A review on five recent and near-future developments in computational processing of emotion in the human voice

Dagmar M. Schuller, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Schuller, Dagmar M., and Björn Schuller. 2021. "A review on five recent and near-future developments in computational processing of emotion in the human voice." *Emotion Review* 13 (1): 44–50. https://doi.org/10.1177/1754073919898526.

emotion review

Emotion Review
Vol. 13, No. 1 (January 2021) 44–50
© The Author(s) 2020
ISSN 1754-0739
DOI: 10.1177/1754073919898526
https://journals.sagepub.com/home/emr

A Review on Five Recent and Near-Future Developments in Computational Processing of Emotion in the Human Voice

Dagmar M. Schuller
audEERING GmbH, Germany

Björn W. Schuller
audEERING GmbH, Germany
GLAM – Group on Language, Audio, and Music, Imperial College London, UK
Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

Abstract

We provide a short review on the recent and near-future developments of computational processing of emotion in the voice, highlighting (a) self-learning of representations moving continuously away from traditional expert-crafted or brute-forced feature representations to end-to-end learning, (b) a movement towards the coupling of analysis and synthesis of emotional voices to foster better mutual understanding, (c) weakly supervised learning at a large scale, (d) transfer learning from related domains such as speech recognition or cross-modal transfer learning, and (e) reinforced learning through interactive applications at a large scale. For each of these trends, we shortly explain their implications and potential use such as for interpretation in psychological studies and usage in digital health and digital psychology applications. We also discuss further potential development.

Keywords
affective computing, emotion, review, speech, voice

Introduction

The idea of masking the verbal content of spoken language but keeping the emotional tone is decades old (Rogers, Schererf, & Rosenthal, 1971). However, as in Steven Spielberg's recent 2018 cinematic film *Ready Player One*, the opposite idea also exists: masking the emotion but keeping the verbal content. There, the protagonist is caught by his evil opponent in a telepresence conversation after receiving a generous offer: "Twenty-five million dollars. I can see you're using emotion-suppressing software right now, and why wouldn't you?" But where are computers really these days on analysing, synthesising, or converting emotions in the voice? Analysing emotions in the voice has a broad application field in industry already—call centre analysis, interview analysis of emotions for recruiting and assessing the right candidate for a position, or tracking one's emotions over the course of the day are just a few examples that are already commonly used. Thereby, the emotion as conveyed in spoken language is recognised by means of acoustic and linguistic analyses. The synthesis of emotional speech by machines bears huge application potential, in particular in human–machine communication. Speech is already broadly synthesised in devices such as Alexa, Cortana, and Siri, but not yet in different emotional flavours. Finally, converting the emotion in the voice by changing acoustic parameters allows, for example, to hide one's genuine emotion, for example, to appear always friendly in a phone call.

Such computational processing of speech either in search of emotion or to synthesise emotion and even convert emotion has matured considerably over its more than two decades of history (B. Schuller, 2018). From a machine learning perspective, different approaches that require more or less human help exist. These include the prevailing approach in the field of supervised learning, where a machine learns only from human-annotated data. Essentially, this means that humans let the machine know by "labelling" which emotion is contained in speech data examples selected for training. In self-learning, the machine itself annotates speech samples with the emotion. This way, the machine learns from having seen more, even if these data are not labelled by humans and could have erroneous labels coming from errors made by the machine. Weakly supervised learning

Author note: We acknowledge funding from the EU's HORIZON 2020 Grants Nos. 754657 (ECoWeB) and 210487208 (WorkingAge).

Corresponding author: Björn W. Schuller, Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Eichleitnerstrasse 30, Augsburg, 86159, Germany.
Email: bjoern.schuller@imperial.ac.uk

can be seen as a mix of these two strategies, where usually the machine is trained first by humans, but then improves itself. To reduce the workload of human annotation, further alternatives exist: transfer learning aims at using knowledge on related tasks available to a machine to exploit. For example, when learning to speak English but already knowing another Germanic language, humans would use this advantage. Machines, too, can profit from knowledge on other speech analysis or computer perception tasks when learning about speech emotion recognition. Reinforced learning has practically not been used in the field yet, but offers to train a machine by interaction with humans and the "world" rather than having humans explicitly annotate emotion in speech samples. These different approaches and others will be outlined in what follows.

