The ACM Multimedia 2023 Computational Paralinguistics Challenge: emotion share & requests

Björn W. Schuller, Anton Batliner, Shahin Amiriparian, Alexander Barnhill, Maurice Gerczuk, Andreas Triantafyllopoulos, Alice Baird, Panagiotis Tzirakis, Chris Gagne, Alan S. Cowen, Nikola Lackovic, Marie-José Caraty, Claude Montacié

Angaben zur Veröffentlichung / Publication details:

Schuller, Björn W., Anton Batliner, Shahin Amiriparian, Alexander Barnhill, Maurice Gerczuk, Andreas Triantafyllopoulos, Alice Baird, et al. 2023. "The ACM Multimedia 2023 Computational Paralinguistics Challenge: emotion share & requests." In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29 – November 3, 2023, Ottawa, Canada*, edited by Abdulmotaleb El Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain, 9635–39. New York, NY: ACM. https://doi.org/10.1145/3581783.3612835.

The ACM Multimedia 2023 Computational Paralinguistics Challenge: Emotion Share & Requests

Björn W. Schuller
GLAM, Imperial College London
London, United Kingdom

Anton Batliner
EIHW, University of Augsburg
Augsburg, Germany

Shahin Amiriparian
EIHW, University of Augsburg
Augsburg, Germany

Alexander Barnhill
FAU
Erlangen-Nuremberg, Germany

Maurice Gerczuk
EIHW, University of Augsburg
Augsburg, Germany

Andreas Triantafyllopoulos
EIHW, University of Augsburg
Augsburg, Germany

Alice Baird
Hume AI
New York, USA

Panagiotis Tzirakis
Hume AI
New York, USA

Chris Gagne
Hume AI
New York, USA

Alan S. Cowen
Hume AI
New York, USA

Nikola Lackovic
Malakoff Humanis
Paris, France

Marie-José Caraty
Sorbonne University
Paris, France

Claude Montacié
Sorbonne University
Paris, France

ABSTRACT

The ACM Multimedia 2023 Computational Paralinguistics Challenge addresses two different problems for the first time in a research competition under well-defined conditions: In the *Emotion Share* Sub-Challenge, a regression on speech has to be made; and in the *Requests* Sub-Challenges, *requests* and *complaints* need to be detected. We describe the Sub-Challenges, baseline feature extraction, and classifiers based on the 'usual' ComParE features, the auDeep toolkit, and deep feature extraction from pre-trained CNNs using the DeepSpectrum toolkit; in addition, wav2vec2 models are used.

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**; • **Computing methodologies → Artificial intelligence**.

KEYWORDS

Computational Paralinguistics; Emotion Share; Requests; Complaints; Challenge; Benchmark

ACM Reference Format:
Björn W. Schuller, Anton Batliner, Shahin Amiriparian, Alexander Barnhill, Maurice Gerczuk, Andreas Triantafyllopoulos, Alice Baird, Panagiotis Tzirakis, Chris Gagne, Alan S. Cowen, Nikola Lackovic, Marie-José Caraty, and Claude Montacié. 2023. The ACM Multimedia 2023 Computational Paralinguistics Challenge: Emotion Share & Requests. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, Canada.* ACM, New York, NY, USA, 5 pages.

MM '23, October 29–November 3, 2023, Ottawa, Canada

1 INTRODUCTION

In this ACM Multimedia 2023 COMputational PARalinguistics challengE (ComParE) – the 15th since 2009 [22, 23], we address two new problems within the field of Computational Paralinguistics [21] in a challenge setting:

1.1 Emotion Share Sub-Challenge

As the ComParE challenge has been making efforts to address for the past decade [23], the lack of data in the domain of emotion modelling continues to be an issue for real-world development, particularly with the rise of data-hungry deep learning methods [6]. To address this, for the **Emotion Share Sub-Challenge**, Hume AI has provided the Hume-Prosody dataset. The Sub-Challenge features a multi-label regression task. For each of the nine different emotions, a proportion or 'share' has been assigned to the nine emotions based on the proportion of raters that rated that emotion for the 'seed' sample. For further details on the data collection methodology, see [7] and Section 2.1.

