

Guest editorial: special issue on affective speech and language synthesis, generation, and conversion

Shahin Amiriparian, Bjorn W. Schuller, Nabiha Asghar, Heiga Zen, Felix Burkhardt

Angaben zur Veröffentlichung / Publication details:

Amiriparian, Shahin, Bjorn W. Schuller, Nabiha Asghar, Heiga Zen, and Felix Burkhardt. 2023. "Guest editorial: special issue on affective speech and language synthesis, generation, and conversion." IEEE Transactions on Affective Computing 14 (1): 3-5. <https://doi.org/10.1109/taffc.2022.3233120>.

Nutzungsbedingungen / Terms of use:

licgercopyright

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:

Deutsches Urheberrecht

Weitere Informationen finden Sie unter: / For more information see:

<https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publiz/>



Guest Editorial: Special Issue on Affective Speech and Language Synthesis, Generation, and Conversion

Shahin Amiriparian , Björn W. Schuller, *Fellow, IEEE*, Nabiha Asghar, Heiga Zen, and Felix Burkhardt

As an inseparable and crucial part of spoken language, emotions play a substantial role in human-human and human-technology conversation. They convey information about a person's needs, how one feels about the objectives of a conversation, the trustworthiness of one's verbal communication, and more. Accordingly, substantial efforts have been made to generate affective text and speech for conversational AI, artificial storytelling, and machine translation. Similarly, there is a push for converting the affect in text and speech, ideally, in real-time and fully preserving intelligibility, e. g., to hide one's emotion, for creative applications and in entertainment, or even to augment training data for affect analysing AI. The rapid development of deep neural networks has increased the ability of computers to produce natural speech and language in many languages. Novel methodologies, including attention-based and sequence-to-sequence Text-to-Speech (TTS), have shown promise in synthesising high-quality speech directly from text inputs. However, most TTS systems do not convey the emotional context that is omnipresent in human-human interaction. The lack of emotions in the generated speech can be assumed as a major reason for the low perceived likeability of such systems. Conversely, generative models such as WaveNet and its derivatives, which use raw waveforms of the audio signals instead of the text input for speech generation, can help to condition the emotions of the produced speech. Further, variations of generative adversarial

networks (GANs), such as StarGAN have been successfully applied for speech-based emotion conversion and generation. Similarly, in affective natural language generation and conversion, deep-learning approaches have considerably changed the landscape and opened up new abilities based on massive language corpora and models. Yet, applications are to come at large, featuring human-like real-time generation and conversion of affect in spoken and written language. As the research in this field is still in its infancy, the goal here is to provide the current state-of-the-art in this field and identify any promising new research areas. We, therefore, provide a new perspective when designing neural speech and language synthesis, generation, and conversion models that consider human affects for a more natural human-AI interaction and a rich plethora of further applications.

1 FEATURED WORKS

First, Pataca and Costa in "Hidden bawls, whispers, and yelps: Can text convey the sound of speech, beyond words" [1] propose a novel approach to enrich the typography of text representations by incorporating prosodic information into automatic text transcriptions of spoken utterances. Their methodology allows for the transcription of both the words and the paralinguistic components of an utterance by transforming acoustic cues from speech into visual variations of typography. The introduced approach relies on computing three key prosodic features *magnitude*, *pitch*, and *duration*. These features are then used to modulate the transcription font by applying the following transformations:

- 1) magnitude $\circ \text{---} \bullet$ font-weight, i. e., thickness or "boldness" of a letter
- 2) pitch $\circ \text{---} \bullet$ baseline shift, i. e., vertical displacement of each letter
- 3) duration/rhythm $\circ \text{---} \bullet$ letter-spacing, i. e., how much space is there between each letter

Following that, the authors conduct user research to assess how effectively human volunteers (117 participants) who had no previous training can utilise their approach to select the proper clip (from two candidate audio clips) that matches an annotated speech. The results demonstrate that

-
- Shahin Amiriparian is with the Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, 86159 Augsburg, Germany. E-mail: shahin.amiriparian@uni-a.de.
 - Björn W. Schuller is with the Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, 86159 Augsburg, Germany, and with the GLAM – the Group on Language, Audio, & Music, Imperial College London, SW7 2BX London, U.K., and also with the audEERING GmbH, 82205 Gilching, Germany. E-mail: bjoern.schuller@uni-a.de.
 - Nabiha Asghar is with Microsoft, Redmond, WA 98052-6399 USA. E-mail: ch.nabiha@gmail.com.
 - Heiga Zen is with Google Research, Tokyo 105-0004, Japan. E-mail: heigazen@gmail.com.
 - Felix Burkhardt is with the Speech Communication, Technical University of Berlin, 10623 Berlin, Germany, and also with the audEERING GmbH, 82205 Gilching, Germany. E-mail: f.burkhardt@tu-berlin.de.

