), pp. 1989–1994. DOI: 10.1073/pnas.1418838112 (cited on page 78).

[377] Jon W. Carr, Kenny Smith, Jennifer Culbertson, and Simon Kirby. 'Simplicity and Informativeness in Semantic Category Systems'. In: *Cognition* 202 (2020), p. 104289. DOI: 10.1016/j.cognition.2020.104289 (cited on page 78).

[378] Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. 'Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market'. In: *Science* 311.5762 (2006), pp. 854–856. DOI: 10.1126/science.1121066 (cited on page 78).

[379] Stefano Balietti, Robert L. Goldstone, and Dirk Helbing. 'Peer Review and Competition in the Art Exhibition Game'. In: *Proceedings of the National Academy of Sciences* 113.30 (2016), pp. 8414–8419. DOI: 10.1073/pnas.1603723113 (cited on page 78).

[380] Imre Lahdelma and Tuomas Eerola. 'Cultural Familiarity and Musical Expertise Impact the Pleasantness of Consonance/Dissonance but Not Its Perceived Tension'. In: *Scientific Reports* 10.1 (2020), p. 8693. DOI: 10.1038/s41598-020-65615-8 (cited on page 78).

[381] Jonas Stein, Marc Keuschnigg, and Arnout van de Rijt. 'Network Segregation and the Propagation of Misinformation'. In: *Scientific Reports* 13.1 (2023), p. 917. DOI: 10.1038/s41598-022-26913-5 (cited on page 78).

[382] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 'The Science of Fake News'. In: *Science* 359.6380 (2018), pp. 1094–1096. DOI: 10.1126/science.aao2998 (cited on page 78).

[383] Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. 'PsychoPy2: Experiments in Behavior Made Easy'. In: *Behavior Research Methods* 51.1 (2019), pp. 195–203. DOI: 10.3758/s13428-018-01193-y (cited on page 78).

[384] Joshua R. de Leeuw, Rebecca A. Gilbert, and Björn Luchterhandt. 'jsPsych: Enabling an Open-Source Collaborative Ecosystem of Behavioral Experiments'. In: *Journal of Open Source Software* 8.85 (2023), p. 5351. DOI: 10.21105/joss.05351 (cited on page 78).

[385] Kristian Lange, Simone Kühn, and Elisa Filevich. '"Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies'. In: *PLOS ONE* 10.6 (2015), e0130834. DOI: 10.1371/journal.pone.0130834 (cited on page 78).

[386] Alexander L. Anwyl-Irvine, Jessica Massonnié, Adam Flitton, Natasha Kirkham, and Jo K. Evershed. 'Gorilla in Our Midst: An Online Behavioral Experiment Builder'. In: *Behavior Research Methods* 52.1 (2020), pp. 388–407. DOI: 10.3758/s13428-019-01237-x (cited on page 78).

[387] Sebastiaan Mathot, Daniel Schreij, and Jan Theeuwes. 'OpenSesame: An Open-Source, Graphical Experiment Builder for the Social Sciences'. In: *Behavior Research Methods* 44.2 (2012), pp. 314–324. DOI: 10.3758/s13428-011-0168-7 (cited on page 78).

[388] Sebastiaan Mathot and Jennifer March. 'Conducting Linguistic Experiments Online With OpenSesame and OSWeb'. In: *Language Learning* 72.4 (2022), pp. 1017–1048. DOI: 10.1111/lang.12509 (cited on page 78).

[389] Joshua K. Hartshorne, Joshua R. de Leeuw, Noah D. Goodman, Mariela Jennings, and Timothy J. O'Donnell. 'A Thousand Studies for the Price of One: Accelerating Psychological Science with Pushkin'. In: *Behavior Research Methods* 51.4 (2019), pp. 1782–1803. DOI: 10.3758/s13428-018-1155-z (cited on page 78).

[390] Joshua R. de Leeuw. 'jsPsych: A JavaScript Library for Creating Behavioral Experiments in a Web Browser'. In: *Behavior Research Methods* 47.1 (2015), pp. 1–12. DOI: 10.3758/s13428-014-0458-y (cited on page 78).

[391] Jan Simson and Samuel A. Mehr. *World-Wide-Lab*. https://github.com/world-wide-lab/world-wide-lab. 2024 (cited on page 78).

[392] Abdullah Almaatouq, Joshua Becker, James P. Houghton, Nicolas Paton, Duncan J. Watts, and Mark E. Whiting. 'Empirica: A Virtual Lab for High-Throughput Macro-Level Experiments'. In: *Behavior Research Methods* 53.5 (2021), pp. 2158–2171. DOI: 10.3758/s13428-020-01535-9 (cited on page 78).

[393] *Dallinger*. GitHub, 2022. URL: https://github.com/Dallinger/Dallinger (cited on page 78).

[394] Gerard Saucier et al. 'Cross-Cultural Differences in a Global "Survey of World Views"'. In: *Journal of Cross-Cultural Psychology* 46.1 (2014), pp. 53–70. DOI: 10.1177/0022022114551791 (cited on page 85).

[395] Michele J. Gelfand et al. 'Differences between Tight and Loose Cultures: A 33-Nation Study'. In: *Science* 332.6033 (2011), pp. 1100–1104. DOI: 10.1126/science.1197754 (cited on page 85).

[396] Jingxuan Liu, Courtney B. Hilton, Elika Bergelson, and Samuel A. Mehr. 'Language Experience Predicts Music Processing in a Half-Million Speakers of Fifty-Four Languages'. In: *Current Biology* 33.10 (2023), 1916–1925.e4. DOI: 10.1016/j.cub.2023.03.067 (cited on page 85).

[397] Robert Thomson et al. 'Relational Mobility Predicts Social Behaviors in 39 Countries and Is Tied to Historical Farming and Threat'. In: *Proceedings of the National Academy of Sciences* 115.29 (2018), pp. 7521–7526. DOI: 10.1073/pnas.1713191115 (cited on page 85).

[398] Angela de Bruin. 'Not All Bilinguals Are the Same: A Call for More Detailed Assessments and Descriptions of Bilingual Experiences'. In: *Behavioral Sciences* 9.3 (2019), p. 33. DOI: 10.3390/bs9030033 (cited on page 85).

[399] Viorica Marian and Sayuri Hayakawa. 'Measuring Bilingualism: The Quest for a "Bilingualism Quotient"'. In: *Applied Psycholinguistics* 42.2 (2020), pp. 527–548. DOI: 10.1017/s0142716420000533 (cited on page 85).