1.2 Requests Sub-Challenge

The interaction between an organisation and its customers (Customer Relationship Management CRM) [5] takes often place within a call centre [28], e. g., via a phone call. Speech analysis [13, 20] helps to model this interaction and the interest of the users as well as to pinpoint problems. In the **Requests Sub-Challenge**, we are interested in two tasks related to CRM: the classification of customer *Requests* and *Complaints*. Data and annotations have been provided by STIH Laboratory, Sorbonne University.

1.3 Tasks

For these two challenges, either a target class has to be predicted (for Request and Complaints), or a correlation measure has to be computed (for Emotion Share). Contributors can employ their own features and machine learning (ML) algorithms; standard feature sets and procedures are provided. Participants have to use the pre-defined partitions for each Sub-Challenge. They may report results that they obtain from the **Train**(ing)/**Dev**(elopment) set but have only five trials to upload their results on the **Test** set per Sub-Challenge, whose labels are unknown to them. Each participation must be accompanied by a paper presenting the results, which undergoes peer-review. The organisers preserve the right to re-evaluate the findings, but will not participate in the Challenge. As evaluation measure, we employ the **Unweighted Average Recall (UAR)** for the Requests Sub-Challenges as used since the first Challenge from 2009 [22, 23]; it is more adequate for (unbalanced) multi-class classifications than Weighted Average Recall (i. e., accuracy) [19, 21]. For the Emotion-Share Sub-Challenge, Spearman's ρ [27] is used as most adequate measure for such ranking values. Ethical approval for the studies has been obtained.

2 THE TWO SUB-CHALLENGES

2.1 Emotion Share – The Hume-Prosody Corpus HP-C

The basis for the dataset is more than 5,000 'seed' samples. Seeds consist of various emotional expressions (e. g., *'Tal vez, sea verdad'*), which were gathered from openly available datasets including MELD [18] and VENEC [10, 15, 16]. The seeds are a mixture of 'same' and 'different' sentences – e. g., more than 500 instances of *'Let me tell you something'* [15], where the functional load of prosody is high, and 'different' sentences (each of them with different words and semantics), where the functional load of prosody is lower.

The seed samples were mimicked by speakers recruited via Amazon Mechanical Turk [9]. The Sub-Challenge subset consists of 51,881 'mimic' samples (total of 41:48:55 h of data, mean 2.9 s., range 1.2 – 7.98 s) from 1,004 speakers aged from 20 to 66 years old. It was gathered in 3 countries with broadly differing cultures: the United States, South Africa, and Venezuela. For data processing, files below 1 s and above 8 s were excluded. The data were recorded at home via the speakers' microphones. The full Hume-Prosody dataset consists of 48 dimensions of emotional expression and is based on the semantic-space model for emotion [8]. For this Sub-Challenge, nine emotional classes have been selected due to their more balanced distribution across the valence-arousal space: 'Anger', 'Boredom', 'Calmness', 'Concentration', 'Determination', 'Excitement', 'Interest', 'Sadness', and 'Tiredness'.

Each seed sample was rated by the individual that imitated it using a 'select-all-that-apply' method [9]. Seeds were assigned a mean of 2.03 emotions per rater (max: 7.11, min: 1.00), with a standard deviation of 1.33 emotions. The proportion of times a given seed sample was rated with each emotion was then applied to all mimics of that seed sample. This results in the **share** per emotion assigned by the speakers. For the Sub-Challenge baseline, the labels have been scaled to a maximum of 1 by dividing by the maximum emotion value per sample across the nine emotions.