Recent improvement on machine analysis, synthesis, and conversion of emotion in the voice is particularly owed to the advances in deep learning,[1] that is, the usage of artificial neural networks with several hidden layers known as "deep neural networks." Manifold, different deep neural network types have been considered in this field, such as convolutional, recurrent, or generative adversarial ones, which can only partially be explained here. Overall, these networks can provide a hierarchical learning of aspects contained in the speech signal. The technique has been embraced early on in this field (Wöllmer et al., 2008).

Deep learning has in particular led to a shift from representing the voice by acoustic features suggested by experts such as phoneticians, to increasingly representing it by acoustic features automatically learnt from exemplary speech samples, for example, by deep neural networks. In fact, such representations are increasingly transferred, including using representations learnt by neural networks for automatic speech recognition for speech emotion recognition (Pratap et al., 2018). Furthermore, the advent of deep neural networks that can also generate speech data, such as generative adversarial networks (GANs; Goodfellow et al., 2014) and deep neural networks trained on large data sets such as WaveNet (van den Oord et al., 2016) allow more and more to synthesise speech not only of much better quality, but to increasingly condition it to speaker states and traits of interest. Ultimately, one can expect future emotionally intelligent AI to improve itself on these tasks, likely first by trying to learn by itself from speech without emotion labels, then increasingly by reinforcement learning when turning its artificial emotional intelligence into an advantage in real-world interactions (D. Schuller & Schuller, 2018).

Here, we aim to quickly introduce these trends and discuss their potential use in getting more insights into human behaviour. In particular, we explain the interpretability of computational processing of emotion in the human voice in the age of deep learning. While we are mostly focussing on the acoustic factor, most principles apply similarly for linguistic analysis of emotion in spoken and even written language.

Self-Learning of Representations

The computational analysis of emotion as manifested in the voice dates back long before the first attempts at automatic speech emotion recognition. While the first patent on emotion recognition from speech dates from 1978 (U.S. Patent No. 4,093,821, 1978), and related attempts at affective speaker state assessment from even earlier (U.S. Patent No. 3,855,416, 1974), computers had already been used for the analysis of emotion before (Williams & Stevens, 1972). In these efforts and the later automatic recognition of emotion in the voice, automatically extracted acoustic features designed by experts were the state-of-the-art technology for the representation of the voice as a basis for decision making on emotional states. This seems, however, suboptimal and unnatural to some degree—suboptimal, as features adapted to the exact problem and learnt by computers from data themselves could be expected to lead to better representations; unnatural, as infants as young as 5 months of age can discriminate emotions in the voice (Walker-Andrews & Lennon, 1991). At this age, however, they can hardly understand spoken language since they have not been told which acoustic features to pay attention to, but rather have learnt to represent the characteristics themselves, likely by a mixture of reinforced learning and transfer learning from context and visual emotion recognition. A first step towards "learning" features from data may be the move towards "bags of audio words." There, acoustic features are quantised into a number of representative examples stored in a codebook. Then, often only histograms of their frequency of occurrence are used for classification of emotion in speech (e.g., the "openXBOW" toolkit; Schmitt & Schuller, 2016). In addition, complete representation learning arrived with the advances in deep learning. In fact, in 2016, the first computational approach was successful to learn "end-to-end" from the raw speech waveform through to the continuous emotion target in terms of arousal and valence values (Trigeorgis et al., 2016). Since then, several related end-to-end learning approaches have appeared (e.g., Chang & Scherer, 2017; Sarma et al., 2018).

In addition, next to toolkits such as "openSMILE"[2] for the extraction of "traditional" feature sets for emotion in speech, such as the compact "GeMAPS" set (Eyben et al., 2016), also toolkits for self-learning of representations such as "auDeep" (Freitag, Amiriparian, Pugachevskiy, Cummins, & Schuller, 2018) and end-to-end learning from the raw speech time signal such as the "End2You" toolkit (Tzirakis, Zafeiriou, & Schuller, 2018) have appeared. Since, related approaches have also been based on spectral representations (Ghosh, Laksana, Morency, & Scherer, 2016). Rather than learning from the raw time signal, these do not learn a representation from the time signal, instead, they learn a representation from a time-frequency transform such as the short-time Fourier transform. This seems natural in a way, relating to human hearing in our inner ear, where a time frequency transformation is the basis for further hearing.

When it comes to decision-making, for example, deciding on "joy" or "anger," or the degree of valence based on the representation, the field has largely moved to modelling also the time history of emotion. An often taken approach to this end are recurrent neural networks, that is, neural networks that take past (and potentially also future) events into account—often considering memory such as by long short-term memory (LSTM; Wöllmer et al., 2008) or the related simpler gated recurrent units (GRUs; Rana, 2016).