[400] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 'Data Quality of Platforms and Panels for Online Behavioral Research'. In: *Behavior Research Methods* 54.4 (2021), pp. 1643–1662. DOI: 10.3758/s13428-021-01694-3 (cited on page 85).

[401] Benjamin D. Douglas, Patrick J. Ewell, and Markus Brauer. 'Data Quality in Online Human-Subjects Research: Comparisons between Mturk, Prolific, CloudResearch, Qualtrics, and SONA'. In: *PLOS ONE* 18.3 (2023). Ed. by Jeffrey S. Hallam, e0279720. DOI: 10.1371/journal.pone.0279720 (cited on page 85).

[402] Siddharth Suri, Daniel G Goldstein, and Winter A Mason. 'Honesty in an Online Labor Market'. In: Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence. 2011 (cited on page 85).

[403] J Charles Alderson and Ari Huhta. 'The Development of a Suite of Computer-Based Diagnostic Tests Based on the Common European Framework'. In: *Language Testing* 22.3 (2005), pp. 301–320 (cited on page 86).

[404] Lloyd M. Dunn and Douglas M. Dunn. 'Peabody Picture Vocabulary Test–Fourth Edition'. In: (2007). DOI: 10.1037/t15144-000 (cited on page 86).

[405] University. Cambridge. Local Examinations Syndicate and Oxford University Press. *Quick Placement Test*. Oxford University Press, 2001 (cited on page 86).

[406] Alaa Alzahrani. 'LexArabic: A Receptive Vocabulary Size Test to Estimate Arabic Proficiency'. In: *Behavior Research Methods* (2023). DOI: 10.3758/s13428-023-02286-z (cited on page 86).

[407] I Lei Chan and Charles Chang. 'LEXTALE-CH: A Quick, Character-Based Proficiency Test for Mandarin Chinese'. In: *Proceedings of the Annual Boston University Conference on Language Development (BUCLD)* 42 (2018), pp. 114–130 (cited on pages 86, 93).

[408] Yun Wen, Yicheng Qiu, Christine Xiang Ru Leong, and Walter J. B. van Heuven. 'LexCHI: A Quick Lexical Test for Estimating Language Proficiency in Chinese'. In: *Behavior Research Methods* (2023). DOI: 10.3758/s13428-023-02151-z (cited on page 86).

[409] Kaidi Lõo, Katrin Leppik, Agu Bleve, and Anton Malmi. 'Estonian L2 Vocabulary Test LexEst'. In: *20th Annual Conference of Applied Linguistics 2023*. Tallinn, Estonia: Eesti Keele Instituut, 2023 (cited on page 86).

[410] Rosa Salmela, Minna Lehtonen, Sefano Garusi, and Raymond Bertram. 'Lexize: A Test to Quickly Assess Vocabulary Knowledge in Finnish'. In: *Scandinavian Journal of Psychology* 62.6 (2021), pp. 806–819. DOI: 10.1111/sjop.12768 (cited on pages 86, 93).

[411] Marc Brysbaert. 'Lextale_FR a Fast, Free, and Efficient Test to Measure Language Proficiency in French'. In: *Psychologica Belgica* 53.1 (2013), p. 23. DOI: 10.5334/pb-53-1-23 (cited on pages 86, 93).

[412] Simona Amenta, Linda Badan, and Marc Brysbaert. 'LexITA: A Quick and Reliable Assessment Tool for Italian L2 Receptive Vocabulary Size'. In: *Applied Linguistics* 42.2 (2020), pp. 292–314. DOI: 10.1093/applin/amaa020 (cited on pages 86, 93).

[413] Soon Tat Lee, Walter J. B. van Heuven, Jessica M. Price, and Christine Xiang Ru Leong. 'LexMAL: A Quick and Reliable Lexical Test for Malay Speakers'. In: *Behavior Research Methods* (2023). DOI: 10.3758/s13428-023-02202-5 (cited on page 86).

[414] Marek Muszyński, Natalia Banasik-Jemielniak, Tomasz Żółtak, Kaili Rimfeld, Nicholas G Shakeshaft, Kerry L Schofield, Margherita Malanchini, and Artur Pokropek. *Moving Intelligence Measurement Online: Adaptation and Validation of the Polish Version of the Pathfinder General Cognitive Ability Test*. PsyArXiv, 2023. DOI: 10.31234/osf.io/tqyux (cited on page 86).

[415] Chao Zhou and Xinyi Li. 'LextPT: A Reliable and Efficient Vocabulary Size Test for L2 Portuguese Proficiency'. In: *Behavior Research Methods* 54.6 (2021), pp. 2625–2639. DOI: 10.3758/s13428-021-01731-1 (cited on page 86).

[416] Cristina Izura, Fernando Cuetos, and Marc Brysbaert. *Lexical Test for Advanced Learners of Spanish*. American Psychological Association (APA), 2014. DOI: 10.1037/t47086-000 (cited on pages 86, 93).

[417] Maria Ferin, Henrik Gyllstad, Ilaria Venagli, Angelica Zordan, and Tanja Kupisch. 'LexVEN: A Quick Vocabulary Test for Proficiency in Venetan'. In: *Isogloss. Open Journal of Romance Linguistics* 9.1 (2023), pp. 1–31. DOI: 10.5565/rev/isogloss.374 (cited on page 86).

[418] *Prolific*. Prolific Academic Ltd, 2024. URL: https://www.prolific.com/ (cited on pages 86, 92).

[419] *Lucid Marketplace*. Cint, 2024. URL: https://luc.id/marketplace-book (cited on page 86).

[420] Morgane Laouenan, Palaash Bhargava, Jean-Benoit Eyméoud, Olivier Gergaud, Guillaume Plique, and Etienne Wasmer. 'A Cross-Verified Database of Notable People, 3500BC-2018AD'. In: *Scientific Data* 9.1 (2022). DOI: 10.1038/s41597-022-01369-4 (cited on page 87).

[421] Nathan Schneider, Primavera De Filippi, Seth Frey, Joshua Z. Tan, and Amy X. Zhang. 'Modular Politics'. In: *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW1 2021), pp. 1–26. DOI: 10.1145/3449090 (cited on page 87).