2.2 Requests – The HealthCall30 Corpus HC-C: Complaints and Requests

This is a subset (audio-only) of the HealthCall30 corpus, provided by Montacié and colleagues [14]. It is based on real audio interactions between call centre agents and customers who call to solve a problem or to request information. This corpus is designed to study natural spoken conversations and to predict Customer Relationship Management (CRM) annotations made by human agents from various vocal interaction, audio, and linguistic features. The corpus consists of 13,409 chunks of spoken conversations, each lasting 30 seconds. Two different classifications of these conversations were performed by annotators, based on CRM annotations: the presence of customer complaints (*Complaint* – *"yes"* or *"no"*) and the type of customer request (*Request*) – either concerning membership issues (*affil*) or a process (*presta*), e. g., a reimbursement. Each conversation was recorded on two separate and distinct audio channels; the first channel corresponds to the customer's audio, and the second corresponds to the agent's audio. More information can be found in [14]. For the challenge, we provide both the raw dual channel audio files, as well as the normalised mono-conversions, as utilised for feature extraction and wav2vec2 training in the baseline.

3 EXPERIMENTS AND RESULTS

Table 1 shows the number of data for Train, Dev, and Test for the different corpora.

3.1 Approaches

This year, we evaluate four baseline systems. Of those, three consist of a distinct feature extraction step followed by a linear Support Vector Machine (SVM) while the last system employes a pre-trained wav2vec2 model, taking raw audio as input. For a comprehensive view of the chosen hyperparameters, the reader is referred to the official baseline repository.

3.1.1 ***ComParE Acoustic Feature Set:*** The official baseline feature set from openSMILE is the same as has been used in previous editions of the ComParE challenges, starting from 2013 [11, 24].

3.1.2 ***DeepSpectrum:*** This toolkit[1] is applied to obtain deep representations from the input audio data utilising image pre-trained Convolutional Neural Networks (CNNs) [1, 3]. has been used in previous challenges [25, 26] and is described in [3]. The efficacy of DeepSpectrum features have been demonstrated for speech and audio recognition tasks [4]. For this iteration of the challenge, we utilise DenseNet169 to extract features from Mel-spectrograms with 128 bands.

3.1.3 ***auDeep:*** This toolkit[2] is obtained through unsupervised representation learning with recurrent sequence-to-sequence autoencoders [2, 12]. We choose a two-layer architecture utilising

[1]https://github.com/DeepSpectrum/DeepSpectrum
[2]https://github.com/auDeep/auDeep

Table 1: Summary of the databases presented per Sub-Challenge. Number of instances per class in the Train/Dev/Test splits. Test split distributions for HC-C Requests and HC-C Complaints are blinded during the ongoing challenge.

HC-C Requests: classification task (#)					**HC-C Complaints**: classification task (#)					**HP-C**: regression task				
Class	Train	Dev	Test	Σ	Class	Train	Dev	Test	Σ		Train	Dev	Test	Σ
affil	3,690	1,552	––	––	Yes	2,522	1,131	––	––	sample no.	30,133	12,241	9,507	51,881
presta	3,132	1,532	––	––	No	4,300	1,953	––	––	speaker no.	600	202	202	1,004
Σ	6,822	3,084	3,503	13,409	Σ	6,822	3,084	3,503	13,409	gender (f:m)	379:221	117:85	141:61	637:367

Table 2: Results for the Sub-Challenges. The official best results for Test yielding the official baselines are highlighted (bold and greyscale); there are no official baselines for Dev. UAR: Unweighted Average Recall. CI on Test: Confidence Intervals on Test, see explanation in the text.