More recently, so-called "attention modelling" has been used to improve according architectures. The aim is to learn which parts of the input to pay more attention to in order to best recognise the emotion. According algorithms are used at various positions in deep neural networks to focus attention, for example, on time or frequency. These have repeatedly been shown to work well in speech emotion recognition (Mirsamadi, Barsoum, & Zhang, 2017; Z. Zhang, Wu, & Schuller, 2019; Zhao et al., 2018).

Such approaches are at state-of-the-art level (Trigeorgis et al., 2016; Tzirakis et al., 2018) while omitting the need for any explicit acoustic feature extraction. The often-made reproach when it comes to such deep end-to-end learning is the "black box" nature of these approaches (Shwartz-Ziv & Tishby, 2017). This relates to the fact that while deep learning usually leads to great results, it is hard to explain how and why decisions are made, as up to millions of learning parameters are used to form a usually highly nonlinear decision boundary in high-dimensional decision space. This seems a particularly important aspect to overcome for psychological or medical applications, where explainability and interpretability are crucial. For example, in terms of diagnoses, patients crave for information on which these are based. In fact, selected parts can be explained. For example, visualising the learnt attention patterns or the contextual time profiles is possible. Also, one can compare learnt acoustic features with human-designed features, for example, by correlation. In fact, high correlations were observed (Trigeorgis et al., 2016), for example between learnt features and human-designed prosodic features, such as average fundamental frequency and energy. The correlation was not 1, but rather between .7 and .8. This could have stemmed from too little training data available during learning. Yet, it may be that more optimal representations have been learnt. In fact, adequate psycho-acoustic modelling (Zwicker & Fastl, 2013) can go far beyond usual prosody-related feature calculation as applied in speech emotion recognition today. As an example, pitch is commonly modelled in attempts to estimate the (physical) fundamental frequency in the speech signal. This, however, ignores human pitch perception's interdependence with length, intensity, and spectral composition of the acoustic signal. In other words, it appears possible that end-to-end learning from the raw speech signal lead to better representations in the sense of being more like human perception, hence leading to better speech emotion recognition. However, the visualisation and interpretation of such learnt acoustic features seem crucial to lead to a better understanding of the relation to emotion as carried in speech acoustics. This interpretation is not trivial; luckily, algorithms and tools (Ribeiro, Singh, & Guestrin, 2016) that help with explanation of the learnt features are increasingly available. With these, it will hopefully soon be possible to better interpret why machine learnt representations and models work potentially better than those trained on traditional feature representations.

Coupling of Analysis and Synthesis

One could discuss whether when we learn to recognise and understand speech, we simultaneously learn to produce speech (Lenneberg, 1962; Owren, Amoss, & Rendall, 2011). In fact, it has been shown that speaking and understanding speech share the same regions in the brain, except that the parts controlling the vocal tract movement are not needed for spoken language understanding (Menenti, Gierhan, Segaert, & Hagoort, 2011). Hence, it seems a promising avenue of research to couple learning to recognise and synthesise emotion in speech also in computational processing, to learn about both of these related tasks synergistically. In the case of (speech emotion) recognition, this is the case because the generation of new artificial samples can be beneficiary because, then, a larger amount of data would be made available for training. Until recently, this was hardly possible, as synthesis of emotion in speech has been largely rule-based (Schröder, Burkhardt, & Krstulovic, 2010). However, GANs, as mentioned in the Introduction and shown in Figure 1, recently allowed for such coupling. In a GAN, a "generator" neural network iteratively learns to generate realistic data. A second neural network called "discriminator" iteratively learns to improve on (a) distinguishing the artificially generated data by the generator neural network from real (speech) data, and (b) solving its usual task (speech emotion recognition, in our case). Then, they iteratively improve on both tasks, that is, generation and recognition of emotion in an iterative process, each neural network learning from the other. This may be a rather simple architecture but indeed it is possible to generate speech audio with GANs (Donahue, McAuley, & Puckette, 2018). GANs that allow to generate a conditioned target output, such as angry or happy speech, can then allow to generate emotionally targeted speech.