[422] Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 'Wiki-40B: Multilingual Language Model Dataset'. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. LREC 2020. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: European Language Resources Association, 2020, pp. 2440–2452 (cited on page 87).

[423] Wikimedia Foundation. *Wikimedia Downloads*. URL: https://dumps.wikimedia.org (cited on page 87).

[424] Harald Hammarström and Robert Forkel. 'Glottocodes: Identifiers Linking Families, Languages and Dialects to Comprehensive Reference Information'. In: *Semantic Web Journal* 13.6 (2022), pp. 917–924 (cited on pages 87, 95).

[425] Ling Liu, Zach Ryan, and Mans Hulden. 'The Usefulness of Bibles in Low-Resource Machine Translation'. In: *Proceedings of the Workshop on Computational Methods for Endangered Languages* 1.2 (2021). DOI: 10.33011/computel.v1i.957 (cited on page 87).

[426] Željko Agić, Dirk Hovy, and Anders Søgaard. 'If All You Have Is a Bit of the Bible: Learning POS Taggers for Truly Low-Resource Languages'. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. ACL-IJCNLP 2015. Ed. by Chengqing Zong and Michael Strube. Beijing, China: Association for Computational Linguistics, 2015, pp. 268–272. DOI: 10.3115/v1/P15-2044 (cited on page 87).

[427] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 'Scaling Speech Technology to 1,000+ Languages'. In: *Journal of Machine Learning Research* 25.97 (2024), pp. 1–52 (cited on pages 87, 99).

[428] Don Tuggener. ‘Incremental Coreference Resolution for German’. PhD thesis. University of Zurich, 2016 (cited on page 88).

[429] Ulf Hermjakob. *Isi-Nlp/Uroman*. ISI NLP, 2024 (cited on page 89).

[430] Huang Huang. *Mozillazg/Python-Pinyin*. 2024 (cited on page 89).

[431] Hiroshi Miura. *Pykakasi: Kana Kanji Simple Inversion Library*. Version 2.3.0 (cited on page 89).

[432] Shin Haebin. *Jamotools: A Library for Korean Jamo Split and Vectorize.* Version 0.1.10 (cited on page 89).

[433] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. 1st. O’Reilly Media, Inc., 2009 (cited on page 89).

[434] Matthew Honnibal and Ines Montani. ‘spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing’. 2017 (cited on pages 89, 90).

[435] Milan Straka. ‘UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task’. In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 197–207. DOI: 10.18653/v1/K18-2020 (cited on page 89).

[436] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. ‘Enriching Word Vectors with Subword Information’. In: *Transactions of the Association for Computational Linguistics* 5 (Dec. 2017), pp. 135–146. DOI: 10.1162/tacl_a_00051 (cited on page 89).

[437] *LibreOffice/Dictionaries*. LibreOffice, 2024 (cited on page 89).

[438] *Guess_language-Spirit: Guess the Natural Language of a Text*. Version 0.5.3 (cited on page 89).

[439] *Pyenchant/Pyenchant*. pyenchant, 2024 (cited on page 89).

[440] Chris Dyer, Victor Chahuneau, and Noah A. Smith. ‘A Simple, Fast, and Effective Reparameterization of IBM Model 2’. In: North American Chapter of the Association for Computational Linguistics. 2013 (cited on page 90).

[441] *Clab/Fast_align*. Chris Dyer’s lab @ LTI/CMU, 2024 (cited on page 90).

[442] Robert Rouse. *Robertrouse/Theographic-Bible-Metadata*. 2024 (cited on page 90).

[443] Isabelle Dautriche, Kyle Mahowald, Edward Gibson, Anne Christophe, and Steven T Piantadosi. ‘Words Cluster Phonetically beyond Phonotactic Regularities’. In: *Cognition* 163 (2017), pp. 128–145 (cited on page 90).

[444] Sean Trott and Benjamin Bergen. ‘Why Do Human Languages Have Homophones?’ In: *Cognition* 205 (2020), p. 104449. DOI: 10.1016/j.cognition.2020.104449 (cited on page 90).

[445] Adam Cohen. *Thefuzz: Fuzzy String Matching in Python*. Version 0.22.1 (cited on page 91).

[446] Martin Haspelmath, Matthew S Dryer, David Gil, and Bernard Comrie. *The World Atlas of Language Structures*. OUP Oxford, 2005 (cited on page 95).

[447] Elazar Leshem and Moshe Sicron. ‘The Absorption of Soviet Immigrants in Israel’. In: *The American Jewish Year Book* 99 (1999), pp. 484–522 (cited on page 95).

[448] Noah Lewin-Epstein and Yinon Cohen. ‘Ethnic Origin and Identity in the Jewish Population of Israel’. In: *Journal of Ethnic and Migration Studies* 45.11 (2018), pp. 2118–2137. DOI: 10.1080/1369183x.2018.1492370 (cited on page 98).

[449] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. *Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks*. 2023 (cited on page 99).

[450] OpenAI et al. *GPT-4 Technical Report*. 2024 (cited on pages 99, 105).

[451] Benjamin Lee Whorf. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf.* Cambridge, MA: MIT Press, 1956 (cited on pages 101, 136).

[452] Brent Berlin and Paul Kay. *Basic Color Terms: Their Universality and Evolution.* Berkeley, CA: University of California Press., 1969 (cited on page 102).

[453] Eric H. Lenneberg and John M. Roberts. *The Language of Experience: A Study in Methodology*. Indiana University Publications in Anthropology and Linguistics Memoir 13. Baltimore: Waverly Press, 1956. 33 pp. (cited on page 102).

[454] Delwin T. Lindsey and Angela M. Brown. 'Universality of Color Names'. In: *Proceedings of the National Academy of Sciences* 103.44 (2006), pp. 16608–16613. DOI: 10.1073/pnas.0607708103 (cited on pages 103, 136).

[455] Terry Regier, Paul Kay, and Naveen Khetarpal. 'Color Naming Reflects Optimal Partitions of Color Space'. In: *Proceedings of the National Academy of Sciences* 104.4 (2007), pp. 1436–1441. DOI: 10.1073/pnas.0610341104 (cited on pages 103, 136).

[456] Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 'Efficient Compression in Color Naming and Its Evolution'. In: *Proceedings of the National Academy of Sciences* 115.31 (2018), pp. 7937–7942. DOI: 10.1073/pnas.1800521115 (cited on pages 103, 136).