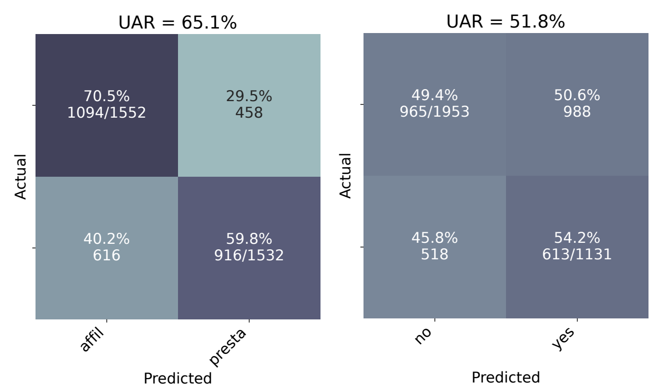
[UAR %]	**HC-C Request**			**HC-C Complaint**			**HP-C** [ρ]		
Approach	Dev	Test	CI on Test	Dev	Test	CI on Test	Dev	Test	CI on Test
wav2vec2	65.1	**67.2**	65.7 – 68.7	50.9	52.2	50.7 – 53.7	.500	**.514**	.499 – .529
auDeep	54.9	52.4	50.7 – 54.0	50.9	50.9	49.0 – 52.7	.347	.357	.341 – .374
DeepSpectrum	58.2	55.6	54.0 – 57.2	49.6	51.7	50.1 – 53.4	.335	.331	.313 – .349
ComParE	60.9	58.8	57.2 – 60.5	51.8	**52.9**	51.2 – 54.7	.359	.365	.347 – .382
Late Fusion	60.5	59.5	57.8 – 61.2	51.2	51.8	50.1 – 53.5	.470	.476	.461 – .492

Gated Recurrent Unit (GRU) cells with 256 hidden units and a latent feature size of 1024 and train it for 64 epochs on Mel-spectrograms with 128 Mel-bands. Four variants of this autoencoder are trained, each clipping out low amplitudes from the input signals below -30 dB, -45 dB, -60 dB and -75 dB, respectively. The final features are obtained by concatenating the hidden representations of these four autoencoders.

3.1.4 ***SVM:*** We train and evaluate SVMs with linear kernels on the three feature sets described above. The choice of feature normalisation, either min-max scaling or normalisation to zero mean and unit variance, is jointly optimised with the cost parameter C of the SVM based on the performance on the Dev set. After this optimisation, a final model is fit on the concatenated training and Dev sets for evaluation on the Test partition.

3.1.5 ***Wav2Vec2:*** For HC-C Requests and HC-C Complaints, we fine-tune a pre-trained Wav2Vec2 model for the challenge tasks on the raw audio files. We obtain the weights for the XLSR pre-trained model from huggingface hub [3]. We freeze the convolutional feature extractor and add final layers for classification to the output of the Transformer encoder. The model is trained for a maximum of 15 epochs with a batch size of 64 and learning rate of $3e^{-4}$. The best model, measured by Unweighted Average Recall (UAR) on the development set, is saved and restored for the final evaluation on the test partition. For Emotion share, no fine-tuning for the challenge tasks was done – instead, we use a model fine-tuned on the MSP-Podcast [17] dataset[4] to extract features; these are then fed to an SVM following the same procedure as detailed in Section 3.1.4.

[3]https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english
[4]https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim



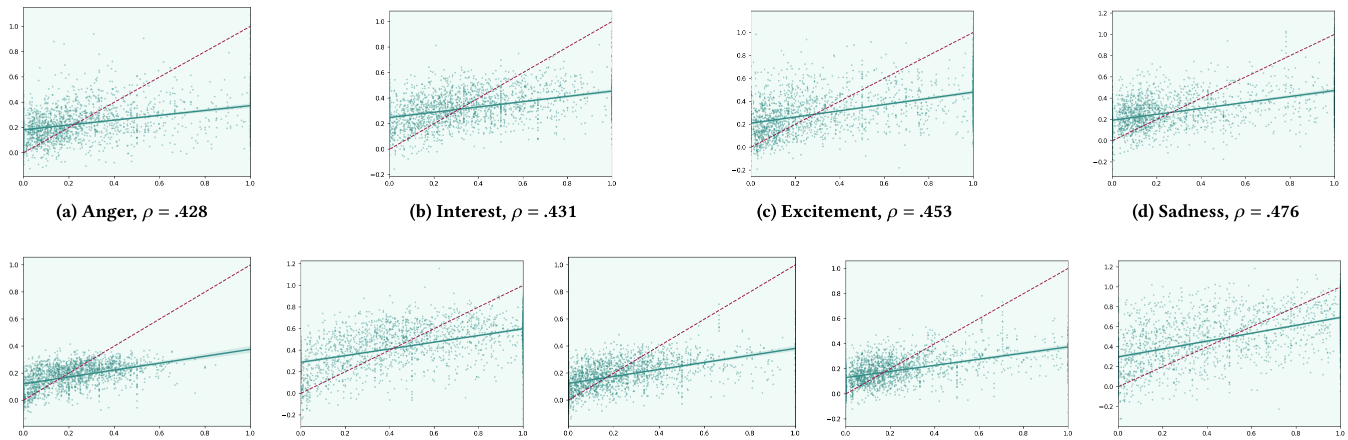
(a) HC-C Requests Confusion Matrix **(b) HC-C Complaints Confusion Matrix**