(Corresponding author: Shahin Amiriparian.)

Digital Object Identifier no. 10.1109/TAFFC.2022.3233120

humans can address this task with an average accuracy rate of 65 %. The authors also display and analyse participant (free text) replies.

This paper fits the theme, as it can be seen as a form of text conversion in an attempt to incorporate prosodic information that can help the reader derive the emotionality of the original speech. The finding in this work can help develop applications that assist hearing-impaired individuals in better understanding latent emotions in the transcribed speech. Further, the (multimodal) affective computing community can benefit from this study which provides valuable insights into how prosodic information can be embedded in the text.

Next, Hu, Huang, Hu, and Xu introduce “The Acoustically Emotion-aware Conversational Agent with Speech Emotion Recognition and Empathetic Responses” [2]. They equip a conversational agent with a speech emotion recognition component and different strategies to create emotionally appropriate responses considering the detected user emotion. More specifically, the authors experiment with 1) adding interjections such as “Ha-ha” to initially neutral responses, and 2) returning specifically designed empathetic responses. The thus obtained textual responses are then synthesised by a TTS system. The responses aim at reinforcing a positive or alleviating a negative mood, respectively.

The approach is evaluated via a user study based on a game moderated by either variants of the emotional conversational agent or a non-emotional control agent. In this scenario, both positive and negative emotions are elicited in 75 subjects, as verified by measuring the mouse click pressure. Afterwards, the participants are interviewed regarding the perceived emotional abilities of the agent. Overall, the interviews confirm that the proposed approaches lead to the agent being experienced as more emotionally intelligent than the non-emotional baseline. The mouse click pressure measurements indicate that the proposed systems are effective in influencing the user’s emotion in the desired manner.

Since this study is concerned with generating emotionally appropriate responses to (potentially) emotional input speech, it is a valuable contribution to the special issue at hand. Especially, it highlights the necessity of the interplay of recognition and synthesis components in empathetic user agents and demonstrates that user experience in human-computer interaction can benefit considerably from emotionally informed systems.

Zhou, Sisman, Rana, Schuller, and Li in “Emotion Intensity and its Control for Emotional Voice Conversion” [3] propose a novel sequence-to-sequence emotional voice conversion (EVC) framework, denoted as *Emovox* for explicit characterisation and control of emotions’ intensity. EVC aims to convert the emotional state of an utterance while preserving the linguistic content and speaker identity. It typically treats emotions as discrete categories, disregarding the fact that speech conveys emotions with various intensity levels that the listener can perceive. This contribution seeks to explicitly characterise and control the intensity of emotion by disentangling the speaker’s style from the linguistic content and encoding it into a style embedding in a continuous space, forming the prototype of emotion embedding. *Emovox* learns actual emotion encoder from an emotion-labelled database and uses relative attributes to represent

fine-grained emotion intensity. To ensure emotional intelligibility, the authors incorporate emotion classification loss and emotion embedding similarity loss into the training of their EVC framework. The proposed architecture successfully controls the fine-grained emotion intensity in the output speech, as validated by objective and subjective evaluations. Moreover, the paper investigates how emotion intensity interacts with various prosodic features such as *speech duration*, *pitch envelope*, and *speech energy* and analysis their importance for the creation of an emotional response. Whilst *Emovox* does not need a large amount of emotional speech data for the sequence-to-sequence EVC training, it still achieves remarkable performance.

This study fits well the theme of the special issue since it properly addresses some of the research gaps in the affective computing community, including the lack of studies on emotion intensity control for achieving better emotional intelligence, and the lack of focus on modelling prosody style for enhanced regulation of emotion intensity.