[457] Benjamin Pitt, Stephen Ferrigno, Jessica F. Cantlon, Daniel Casasanto, Edward Gibson, and Steven T. Piantadosi. 'Spatial Concepts of Number, Size, and Time in an Indigenous Culture'. In: *Science Advances* 7.33 (2021). DOI: 10.1126/sciadv.abg4141 (cited on page 104).

[458] Stephen C. Levinson and Asifa Majid. 'The Island of Time: Yéli Dnye, the Language of Rossel Island'. In: *Frontiers in Psychology* 4 (2013). DOI: 10.3389/fpsyg.2013.00061 (cited on page 104).

[459] Sarah Dolscheid, Shakila Shayan, Asifa Majid, and Daniel Casasanto. 'The Thickness of Musical Pitch: Psychophysical Evidence for Linguistic Relativity'. In: *Psychological science* 24.5 (2013), pp. 613–621 (cited on page 104).

[460] Ronald Inglehart, Miguel Basanez, Jaime Diez-Medrano, Loek Halman, and Ruud Luijkx. 'World Values Surveys and European Values Surveys, 1981-1984, 1990-1993, and 1995-1997'. In: *Ann Arbor-Michigan, Institute for Social Research, ICPSR version* (2000) (cited on page 104).

[461] Harry C Triandis. *Individualism and Collectivism*. Routledge, 2018 (cited on page 104).

[462] Ben Fine. *The World of Consumption: The Material and Cultural Revisited*. Routledge, 2016 (cited on page 104).

[463] Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. 'In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies'. In: *American Economic Review* 91.2 (2001), pp. 73–78 (cited on page 104).

[464] H Clark Barrett. 'Towards a Cognitive Science of the Human: Cross-Cultural Approaches and Their Urgency'. In: *Trends in cognitive sciences* 24.8 (2020), pp. 620–638 (cited on pages 104, 108, 131).

[465] Jan Nederveen Pieterse. *Globalization and Culture: Global Mélange*. Rowman & Littlefield, 2019 (cited on page 104).

[466] Allison Chen, Ilia Sucholutsky, Olga Russakovsky, and Tom Griffiths. 'Analyzing the Roles of Language and Vision in Learning from Limited Data'. In: *CogSci*. Proceedings of the Annual Meeting of the Cognitive Science Society. 2024 (cited on page 104).

[467] Raja Marjieh, Ilia Sucholutsky, Ted Sumers, Nori Jacoby, and Tom Griffiths. 'Predicting Human Similarity Judgments Using Large Language Models'. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 44. 44. 2022 (cited on page 104).

[468] Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T. Stevens, and Morteza Dehghani. 'Morality beyond the WEIRD: How the Nomological Network of Morality Varies across Cultures'. In: *Journal of Personality and Social Psychology* 125.5 (2023), pp. 1157–1188. DOI: 10.1037/pspp0000470 (cited on page 104).

[469] JH Clark. 'The Ishihara Test for Color Blindness.' In: *American Journal of Physiological Optics* (1924) (cited on page 104).

[470] Richard S Cook, Paul Kay, and Terry Regier. 'The World Color Survey Database'. In: *Handbook of Categorization in Cognitive Science*. Elsevier, 2005, pp. 223–241 (cited on page 105).

[471] *Professional Language Solutions for Your Business - Translated*. URL: https://translated.com/welcome (cited on page 105).

[472] Adrien Barbaresi. *Simplemma*. GitHub, 2024. URL: https://github.com/adbar/simplemma (cited on page 105).

[473] Robin Speer. *Wordfreq*. GitHub, 2024. URL: https://github.com/rspeer/wordfreq (cited on page 106).

[474] Seatgeak. *TheFuzz*. GitHub, 2024. URL: https://github.com/seatgeek/thefuzz (cited on page 106).

[475] Lawrence Hubert and Phipps Arabie. 'Comparing Partitions'. In: *Journal of Classification* 2.1 (1985), pp. 193–218. DOI: 10.1007/bf01908075 (cited on page 106).

[476] Rachael E. Jack and Philippe G. Schyns. 'The Human Face as a Dynamic Tool for Social Communication'. In: *Current Biology* 25.14 (2015), R621–R634. DOI: 10.1016/j.cub.2015.05.052 (cited on page 113).

[477] Carmel Sofer, Ron Dotsch, Daniel H. J. Wigboldus, and Alexander Todorov. 'What Is Typical Is Good'. In: *Psychological Science* 26.1 (2014), pp. 39–47. DOI: 10.1177/0956797614554955 (cited on page 113).

[478] Raphael Angulu, Jules R. Tapamo, and Aderemi O. Adewumi. 'Age Estimation via Face Images: A Survey'. In: *EURASIP Journal on Image and Video Processing* 2018.1 (2018). DOI: 10.1186/s13640-018-0278-6 (cited on page 113).

[479] Mirella Walker and Thomas Vetter. 'Changing the Personality of a Face: Perceived Big Two and Big Five Personality Factors Modeled in Real Photographs.' In: *Journal of Personality and Social Psychology* 110.4 (2016), pp. 609–624. DOI: 10.1037/pspp0000064 (cited on page 113).

[480] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. 'Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis'. In: *Advances in neural information processing systems* 31 (2018) (cited on pages 113, 114, 120).

[481] Daisy Stanton, Matt Shannon, Soroosh Mariooryad, RJ Skerry-Ryan, Eric Battenberg, Tom Bagby, and David Kao. 'Speaker Generation'. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 7897–7901. DOI: 10.1109/ICASSP43922.2022.9747345 (cited on pages 113, 120).

[482] M. Bernard. *Phonemizer*. GitHub, 2021. URL: https://github.com/bootphon/phonemizer (cited on page 114).

[483] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 'HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis'. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020, pp. 17022–17033 (cited on page 114).

[484] Nithin Rao Koluguri, Jason Li, Vitaly Lavrukhin, and Boris Ginsburg. 'SpeakerNet: 1D Depth-Wise Separable Convolutional Network for Text-Independent Speaker Recognition and Verification'. 2020 (cited on page 114).

[485] Justin N. M. Pinkney and Doron Adler. 'Resolution Dependent GAN Interpolation for Controllable Image Synthesis between Domains'. In: *CoRR* abs/2010.05334 (2020) (cited on page 115).

[486] Sato. *Ai Gahaku*. 2022. URL: https://ai-art.tokyo (cited on page 115).