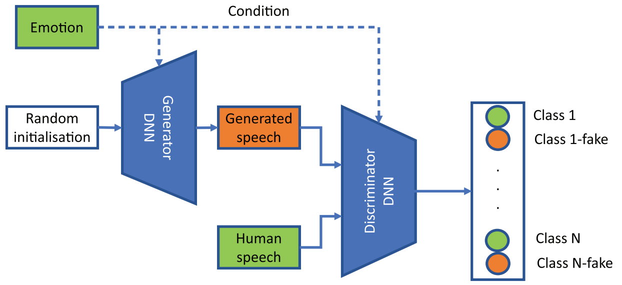

Figure 1. Schema of a generative adversarial network (GAN) with emotion as "condition." A generator deep neural network (DNN) learns to generate artificial speech; a discriminator DNN learns to discriminate artificially generated from human-generated speech while simultaneously learning how to solve its emotion recognition task such that both DNNs improve on their task.

At the same time, it has been shown that GANs can be used to improve machines' performance at emotion recognition tasks (Han, Zhang, Cummins, & Schuller, 2018; Han, Zhang, Ren, Ringeval, & Schuller, 2018). The generative part thereby generates novel learning examples to enrich the training material. While this has so far been executed on the level of the feature vector for speech emotion recognition, that is, the generator produces novel acoustic feature values rather than audio, it basically leaves coupling generation of emotional speech audio for future research as a next step to truly couple analysis and synthesis in this field. A couple of further issues to solve will include improvements on the so-called "mode collapse" problem of GANs (the production of novel examples of data that are too similar to previous instances; Srivastava, Valkov, Russell, Gutmann, & Sutton, 2017). Independent of that, a couple of further architectures such as variational auto-encoders (VAEs) are currently available and able to generate novel data that could be considered for related architectures of coupled analysis and synthesis. As end-to-end learning has been successfully realised also for synthesis of emotional speech (Gao, Chakraborty, Tembine, & Olaleye, 2018), one could aim at coupling end-to-end analysis and synthesis.

Furthermore, such coupled analysis and synthesis may help to improve future emotional voice conversion. The identity of the speaker and the linguistic content could be preserved during conversion (Gao et al., 2018). Yet, the emotion would be changed such as in the introductory example of emotion-suppressing software. Apart from human audible voice conversion, one can also think of addition of human nonaudible noise to a speech signal to fool automatic recognition systems. So far, such "adversarial attacks" have been successfully applied to fool AI unnoticeable to humans mostly in the image domain. This was reached by adding image noise hard to see by humans but triggering a desired recognition outcome from a deep neural network. A potential use of such deception in the speech emotion domain could be to trick an AI to believe one is being friendly despite being rude. This could be of interest, for example, if the AI would be monitoring one's behaviour during an AI-based job interview. Adversarial attacks have also been accomplished to fool automatic speech recognition (Alzantot, Balaji, & Srivastava, 2018), and could soon alter how we appear emotionally to an AI.

Weakly Supervised Learning

Insufficient data with information of the contained emotion is usually the major bottleneck for computational processing of emotion in speech and nonverbal vocalisations. Hence, approaches that exploit also speech data without a known emotion are particularly important in reaching emotion recognition by machines largely independent of the individual speaking, her cultural background, or the language spoken (Sauter, Eisner, Ekman, & Scott, 2015). As speech data are available in abundance on social media, in films, etc., according approaches are in principle able to exploit endless data resources. Most importantly, methods exploiting such speech data without information of the emotion usually need to provide a solid estimation of (a) the algorithm's own confidence in its judgement of the emotion and (b) how informative the current sample is. Based on this, they are able to decide whether the algorithm can add the new sample to the training material based on its own assessment ("semisupervised learning"; Huang et al., 2018), or, if not sufficiently confident in its own assessment, whether it would be important enough to ask for a human labeller, if available ("active learning"). In combination, such "cooperative learning" (as offered, for example, by the "iHEARu-PLAY" platform; Hantke, Eyben, Appel, & Schuller, 2015) is known to bear the potential to drastically reduce the amount of manually labelled data (Y. Zhang et al., 2019). In addition, weakly supervised learning can also exploit other strong modalities. An example is training an automatic speech emotion recogniser with the help of an automatic video emotion recogniser. This makes sense, for example, for valence, where video analysis is known to be more reliable than speech acoustics analysis (Albanie, Nagrani, Vedaldi, & Zisserman, 2018; Han, Zhang, Ren, & Schuller, 2019). As it requires excessive computational power to process big data in such manner, one can expect the potential of such weakly supervised learning to be leveraged soon.

Transfer Learning

Humans are great at transfer learning, that is, using knowledge gained in one domain when learning in another one. For example, when it comes to language learning (Durgunoglu, Nagy, & Hancin- Bhatt, 1993): if speaking English, we exploit this knowledge easily to learn new related languages such as those in the Germanic language family. While transfer learning is gaining momentum in machine learning these days (Pan & Yang, 2010), such knowledge transfer has not been seen much in the field of computational emotion processing in speech until recently.