Figure 1: Confusion matrices for (a) HC-C Requests and (b) HC-C Complaints; the individual approach/hyperparameters performing best on Dev (without fusion) are chosen; see Table 2. In the cells, percents of 'classified as' of the class displayed in the respective row are given, also indicated by colour-scale: the darker, the higher. Cases per class given in Table 1.

3.2 Challenge Baselines and Interpretation

We provide a branch on the official challenge repository[5] for each Sub-Challenge, which includes scripts allowing participants to fully reproduce the baselines (including pre-processing, model training, and model evaluation on Dev). For Requests and Complaints, the

[5]https://github.com/EIHW/ComParE2023



(a) Anger, ρ = .428 (b) Interest, ρ = .431 (c) Excitement, ρ = .453 (d) Sadness, ρ = .476

(e) Concentration, ρ = .524 (f) Determination, ρ = .531 (g) Boredom, ρ = .545 (h) Tiredness, ρ = .550 (i) Calmness, ρ = .559

Figure 2: Emotion Share; Scatterplots for each emotion; reference ('true') values on x-axis, predicted values on y-axes; green: regression line; red: 'perfect prediction'; random selection of data points for better visibility; emotions ordered by size of ρ; for wav2vec2 on Dev, see Table 2.

95 % Confidence Intervals (CI) were computed by 1 000x bootstrapping (random sampling with replacement) based on the same model that was trained with Train and Dev, and UARs for Test. For Emotion Share, the 95 % Confidence Intervals (CI) were also computed the same way, with Spearmans's correlation score for Test. (Note that these CIs are too optimistic because we model single instances which are, however, partly not independent because more than one instance could have been produced by the same speaker.)

3.3 The Requests Sub-Challenge

Requests: We obtain the best **UAR=67.2 %** on Test with wav2vec2, see Table 2. Figure 1(a) shows, for the best Dev result given in Table 2, that affil (requests concerning membership issues) can be better modelled than presta (type of process). The reason might be that affil provides less variance in linguistic content than presta, where different types of processes have to be be modelled.
Complaints: We obtain the best **UAR=52.9 %** on Test with ComParE, see Table 2. Figure 1(b) shows, for the best Dev result given in Table 2, a rather balanced distribution. Yet, UAR is rather low and not really different from chance level.

Overall baseline for the Requests Sub-Challenge is the combined best UAR of Requests and Complaints: (67.2 + 52.9) / 2 = **60.1 %**. Note that our baselines are computed with only acoustic information – concerning wav2vec2, only with the implicit linguistic information entailed in this procedure. Thus, an additional processing of the linguistic content might surely improve performance.

3.4 The Emotion Share Sub-Challenge

We achieve a best UAR on Test of ρ = **.514** for wav2vec2 which is markedly better than the other three procedures, see Table 2. Note that the data consist of 'same' and 'different' utterances. As we mentioned above in Section 2, especially for the 'different' sentences, where the emotional content might at least partly be conveyed with words, the linguistic information entailed in wav2vec2 might contribute to this difference.