[487] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 'The Unreasonable Effectiveness of Deep Features as a Perceptual Metric'. In: *CVPR*. 2018 (cited on page 115).

[488] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 'Common Voice: A Massively-Multilingual Speech Corpus'. In: *LREC*. 2020, pp. 4218–4222 (cited on page 115).

[489] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 'A Lip Sync Expert Is All You Need for Speech to Lip Generation in the Wild'. In: *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 2020. DOI: 10.1145/3394171.3413532 (cited on page 116).

[490] J. L. Fitch and A. Holbrook. 'Modal Vocal Fundamental Frequency of Young Adults'. In: *Archives of Otolaryngology - Head and Neck Surgery* 92.4 (1970), pp. 379–382. DOI: 10.1001/archotol.1970.04310040067012 (cited on page 117).

[491] Bilge Mutlu, Steven Osman, Jodi Forlizzi, Jessica K. Hodgins, and Sara B. Kiesler. 'Task Structure and User Attributes as Elements of Human-Robot Interaction Design'. In: *The 15th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2006, Hatfield, Herthfordshire, UK, 6-8 September, 2006*. IEEE, 2006, pp. 74–79. DOI: 10.1109/ROMAN.2006.314397 (cited on page 119).

[492] Klaus R. Scherer. 'Personality Inference from Voice Quality: The Loud Voice of Extroversion'. In: *European Journal of Social Psychology* 8.4 (1978), pp. 467–487. DOI: 10.1002/ejsp.2420080405 (cited on page 119).

[493] Klaus J. Kohler. 'Communicative Functions Integrate Segments in Prosodies and Prosodies in Segments'. In: *Phonetica* 68.1-2 (2011), pp. 26–56. DOI: 10/d9vhxq (cited on page 119).

[494] Cynthia Breazeal. *Designing Sociable Robots*. Intelligent Robotics and Autonomous Agents Series. MIT Press, 2004 (cited on page 119).

[495] Kathrin Janowski, Hannes Ritschel, and Elisabeth André. 'Adaptive Artificial Personalities'. In: *The Handbook on Socially Interactive Agents*. ACM, 2022, pp. 155–194. DOI: 10.1145/3563659.3563666 (cited on pages 119, 130).

[496] Anna Esposito, Terry Amorese, Marialucia Cuciniello, Maria Teresa Riviello, Antonietta Maria Esposito, Alda Troncone, and Gennaro Cordasco. 'The Dependability of Voice on Elders' Acceptance of Humanoid Agents.' In: *Interspeech*. 2019, pp. 31–35 (cited on page 119).

[497] Silvan Mertes, Thomas Kiderle, Ruben Schlagowski, Florian Lingenfelser, and Elisabeth André. 'On the Potential of Modular Voice Conversion for Virtual Agents'. In: *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2021, pp. 1–7 (cited on page 119).

[498] Sigrid Roehling, Bruce MacDonald, and Catherine Watson. 'Towards Expressive Speech Synthesis in English on a Robotic Platform'. In: *Proceedings of the Australasian International Conference on Speech Science and Technology*. Citeseer, 2006, pp. 130–135 (cited on page 120).

[499] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. *A Survey on Neural Speech Synthesis*. arXiv, 2021. DOI: 10.48550/ARXIV.2106.15561 (cited on page 120).

[500] Katharina Kühne, Martin H Fischer, and Yuefang Zhou. 'The Human Takes It All: Humanlike Synthesized Voices Are Perceived as Less Eerie and More Likable. Evidence from a Subjective Ratings Study'. In: *Frontiers in neurorobotics* (2020), p. 105 (cited on page 120).

[501] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 'Glow-Tts: A Generative Flow for Text-to-Speech via Monotonic Alignment Search'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 8067–8077 (cited on page 120).

[502] Adrian Łańcucki. 'Fastpitch: Parallel Text-to-Speech with Pitch Prediction'. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6588–6592 (cited on page 120).

[503] Fernando Alonso-Martín, María Malfaz, Álvaro Castro-González, José Carlos Castillo, and Miguel A Salichs. 'Online Evaluation of Text to Speech Systems for Three Social Robots'. In: *International Conference on Social Robotics*. Springer, 2019, pp. 155–164 (cited on pages 120, 121).

[504] Fernando Alonso Martin, María Malfaz, Álvaro Castro-González, José Carlos Castillo, and Miguel Ángel Salichs. 'Four-Features Evaluation of Text to Speech Systems for Three Social Robots'. In: *Electronics* 9.2 (2020), p. 267 (cited on pages 120, 121).

[505] Conor McGinn and Ilaria Torre. 'Can You Tell the Robot by the Voice? An Exploratory Study on the Role of Voice in the Perception of Robots'. In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 211–221 (cited on pages 120, 121).

[506] Aaron Powers and Sara Kiesler. 'The Advisor Robot: Tracing People's Mental Model from a Robot's Physical Attributes'. In: *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*. 2006, pp. 218–225 (cited on pages 120, 121, 126, 127).

[507] Ilaria Torre, Jeremy Goslin, and Laurence White. 'If Your Device Could Smile: People Trust Happy-Sounding Artificial Agents More'. In: *Computers in Human Behavior* 105 (2020), p. 106215 (cited on page 120).

[508] Ilaria Torre, Jeremy Goslin, Laurence White, and Debora Zanatto. 'Trust in Artificial Voices: A" Congruency Effect" of First Impressions and Behavioural Experience'. In: *Proceedings of the Technology, Mind, and Society*. 2018, pp. 1–6 (cited on pages 120, 121).

[509] Jennifer Goetz, Sara Kiesler, and Aaron Powers. 'Matching Robot Appearance and Behavior to Tasks to Improve Human-Robot Cooperation'. In: *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003*. Ieee, 2003, pp. 55–60 (cited on page 120).

[510] Jesin James, BT Balamurali, Catherine I Watson, and Bruce MacDonald. ‘Empathetic Speech Synthesis and Testing for Healthcare Robots’. In: *International Journal of Social Robotics* (2020), pp. 1–19 (cited on pages 120, 121).

[511] Matthew P Aylett, Selina Jeanne Sutton, and Yolanda Vazquez-Alvarez. ‘The Right Kind of Unnatural: Designing a Robot Voice’. In: *Proceedings of the 1st International Conference on Conversational User Interfaces*. 2019, pp. 1–2 (cited on pages 121, 131).