2 CHALLENGES AND PERSPECTIVES

The three presented studies address very promising and challenging applications of affective computing, including (i) speech-to-text conversion incorporating prosodic information that can help the reader derive the emotionality of speech, (ii) interplay of emotion recognition and synthesis for more naturalistic and empathetic human-technology interaction, and (iii) characterisation and control of emotions’ intensity for generation of emotional responses. We believe that the technologies proposed in the accepted manuscripts are ready to be used in longer-term research that will demonstrate their efficacy for ‘real-world’ affective human-human-AI interaction. We hope that other researchers and developers will follow the benchmark set by the articles in this special issue and provide insights that can help build affective computing systems improving the quality of our daily life. We also hope that you will find this special issue on affective speech and language synthesis, generation, and conversion as interesting and engaging to read as we did while preparing it.

ACKNOWLEDGMENTS

The guest editors of this special issue would like to thank all the reviewers for their time and their insightful and valuable comments.

REFERENCES

- [1] C. d. L. Pataca and P. D. P. Costa, “Hidden bawls, whispers, and yelps: Can text convey the sound of speech, beyond words,” *IEEE Trans. Affective Comput.*, to be published, doi: [10.1109/TAFFC.2022.3174721](https://doi.org/10.1109/TAFFC.2022.3174721).
- [2] J. Hu, Y. Huang, X. Hu, and Y. Xu, “The acoustically emotion-aware conversational agent with speech emotion recognition and empathetic responses,” *IEEE Trans. Affective Comput.*, to be published, doi: [10.1109/TAFFC.2022.3205919](https://doi.org/10.1109/TAFFC.2022.3205919).
- [3] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, “Emotion intensity and its control for emotional voice conversion,” *IEEE Trans. Affective Comput.*, to be published, doi: [10.1109/TAFFC.2022.3175578](https://doi.org/10.1109/TAFFC.2022.3175578).



Shahin Amiriparian received the doctoral degree with the highest honours (summa cum laude) from the Technical University of Munich, Germany, in 2019. Currently, he is a postdoctoral researcher with the chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany. His main research focus is deep learning, unsupervised representation learning, and transfer learning for affective computing, machine perception, and audio understanding.



Björn W. Schuller (Fellow, IEEE) received the diploma, in 1999, the doctoral degree, in 2006, and the Habilitation and Adjunct Teaching Professorship in the subject area of signal processing and machine intelligence, in 2012, all in electrical engineering and information technology from Technische Universität München (TUM), Munich, Germany. He is professor of AI and head of GLAM, Imperial College London, U.K., and full professor and head of the chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany. He (co-)authored five books and more than 1 000 publications in peer-reviewed books, journals, and conference proceedings leading to more than 50 000 citations (h-index > 100). He is fellow of the AAAC, BCS, ELLIS, and ISCA.



Nabiha Asghar received the PhD degree in computer science from the University of Waterloo, Canada, in 2020. Her research focuses on developing empathetic and emotionally cognizant conversational systems using neural-network-based NLP techniques as well as established socio-mathematical frameworks like Affect Control Theory. Her work has been published with prominent venues like ICLR, UAI, and ECIR, and she routinely serves on the program committees of several AI conferences. She is currently an applied machine learning scientist with Microsoft USA, where she develops massive-scale NLP and graph-based ML models used by billions of people.



Heiga Zen received the AE degree from the Suzuka National College of Technology, Suzuka, Japan, in 1999, and the PhD degree from the Nagoya Institute of Technology, Nagoya, Japan, in 2006. He was an Intern/Co-Op researcher with the IBM T.J. Watson Research Center, Yorktown Heights, NY (2004–2005), and a research engineer with Toshiba Research Europe Ltd. Cambridge Research Laboratory, Cambridge, UK (2008–2011). With Google, he was in the Speech team from July 2011 to July 2018, then joined the Brain team in August 2018. His research interests include speech technology and machine learning. He was one of the original authors and the first maintainer of the HMM-based speech synthesis system (HTS).



Felix Burkhardt is a contract teacher with the Chair of Speech Communication, TU Berlin since October 2022. From 1 October 2020 to 31 March 2022, he was a temporary professor with TU Berlin, serving as acting head of the Chair of Speech Communication at the Institute of Language and Communication. He has been head of research with audEERING GmbH since September 2018, working in the field of machine audio analyses. Following studies and a doctorate with TU Berlin in speech communication and computer science, he worked on several DFG projects as a research associate. From 2000 to 2018 he worked as a language technology expert with T-Systems and Deutsche Telekom AG. He also serves as an expert reviewer with conferences and editor for the World Wide Web Consortium. He was also a reviewer for EU Horizon 2020 tenders.