Figure 2 displays scatterplots and regression lines for each emotion, obtained for Dev with wav2vec2. We see that especially calmness and determination are distributed over all reference values. In contrast, especially the prototypical emotions (two of the big *n*) anger and sadness – we can attribute excitement to these emotions as well – are below ρ = .50. This might be explained by the experimental design and the choice of samples which do not favour such prototypical, rather extreme emotions, in contrast to less pronounced ones. The skewed distribution might as well be responsible for the slightly lower performance of the prototypical emotions.

4 CONCLUDING REMARKS

This year's challenge is new by two new tasks, all of them highly relevant for applications. We feature our 'classic' approaches **ComParE**, **auDeep**, and **DeepSpectrum**, and introduce (fine-tuning of) **Wav2Vec2** model as an additional baseline. For all computation steps, scripts are provided that can, but need not be used by the participants. We expect participants to obtain better performance measures by employing novel (combinations of) procedures and features, including such tailored to the particular tasks. For both Sub-Challenges – maybe more for the Requests Sub-Challenge – additional linguistic modelling might improve performance as well.

5 ACKNOWLEDGMENTS

We acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG) under grant agreement No. 421613952 (ParaStiChaD) as well as from the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS).

REFERENCES

[1] Shahin Amiriparian. 2019. *Deep Representation Learning Techniques for Audio Signal Processing*. Ph.D. Dissertation. Technische Universität München.

[2] Shahin Amiriparian, Michael Freitag, Nicholas Cummins, and Björn Schuller. 2017. Sequence to Sequence Autoencoders for Unsupervised Representation Learning from Audio. In *Proc. DCASE 2017*. IEEE, Munich, Germany, 17–21.

[3] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, and Björn Schuller. 2017. Snore Sound Classification Using Image-based Deep Spectrum Features. In *Proc. Interspeech*. ISCA, Stockholm, Sweden, 3512–3516.

[4] Shahin Amiriparian, Maurice Gerczuk, Lukas Stappen, Alice Baird, Lukas Koebe, Sandra Ottl, and Björn Schuller. 2020. Towards Cross-Modal Pre-Training and Learning Tempo-Spatial Characteristics for Audio Recognition with Convolutional and Recurrent Neural Networks. *EURASIP Journal on Audio, Speech, and Music Processing* 2020, 19 (2020), 1–11.

[5] Muhammad Anshari, Mohammad Nabil Almunawar, Syamimi Ariff Lim, and Abdullah Al-Mudimigh. 2019. Customer relationship management and big data enabled: Personalization & customization of services. *Applied Computing and Informatics* 15 (2019), 94–101.

[6] Anton Batliner, Simone Hantke, and Björn Schuller. 2020. Ethics and good practice in computational paralinguistics. *IEEE Transactions on Affective Computing* 13, 3 (2020), 1236–1253.

[7] Alan S. Cowen, Hillary Anger Elfenbein, Petri Laukka, and Dacher Keltner. 2019. Mapping 24 emotions conveyed by brief human vocalization. *Am Psychol.* 74 (2019), 698–712.

[8] Alan S Cowen and Dacher Keltner. 2021. Semantic space theory: A computational approach to emotion. *Trends in Cognitive Sciences* 25, 2 (2021), 124–136.

[9] Alan S Cowen, Petri Laukka, Hillary Anger Elfenbein, Runjing Liu, and Dacher Keltner. 2019. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature human behaviour* 3, 4 (2019), 369–382.

[10] Hillary Anger Elfenbein, Petri Laukka, Jean Althoff, Wanda Chui, Frederick K Iraki, Thomas Rockstuhl, and Nutankumar S. Thingujam. 2022. What Do We Hear in the Voice? An Open-Ended Judgment Study of Emotional Speech Prosody. *Personality and Social Psychology Bulletin* 48, 7 (2022), 1087–1104.