[512] Markus Häring, Dieta Kuchenbrandt, and Elisabeth André. ‘Would You like to Play with Me?: How Robots’ Group Membership and Task Features Influence Human-Robot Interaction’. In: *ACM/IEEE International Conference on Human-Robot Interaction, HRI’14, Bielefeld, Germany, March 3-6, 2014*. Ed. by Gerhard Sagerer, Michita Imai, Tony Belpaeme, and Andrea Lockerd Thomaz. ACM, 2014, pp. 9–16. DOI: 10.1145/2559636.2559673 (cited on pages 121, 131).

[513] Dieta Kuchenbrandt, Markus Häring, Jessica Eichberg, Friederike Eyssel, and Elisabeth André. ‘Keep an Eye on the Task! How Gender Typicality of Tasks Influence Human-Robot Interactions’. In: *Int. J. Soc. Robotics* 6.3 (2014), pp. 417–427. DOI: 10.1007/s12369-014-0244-0 (cited on pages 121, 131).

[514] Hannes Ritschel, Ilhan Aslan, Silvan Mertes, Andreas Seiderer, and Elisabeth André. ‘Personalized Synthesis of Intentional and Emotional Non-Verbal Sounds for Social Robots’. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 1–7 (cited on page 121).

[515] Junichi Yamagishi, Christophe Veaux, and Kirsten. MacDonald. *CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (Version 0.92)*. University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2019. DOI: 10.7488/DS/2645 (cited on page 121).

[516] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. ‘Librosa: Audio and Music Signal Analysis in Python’. In: *Proceedings of the 14th Python in Science Conference*. Vol. 8. Citeseer, 2015, pp. 18–25 (cited on page 122).

[517] Marco Costa, Pio Enrico Ricci Bitti, and Luisa Bonfiglioli. ‘Psychological Connotations of Harmonic Musical Intervals’. In: *Psychology of Music* 28.1 (2000), pp. 4–22 (cited on page 122).

[518] Nicolas Ramirez-Guevara. ‘Robotization Effect Using Phase Vocoder Processing’. In: (2017) (cited on page 122).

[519] TAL Software GmbH. *TAL-Vocoder*. 2024 (cited on page 122).

[520] Peter Sobot. *Pedalboard*. Version 0.7.3. Zenodo, 2023. DOI: 10.5281/zenodo.7817839 (cited on page 122).

[521] Paul D. Trapnell and Jerry S. Wiggins. ‘Extension of the Interpersonal Adjective Scales to Include the Big Five Dimensions of Personality.’ In: *Journal of Personality and Social Psychology* 59.4 (1990), pp. 781–790. DOI: 10.1037/0022-3514.59.4.781 (cited on page 122).

[522] RR McCrae and OP John. ‘An Introduction to the Five-Factor Model and Its Applications.’ In: *Journal of personality* 60.2 (1992), p. 175 (cited on page 122).

[523] Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann Jr. ‘A Very Brief Measure of the Big-Five Personality Domains’. In: *Journal of Research in Personality* 37.6 (2003), pp. 504–528. DOI: http://dx.doi.org/10.1016/S0092-6566(03)00046-1 (cited on page 122).

[524] Beatrice Rammstedt and Oliver P. John. ‘Measuring Personality in One Minute or Less: A 10-Item Short Version of the Big Five Inventory in English and German’. In: *Journal of Research in Personality* 41.1 (2007), pp. 203–212. DOI: http://dx.doi.org/10.1016/j.jrp.2006.02.001 (cited on page 122).

[525] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. ‘Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots’. In: *International Journal of Social Robotics* 1.1 (2009), pp. 71–81. DOI: 10.1007/s12369-008-0001-3 (cited on page 122).

[526] Marc Hassenzahl, Michael Burmester, and Franz Koller. ‘AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität’. In: *Mensch & Computer 2003*. Ed. by Gerd Szwillus and Jürgen Ziegler. Vol. 57. Wiesbaden: Vieweg+Teubner Verlag, 2003, pp. 187–196. DOI: 10.1007/978-3-322-80058-9_19 (cited on page 122).

[527] Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Verena Winterhalter, Markus Bühner, and Heinrich Hussmann. 'Developing a Personality Model for Speech-Based Conversational Agents Using the Psycholexical Approach'. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–14. DOI: 10.1145/3313831.3376210 (cited on page 122).

[528] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. *Gephi: An Open Source Software for Exploring and Manipulating Networks*. 2009 (cited on page 125).

[529] Susan R. Fussell, Sara B. Kiesler, Leslie D. Setlock, and Victoria Yew. 'How People Anthropomorphize Robots'. In: *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction, HRI 2008, Amsterdam, the Netherlands, March 12-15, 2008*. Ed. by Terry Fong, Kerstin Dautenhahn, Matthias Scheutz, and Yiannis Demiris. ACM, 2008, pp. 145–152. DOI: 10.1145/1349822.1349842 (cited on page 126).

[530] Julia Cambre and Chinmay Kulkarni. 'One Voice Fits All?: Social Implications and Research Challenges of Designing Voices for Smart Devices'. In: *Proc. ACM Hum. Comput. Interact.* 3 (CSCW 2019), 223:1–223:19. DOI: 10.1145/3359325 (cited on page 128).

[531] Elizabeth Phillips, Xuan Zhao, Daniel Ullman, and Bertram F. Malle. 'What Is Human-like?: Decomposing Robots' Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database'. In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. Hri '18. ACM, 2018. DOI: 10.1145/3171221.3171268 (cited on pages 128, 140).

[532] Ian J. Goodfellow et al. 'Challenges in Representation Learning: A Report on Three Machine Learning Contests'. In: *Neural Information Processing*. Ed. by Minho Lee, Akira Hirose, Zeng-Guang Hou, and Rhee Man Kil. Berlin, Heidelberg: Springer, 2013, pp. 117–124. DOI: 10.1007/978-3-642-42051-1_16 (cited on page 135).

[533] Vikram V. Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. 'GeoDE: A Geographically Diverse Evaluation Dataset for Object Recognition'. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 66127–66137 (cited on page 135).

[534] Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. 'THINGS: A Database of 1,854 Object Concepts and More than 26,000 Naturalistic Object Images'. In: *PLOS ONE* 14.10 (), e0223792. DOI: 10.1371/journal.pone.0223792 (cited on page 135).