The idea to use networks trained on other tasks in the recognition of emotion in speech kicked off using deep neural networks that, surprisingly, had been pretrained to do image classification (Cummins et al., 2017) based on speech spectrograms as images (see Figure 2). Cummins et al. (2017) showed an image-recognition pretrained network to be on a par or better than expert-designed "traditional" speech features, such as mean fundamental frequency, and alike for representation of recognition of emotion in speech in a recognition task. Publicly shared automatic speech recognition pretrained networks have started to be used for emotion recognition in speech with great success (Adigwe, Tits, Haddad, Ostadabbas, & Dutoit, 2018). In addition, whole toolkits such as the "DeepSpectrum" toolkit have recently been devoted to this task, where one can choose amongst trained deep neural networks to use for representation of speech. Following this trend of shared pretrained networks in computer vision and speech recognition, one can expect "emotion-nets" pretrained for the specific task of emotion processing in speech or other modalities to appear soon. Interpretation of why such transfer learning works can be difficult, but visualisation of the neural network activations can give some insight. For example, vision deep neural networks often learn to recognise edges or certain shapes, which may also appear in audio spectra. It remains to be evaluated how well such transfer learning will generalise across languages and levels of specificity (e.g., positive valence vs. happy).

Beyond such transfer learning on the level of acoustic features, further forms of transfer learning exist, such as transfer learning for decision-making models (e.g., Abdelwahab & Busso, 2015; Mao, Xue, Rao, Zhang, & Zhan, 2016; Song, Jin, Zhao, & Xin, 2014). Interpretation of successful transfer of knowledge can be challenging in these cases as well, in particular if the tasks are not obviously related to each other.

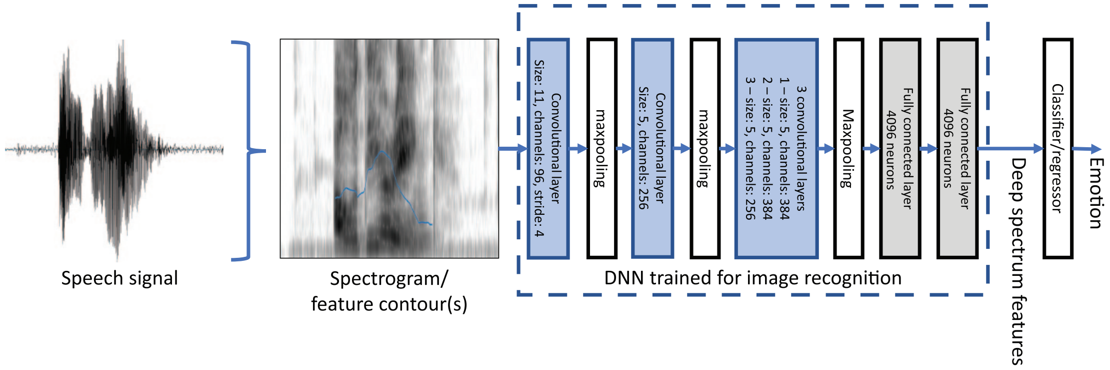

Figure 2. Transfer learning using DNNs trained on images for speech emotion recognition by representing a speech signal as image (e.g., by spectral transformation and/or inclusion of other feature contours).

Reinforced Learning

As humans, it appears natural that we learn about emotions by interacting with the real world. For example, we learn to interpret vocal patterns conveying emotion or facial expressions of others without being given overly repeated explanations of these patterns to learn in a supervised manner to generalise to future unseen patterns. In current computational speech emotion recognition, however, supervised learning is still the prevailing reality, apart from the sparse attempts at weakly supervised learning mentioned before. But in a world where AI empowered with artificial emotional intelligence will become an ever-present reality, it seems obvious that also computers will start learning to recognise our emotions to best solve their tasks while in interaction with us (Broekens, 2007). A positive emotion of a user of AI could then serve as a positive reward, and vice versa for a user's negative emotion for an AI relying on reinforcement learning. First attempts in this direction exist (Motamed, Setayeshi, & Rabiee, 2017), but more is expected to come.