[11] Florian Eyben, Felix Weninger, Florian Groß, and Björn Schuller. 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In *Proc. ACM Multimedia*. ACM, Barcelona, Spain, 835–838.

[12] Michael Freitag, Shahin Amiriparian, Sergey Pugachevskiy, Nicholas Cummins, and Björn Schuller. 2018. auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks. *Journal of Machine Learning Research* 18 (2018), 1–5.

[13] Christian Hildebrand, Fotis Efthymiou, Francesc Busquet, William H. Hampton, Donna L. Hoffman, and Thomas P. Novak. 2020. Voice analytics in business research. Conceptual foundations, acoustic feature extraction, and applications. *Journal of Business Research* 121 (2020), 364–374.

[14] Nikola Lackovic, Claude Montacié, Gauthier Lalande, and Marie-José Caraty. 2022. Prediction of User Request and Complaint in Spoken Customer-Agent Conversations. *preprint arXiv:2208.10249* (2022).

[15] Petri Laukka, Hillary Anger Elfenbein, Wanda Chui, Nutankumar S. Thingujam, Frederick K. Iraki, Thomas Rockstuhl, and Jean Althoff. 2010. Presenting the VENEC corpus: Development of a cross-cultural corpus of vocal emotion expressions and a novel method of annotating emotion appraisals. In *Proc. LREC 2010 Workshop on Corpora for Research on Emotion and Affect*. LREC, Marrakesh, Morocco, 53–57.

[16] Petri Laukka, Hillary Anger Elfenbein, Nutankumar S Thingujam, Thomas Rockstuhl, Frederick K Iraki, Wanda Chui, and Jean Althoff. 2016. The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features. *Journal of personality and social psychology* 111, 5 (2016), 686.

[17] Reza Lotfian and Carlos Busso. 2019. Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings. *IEEE Transactions on Affective Computing* 10, 4 (2019), 471–483.

[18] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *preprint arXiv:1810.02508* (2018).

[19] Andrew Rosenberg. 2012. Classifying skewed data: importance weighting to optimize average recall. In *Proc. Interspeech*. ISCA, Portland, USA, 2242–2245.

[20] Scott Scheidt and Q. B. Chung. 2019. Making a case for speech analytics to improve customer service quality: Vision, implementation and evaluation. *International Journal of Information Management* 45 (2019), 223–232.

[21] Björn Schuller and Anton Batliner. 2014. *Computational Paralinguistics – Emotion, Affect, and Personality in Speech and Language Processing*. Wiley, Chichester, UK.

[22] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge. *Speech Communication* 53 (2011), 1062–1087.

[23] Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The INTERSPEECH 2009 Emotion Challenge. In *Proc. Interspeech*. ISCA, Brighton, UK, 312–315.

[24] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim. 2013. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In *Proc. Interspeech*. ISCA, Lyon, France, 148–152.

[25] Björn W. Schuller, Anton Batliner, Christian Bergler, Cecilia Mascolo, Jing Han, Iulia Lefter, Heysem Kaya, Shahin Amiriparian, Alice Baird, Lukas Stappen, Sandra Ottl, Maurice Gerczuk, Panagiotis Tzirakis, Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, Leon J. M. Rothkrantz, Joeri Zwerts, Jelle Treep, and Casper Kaandorp. 2021. The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates. In *Proc. Interspeech*. ISCA, Brno, Czechia, 431–435.

[26] Björn W Schuller, Anton Batliner, Christian Bergler, Eva-Maria Messner, Antonia Hamilton, Shahin Amiriparian, Alice Baird, Georgios Rizos, Maximilian Schmitt, Lukas Stappen, et al. 2020. The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks. In *Proc. Interspeech*. ISCA, Shanghai, China, 2042–2046.

[27] Charles Spearman. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* 15 (1904), 72–101.

[28] Anita Whiting and Naveen Donthu. 2006. Managing voice-to-voice encounters: reducing the agony of being put on hold. *Journal of Service Research* 8 (2006), 234–244.