[535] Laura Martin. '"Eskimo Words for Snow": A Case Study in the Genesis and Decay of an Anthropological Example'. In: *American Anthropologist* 88.2 (1986), pp. 418–423 (cited on page 135).

[536] Piotr Cichocki and Marcin Kilarski. 'On "Eskimo Words for Snow": The Life Cycle of a Linguistic Misconception'. In: *Historiographia Linguistica* 37.3 (2010), pp. 341–377. DOI: 10.1075/hl.37.3.03cic (cited on page 135).

[537] Lawrence W. Barsalou. 'Grounded Cognition'. In: *Annual Review of Psychology* 59 (Volume 59, 2008 2008), pp. 617–645. DOI: 10.1146/annurev.psych.59.103006.093639 (cited on page 136).

[538] Katie Hoemann. 'Beyond Linguistic Relativity, Emotion Concepts Illustrate How Meaning Is Contextually and Individually Variable'. In: *Topics in Cognitive Science* 15.4 (2023), pp. 668–675. DOI: 10.1111/tops.12659 (cited on page 136).

[539] Eline Van Geert and Nori Jacoby. 'Using Gibbs Sampling with People to Characterize Perceptual and Aesthetic Evaluations in Multidimensional Visual Stimulus Space'. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 46. 2024 (cited on page 136).

[540] Samuel A. Mehr, Manvir Singh, Dean Knox, Daniel M. Ketter, Daniel Pickens-Jones, S. Atwood, Christopher Lucas, Nori Jacoby, Alena A. Egner, Erin J. Hopkins, Rhea M. Howard, Joshua K. Hartshorne, Mariela V. Jennings, Jan Simson, Constance M. Bainbridge, Steven Pinker, Timothy J. O'Donnell, Max M. Krasnow, and Luke Glowacki. 'Universality and Diversity in Human Song'. In: *Science* 366.6468 (2019), eaax0868. DOI: 10.1126/science.aax0868 (cited on page 138).

[541] Tero Karras, Samuli Laine, and Timo Aila. 'A Style-Based Generator Architecture for Generative Adversarial Networks'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.12 (2021), pp. 4217–4228. DOI: 10.1109/TPAMI.2020.2970919 (cited on page 139).

[542] Albert E. Mannes. 'Shorn Scalps and Perceptions of Male Dominance'. In: *Social Psychological and Personality Science* 4.2 (2013), pp. 198–205. DOI: 10.1177/1948550612449490 (cited on page 139).

[543] Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 'Measuring Individual Differences in Implicit Cognition: The Implicit Association Test.' In: *Journal of Personality and Social Psychology* 74.6 (1998), pp. 1464–1480. DOI: 10.1037/0022-3514.74.6.1464 (cited on page 139).

[544] Konrad Schnabel, Jens B. Asendorpf, and Anthony G. Greenwald. 'Using Implicit Association Tests for the Assessment of Implicit Personality Self-Concept'. In: *The SAGE Handbook of Personality Theory and Assessment: Volume 2 — Personality Measurement and Testing*. SAGE Publications Ltd, 2008, pp. 508–528. DOI: 10.4135/9781849200479.n24 (cited on page 139).

[545] Calvin K. Lai et al. 'Reducing Implicit Racial Preferences: I. A Comparative Investigation of 17 Interventions.' In: *Journal of Experimental Psychology: General* 143.4 (2014), pp. 1765–1785. DOI: 10.1037/a0036260 (cited on page 140).

[546] Giulia Perugia, Stefano Guidi, Margherita Bicchi, and Oronzo Parlangeli. 'The Shape of Our Bias: Perceived Age and Gender in the Humanoid Robots of the ABOT Database'. In: *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022. DOI: 10.1109/hri53351.2022.9889366 (cited on page 140).

[547] Londa Schiebinger. 'The Robots Are Coming! But Should They Be Gendered'. In: *AWIS Magazine, Winter* (2019), pp. 18–21, 58 (cited on page 140).

[548] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 'Semantics Derived Automatically from Language Corpora Contain Human-like Biases'. In: *Science* 356.6334 (2017), pp. 183–186. DOI: 10.1126/science.aal4230 (cited on page 140).

[549] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 'Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings'. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, pp. 4356–4364 (cited on page 140).

[550] Beau Sievers, Larry Polansky, Michael Casey, and Thalia Wheatley. 'Music and Movement Share a Dynamic Structure That Supports Universal Expressions of Emotion'. In: *Proceedings of the National Academy of Sciences* 110.1 (2013), pp. 70–75. DOI: 10.1073/pnas.1209023110 (cited on page 141).

[551] Beau Sievers, Caitlyn Lee, William Haslett, and Thalia Wheatley. 'A Multi-Sensory Code for Emotional Arousal'. In: *Proceedings of the Royal Society B: Biological Sciences* 286.1906 (2019), p. 20190513. DOI: 10.1098/rspb.2019.0513 (cited on page 141).

[552] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 'Can AI Language Models Replace Human Participants?' In: *Trends in Cognitive Sciences* 27.7 (2023), pp. 597–600. DOI: 10.1016/j.tics.2023.04.008 (cited on page 141).

[553] Mathew Hardy, Ilia Sucholutsky, Bill Thompson, and Tom Griffiths. 'Large Language Models Meet Cognitive Science: LLMs as Tools, Models, and Participants'. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 45. 45. 2023 (cited on page 141).

[554] Marcel Binz et al. *Centaur: A Foundation Model of Human Cognition*. 2024. DOI: 10.48550/arXiv.2410.20268. Pre-published (cited on page 142).

[555] Stephen E Palmer, Karen B Schloss, Zoe Xu, and Lilia R Prado-León. 'Music-Color Associations Are Mediated by Emotion'. In: *Proceedings of the National Academy of Sciences of the United States of America* 110.22 (2013), pp. 8836–8841 (cited on page 143).

[556] Francis M. Adams and Charles E. Osgood. 'A Cross-Cultural Study of the Affective Meanings of Color'. In: *Journal of Cross-Cultural Psychology* 4.2 (1973), pp. 135–156. DOI: 10.1177/002202217300400201 (cited on page 143).