Beyond analysis of emotional speech, one can also easily imagine reinforcement learning to play a role in future emotional speech synthesis; for example, a speech interface may learn how to shape its acoustic parameters to better simulate empathy to be perceived as more sympathetic by the user. Integrated into mass consumer products, such an AI could learn reinforced from interactions with thousands or millions of users, just like "Samantha" in the 2013 film *Her*. It will be interesting to analyse the vocal behaviour such a system will learn to use to best accomplish its goals in its interaction with us.

Conclusion

We discussed five major trends currently reshaping the computational processing of emotion in the human voice, or that can be expected to soon do so. These were, in short, (a) self-learning of representations from or to raw speech signals in analysis and synthesis of emotional speech, (b) the coupling of analysis and synthesis into a topology that learns about both ends and hence gains emergent knowledge beyond separated handling of these tasks, (c) weakly supervised learning at a large scale, (d) the transfer of knowledge from related tasks, domains, and modalities, and (e) reinforcement learning of analysis and synthesis of emotion in the voice. From these trends, one can assume that such systems will soon be able to reach human parity level not only for arousal recognition. How accurate they become, presumably depends on the granularity at which they are tested. However, as self-learning and reinforcement learning at a large scale can be expected to soon shape the landscape of computational engines in this context, machines could indeed do without learning from human-annotated data, which is usually highly ambiguous in this field. Rather, machines could develop an improved beyond-human understanding of emotion, potentially also profiting from closer to ground-truth emotion labels that could come from increasingly better brain–computer interfaces. This would potentially lead to superhuman speech emotion recognition performance. At the same time, recent progress in speech separation from noise and other speakers (Keren, Han, & Schuller, 2018) will help reach superhuman abilities also when it comes to processing of speech in adverse "in-the-wild" conditions. As a side stream, computational processing may also broaden out beyond human vocalisations and lead to analysis and synthesis of animal emotion (e.g., Hantke, Cummins, & Schuller, 2018). In short, we likely look into a near future where machines can recognise, synthesise, and convert emotion in human and animal voices beyond our skills. The challenge then will be to interpret and explain what the machine has learnt. It will remain to be seen how well such AI will recognise regulated or insincere emotion. The current belief that women are more emotionally expressive than men (Fischer & LaFrance, 2015), and computers are not emotionally expressive, could hence soon change—at least when it comes to computers.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: both authors are chief executives of audEERING, an audio intelligence company providing industrial emotional speech processing solutions.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Björn W. Schuller https://orcid.org/0000-0002-6478-8699

Notes

[1] Resources to learn about backgrounds of deep learning or how to implement solutions can be found, for example at websites https://pytorch.org/ and https://www.tensorflow.org/. An example of a tutorial on machine learning can be found at https://developers.google.com/machine-learning/crash-course/

[2] For a tutorial allowing to implement a speech emotion recogniser, see: https://www.audeering.com/download/opensmile-book-latest/

References

Abdelwahab, M., & Busso, C. (2015). Supervised domain adaptation for emotion recognition from speech. In *Proceedings of the 40th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP* (pp. 5058–5062). New York, NY: IEEE.

Adigwe, A., Tits, N., Haddad, K. E., Ostadabbas, S., & Dutoit, T. (2018). *The emotional voices database: Towards controlling the emotion dimension in voice generation systems*. Retrieved from https://arxiv.org/abs/1806.09514

Albanie, S., Nagrani, A., Vedaldi, A., & Zisserman, A. (2018). *Emotion recognition in speech using cross-modal transfer in the wild*. Retrieved from https://arxiv.org/abs/1808.05561

Alzantot, M., Balaji, B., & Srivastava, M. (2018). *Did you hear that? Adversarial examples against automatic speech recognition*. Retrieved from https://arxiv.org/abs/1801.00554

Broekens, J. (2007). Emotion and reinforcement: Affective facial expressions facilitate robot learning. In T. S. Huang, A. Nijholt, M. Pantic, & A. Pentland (Eds.), *Artificial intelligence for human computing* (pp. 113–132). Berlin, Germany: Springer.

Chang, J., & Scherer, S. (2017). Learning representations of emotional speech with deep convolutional generative adversarial networks. In *Proceedings of the 42nd IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP* (pp. 2746–2750). New York, NY: IEEE.

Cummins, N., Amiriparian, S., Hagerer, G., Batliner, A., Steidl, S., & Schuller, B. (2017). An image-based deep spectrum feature representation for the recognition of emotional speech. In *Proceedings of the 25th ACM International Conference on Multimedia, MM 2017* (pp. 478–484). New York, NY: ACM.

Donahue, C., McAuley, J., & Puckette, M. (2018). *Synthesizing audio with generative adversarial networks*. Retrieved from https://arxiv.org/abs/1802.04208

Durgunoglu, A. Y., Nagy, W. E., & Hancin-Bhatt, B. J. (1993). Cross-language transfer of phonological awareness. *Journal of Educational Psychology*, *85*(3), 453–465.

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., . . . Truong, K. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, *7*(2), 190–202.

Fischer, A., & LaFrance, M. (2015). What drives the smile and the tear: Why women are more emotionally expressive than men. *Emotion Review*, *7*, 22–29.

Freitag, M., Amiriparian, S., Pugachevskiy, S., Cummins, N., & Schuller, B. (2018). auDeep: Unsupervised learning of representations from audio with deep recurrent neural networks. *Journal of Machine Learning Research*, *18*, 1–5.

Fuller, F. (1974). *U.S. Patent No. 3,855,416*. Washington, DC: Patent and Trademark Office.

Gao, J., Chakraborty, D., Tembine, H., & Olaleye, O. (2018). *Nonparallel emotional speech conversion*. Retrieved from https://arxiv.org/abs/1811.01174

Ghosh, S., Laksana, E., Morency, L.-P., & Scherer, S. (2016). Representation learning for speech emotion recognition. In *Proceedings Interspeech* (pp. 3603–3607). http://dx.doi.org/10.21437/Interspeech.2016-692

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). *Generative adversarial nets*. Retrieved from https://arxiv.org/pdf/1406.2661.pdf

Han, J., Zhang, Z., Cummins, N., & Schuller, B. (2018). Adversarial training in affective computing and sentiment analysis: Recent advances and prospectives. *IEEE Computational Intelligence Magazine*, *14*(2), 68–81.

Han, J., Zhang, Z., Ren, Z., Ringeval, F., & Schuller, B. (2018). Towards conditional adversarial training for predicting emotions from speech. In *Proceedings of the 43rd IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP* (pp. 6822–6826). New York, NY: IEEE.

Han, J., Zhang, Z., Ren, Z., & Schuller, B. (2019). Implicit fusion by joint audiovisual training for emotion recognition in mono modality. In *Proceedings of the 44th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP* (pp. 5861–5865). New York, NY: IEEE.

Hantke, S., Cummins, N., & Schuller, B. (2018). What is my dog trying to tell me? The automatic recognition of the context and perceived emotion of dog barks. In *Proceedings of the 43rd IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP* (pp. 5134–5138). New York, NY: IEEE.

Hantke, S., Eyben, F., Appel, T., & Schuller, B. (2015). iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing. In *Proceedings of the 1st International Workshop on Automatic Sentiment Analysis in the Wild (WASA 2015) held in conjunction with the 6th biannual Conference on Affective Computing and Intelligent Interaction (ACII 2015)* (pp. 891–897). New York, NY: IEEE.

Huang, J., Li, Y., Tao, J., Lian, Z., Niu, M., & Yi, J. (2018). Speech emotion recognition using semi-supervised learning with ladder networks. In *Proceedings of the First Asian Conference on Affective Computing and Intelligent Interaction, ACII Asia* (pp. 1–5). New York, NY: IEEE.

Keren, G., Han, J., & Schuller, B. (2018). *Scaling speech enhancement in unseen environments with noise embeddings*. Retrieved from https://arxiv.org/abs/1810.12757

Lenneberg, E. H. (1962). Understanding language without ability to speak: A case report. *The Journal of Abnormal and Social Psychology*, *65*(6), 419–425.

Mao, Q., Xue, W., Rao, Q., Zhang, F., & Zhan, Y. (2016). Domain adaptation for speech emotion recognition by sharing priors between related source and target classes. In *Proceedings of the 41st IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP* (pp. 2608–2612). New York, NY: IEEE.

Menenti, L., Gierhan, S. M., Segaert, K., & Hagoort, P. (2011). Shared language: Overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. *Psychological Science*, *22*(9), 1173–1182.

Mirsamadi, S., Barsoum, E., & Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *Proceedings of the 42nd IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP* (pp. 2227–2231). New York, NY: IEEE.

Motamed, S., Setayeshi, S., & Rabiee, A. (2017). Speech emotion recognition based on a modified brain emotional learning model. *Biologically Inspired Cognitive Architectures*, *19*, 32–38.

Owren, M. J., Amoss, R. T., & Rendall, D. (2011). Two organizing principles of vocal production: Implications for nonhuman and human primates. *American Journal of Primatology*, *73*(6), 530–544.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.