Abbreviations

A

AI Artificial Intelligence. 78
API Application Programming Interface. 8, 56, 80–83, 105
ARI Adjusted Rand Index. 106
ASR Automatic Speech Recognition. 60, 87

B

BCT Basic Color Term. 102, 103
BLEU Bilingual Evaluation Understudy. 83

C

CIELAB Commission Internationale de l'Eclairage: lightness (L*), chroma (a*), and hue (b*). 101–103
CJK Chinese, Japanese, and Korean. 88
CMYK Cyan, Magenta, Yellow, Key (Black). 101
CPU Central Processing Unit. 80

D

DNN Deep Neural Network. 20, 68, 141

F

f_0 Fundamental frequency (the vibration rate of the vocal folds). 5–8, 117

G

GAP Genetic Algorithm with People. iii, v, 2, 4, 39, 51, 54, 56–60, 133–135, 137, 140–144
GSP Gibbs Sampling with People. iii, v, 2–4, 39, 41–49, 54, 80, 81, 111, 113, 117, 119, 123, 124, 129, 130, 133, 135–138, 140, 141, 143, 144
GST Global Style Token. 46, 47, 114, 115, 117, 118
GWAP Games With A Purpose. 63–65, 72

H

HIT Human Intelligence Task. 56
HITL Human-In-The-Loop. iii, v, 2–5, 17, 18, 22, 27, 39, 41, 42, 51, 63, 65, 72, 74, 111, 113, 119, 133–138, 140–144
HNR Harmonic-to-Noise Ratio. 8, 9, 32
HTML HyperText Markup Language. 79, 81

K

KDE Kernel Density Estimation. 57, 58, 123, 124

L

LLM Large Language Model. 83, 104, 105, 107, 141, 143
Lucid Lucid Marketplace recruitment platform. 3, 79, 81–83, 92–95, 97, 98, 104, 106, 142, 143

M

MCMC Markov Chain Monte Carlo. 21, 42
MCMCP Markov Chain Monte Carlo with People. 21, 22, 42–44, 49, 53, 54
MDS Multidimensional Scaling. 41, 67, 117, 142
MFCC Mel-Frequency Cepstral Coefficients. 8, 9, 32
MOS Mean Opinion Score. 116, 117
MTurk Amazon Mechanical Turk recruitment platform. 3, 43, 45, 48, 49, 56, 70, 78, 79, 116

N

NLP Natural Language Processing. 65, 87, 89, 96

O

OCC Ortony, Clore, and Collins. 13

P

PAD Pleasure-Arousal-Dominance. 14

PCA Principal Component Analysis. 48, 121, 123, 124, 138

POS Part-Of-Speech. 87, 89, 90

Prolific Prolific recruitment platform. 3, 77–79, 81, 82, 92–94, 97, 98, 104, 106, 123, 142

R

RGB Red, Green, Blue. 101, 102, 106

S

SOTA State-Of-The-Art. 67

SSL Secure Sockets Layer. 82

STEP Sequential Transmission Evaluation Pipeline. iii, v, 2–4, 39, 63, 66, 68–70, 72–74, 80, 81, 111, 119, 124–126, 130, 133–135, 137, 139, 141–144

SVM Support Vector Machine. 10

T

TTS Text-To-Speech. 46, 47, 55, 57, 83, 87, 99, 114, 118–121, 130, 131, 133, 137, 138, 142

U

UML Unified Modeling Language. 80

V

VLM Vision Language Model. 105, 106

W

WAIC Widely Applicable Information Criterion. 32, 34, 35

WCS World Color Survey. 102, 103, 107

WEIRD Western, Educated, Industrialized, Rich, and Democratic. 23, 24, 77, 85, 142

WSGI Web Server Gateway Interface. 78

Speech conveys more than just words—it carries emotional cues through prosody, which describes variations in pitch, loudness, timing, and voice quality. Detecting emotions from speech and expressive speech synthesis are crucial for successful communication in human-computer and human-robot interaction. This requires large datasets of emotional recordings. While corpora only indirectly capture the association between prosody and emotions, the actual association is stored in the minds of humans.

This thesis introduces three human-in-the-loop algorithms to efficiently characterize these associations through human experiments:

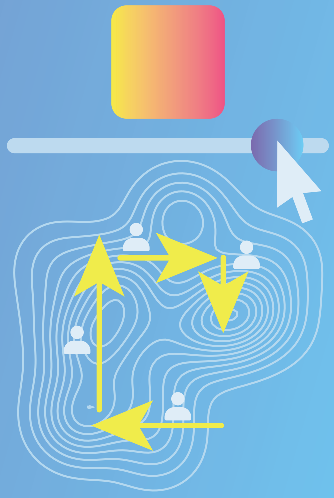
- **Gibbs Sampling with People (GSP):** Participants adjust speech dimensions (e.g., speed) using a slider to match a target emotion. Each participant only changes one dimension at a time. Over iterations, this converges to representative prosodies for particular emotions.
- **Genetic Algorithm with People (GAP):** Participants imitate and refine emotional speech samples through a process of mutation and selection. This yields a diverse set of expressive recordings.
- **Sequential Transmission Evaluation Pipeline (STEP):** Participants label emotional recordings, and can rate the relevance of labels provided by others. This process converges to a weighted taxonomy of emotions expressed through prosody.

To study the associations between emotions and prosody across cultures, I develop an infrastructure to run massive online experiments across the globe. I demonstrate the use and efficacy of it by running a large-scale, cross-lingual experiment.

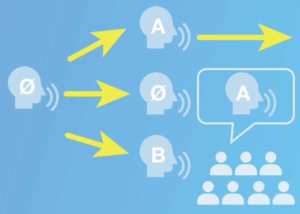
Beyond emotional prosody, these algorithms have broader applications. I show how GSP can be used for voice personalization for digital agents and avatars, and I demonstrate how the combination of GSP and STEP can be used to align impressions of robots across the auditory and visual modality.

In a broader context, these algorithms allow the creation of more representative corpora to train machine learning models that are more balanced and diverse and to benchmark the performance of state-of-the-art models.

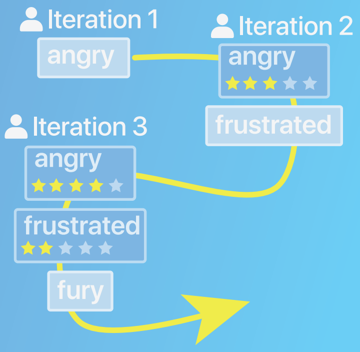
GSP interface



GAP process



STEP process



Computational
Auditory
Perception

Max Planck Institute
for Empirical Aesthetics

Human-Centered
Artificial Intelligence