Pratap, V., Hannun, A., Xu, Q., Cai, J., Kahn, J., Synnaeve, G., . . .Collobert, R. (2018). *wav2letter++: The fastest open-source speech recognition system*. Retrieved from https://arxiv.org/abs/1812.07625

Rana, R. (2016). *Gated recurrent unit (GRU) for emotion classification from noisy speech*. Retrieved from https://arxiv.org/abs/1612.07778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). New York, NY: ACM.

Rogers, P. L., Schererf, K. R., & Rosenthal, R. (1971). Content filtering human speech: A simple electronic system. *Behavior Research Methods & Instrumentation*, *3*(1), 16–18.

Sarma, M., Ghahremani, P., Povey, D., Goel, N. K., Sarma, K. K., & Dehak, N. (2018). Emotion identification from raw speech signals using DNNs. In *Proceedings Interspeech* (pp. 3097–3101). Retrieved from https://danielpovey.com/files/2018_interspeech_emotion_id.pdf

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2015). Emotional vocalizations are recognized across cultures regardless of the valence of distractors. *Psychological Science*, *26*(3), 354–356.

Schmitt, M., & Schuller, B. (2016). *openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit*. Retrieved from https://arxiv.org/abs/1605.06778

Schröder, M., Burkhardt, F., & Krstulovic, S. (2010). Synthesis of emotional speech. In K. R. Scherer, T. Bänziger, & E. B. Roesch (Eds.), *Blueprint for affective computing* (pp. 222–231). Oxford, UK: Oxford University Press.

Schuller, B. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, *61*(5), 90–99.

Schuller, D., & Schuller, B. (2018). The age of artificial emotional intelligence. *IEEE Computer Magazine*, *51*(9), 38–46.

Shwartz-Ziv, R., & Tishby, N. (2017). *Opening the black box of deep neural networks via information*. Retrieved from https://arxiv.org/abs/1703.00810

Song, P., Jin, Y., Zhao, L., & Xin, M. (2014). Speech emotion recognition using transfer learning. *IEICE Transactions on Information and Systems*, *97*(9), 2530–2532.

Srivastava, A., Valkov, L., Russell, C., Gutmann, M. U., & Sutton, C. (2017). *VEEGAN: Reducing mode collapse in GANs using implicit variational learning*. Retrieved from https://arxiv.org/abs/1705.07761

Trigeorgis, G., Ringeval, F., Brückner, R., Marchi, E., Nicolaou, M., Schuller, B., & Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of the 41st IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP* (pp. 5200–5204). New York, NY: IEEE.

Tzirakis, P., Zafeiriou, S., & Schuller, B. (2018). *End2You – The Imperial Toolkit for Multimodal Profiling by End-to-End Learning*. Retrieved from https://arxiv.org/abs/1802.01115

Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., . . . Kavukcuoglu, K. (2016). *WaveNet: A generative model for raw audio*. Retrieved from https://arxiv.org/abs/1609.03499

Walker-Andrews, A. S., & Lennon, E. (1991). Infants' discrimination of vocal expressions: Contributions of auditory and visual information. *Infant Behavior and Development*, *14*(2), 131–142.

Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, *52*(4B), 1238–1250.

Williamson, J. D. (1978). *U.S. Patent No. 4,093,821*. Washington, DC: Patent and Trademark Office.

Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., & Cowie, R. (2008). Abandoning emotion classes – Towards continuous emotion recognition with modelling of long-range dependencies. In *Proceedings Interspeech* (pp. 597–600). Retrieved from https://pdfs.semanticscholar.org/2096/6cbf8d1813e585131d63aa27b01c622df04e.pdf

Zhang, Y., Weninger, F., Michi, A., Wagner, J., André, E., & Schuller, B. (2019). A generic human–machine annotation framework using dynamic cooperative learning with a deep learning-based confidence measure. *IEEE Transactions on Cybernetics*. Advance online publication. http://doi.org/10.1109/TCYB.2019.2901499

Zhang, Z., Wu, B., & Schuller, B. (2019). Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In *Proceedings of the 44th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP* (pp. 6705–6709). New York, NY: IEEE.

Zhao, Z., Zheng, Y., Zhang, Z., Wang, H., Zhao, Y., & Li, C. (2018). Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition. In *Proceedings Interspeech* (pp. 272–276). http://doi.org/10.21437/Interspeech.2018-1477

Zwicker, E., & Fastl, H. (2013). *Psychoacoustics: Facts and models*. Berlin, Germany: Springer Science & Business Media.