Deep speaker conditioning for speech emotion recognition

Andreas Triantafyllopoulos, Shuo Liu, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Triantafyllopoulos, Andreas, Shuo Liu, and Björn W. Schuller. 2021. "Deep speaker conditioning for speech emotion recognition." In *IEEE International Conference on Multimedia and Expo (ICME 2021), Shenzhen, China, 5-9 July 2021*, 1–6. New York, NY: IEEE. https://doi.org/10.1109/icme51207.2021.9428217.

DEEP SPEAKER CONDITIONING FOR SPEECH EMOTION RECOGNITION

Andreas Triantafyllopoulos[*†], *Shuo Liu*[†] , *Björn W. Schuller*[*†‡]

[*]audEERING GmbH, Gilching, Germany
[†]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[‡]GLAM – Group on Language, Audio & Music, Imperial College London, UK
atriant@audeering.com

ABSTRACT

In this work, we explore the use of speaker conditioning sub-networks for speaker adaptation in a deep neural network (DNN) based speech emotion recognition (SER) system. We use a ResNet architecture trained on log spectrogram features, and augment this architecture with an auxiliary network providing speaker embeddings, which conditions multiple layers of the primary classification network on a single neutral speech sample of the target speaker. The whole system is trained end-to-end using a standard cross-entropy loss for utterance-level SER. Relative to the same architecture without the auxiliary embedding sub-network, we are able to improve by 8.3% on IEMOCAP, and by 5.0% and 30.9% on the 2-class and 5-class SER tasks on FAU-AIBO, respectively.

Index Terms— speech emotion recognition, affective computing

1. INTRODUCTION

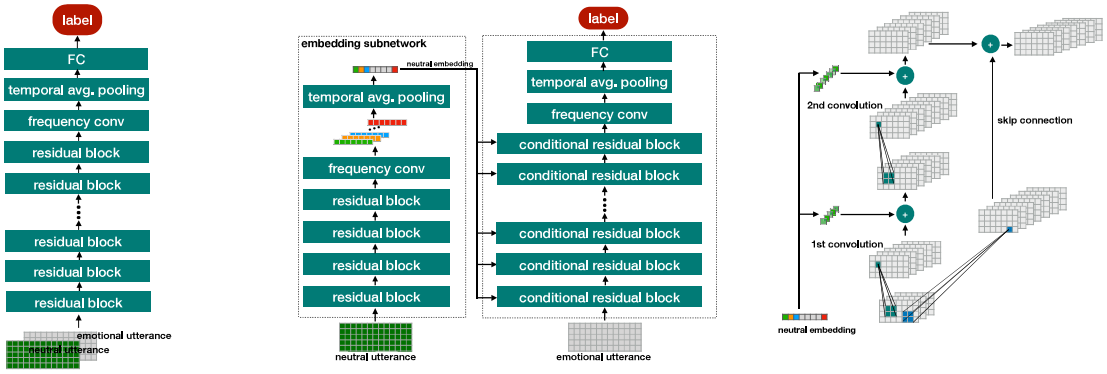
Speech emotion recognition (SER) architectures are highly dependent on speaker-specific characteristics [1, 2, 3]. That is a natural byproduct of the emotional expression process, which is highly individualised [4, 5]. Different people both *experience* and *express* emotions in different ways, which causes differences in the bio-signals they emit during an emotional episode. For the audio modality in particular, the acoustic properties of the speaker's voice (e. g., their fundamental frequency) additionally influence the signal that is captured by an SER system – an effect which further complicates the analysis process. Therefore, when an automatic framework is utilised to identify a person's emotional state, it needs to take these individual differences into account, even more so when it relies on the audio modality.

This has led the SER community to adopt speaker-independent test sets in order to measure model performance in an unbiased way, as seen, for example, in the Interspeech Computational Paralinguistics Challenge (ComParE) series [6]. On the other hand, different modelling approaches have been proposed for taking speaker dependence into account; either by incorporating it directly into the algorithm to improve performance, or by trying to remove it to boost generalisation.

One approach is to try and remove speaker-dependent characteristics from the input features. Recently, Tu *et al.* [7] used the well-known deep adversarial domain adaptation framework [8], jointly training a network to predict emotion but mis-classify speaker. Li *et al.* [9] extend that framework by disentangling the speaker classification from the emotion recognition sub-networks into two separate steps, and then training the latter to cause more mis-classifications in the former, thus removing speaker information.

More relevant to this work, a typical methodology is to apply some form of speaker-specific normalisation. Vlasenko *et al.* [2] utilise speaker normalisation in the feature space and report a 4 % accuracy increase. They normalised using the mean and standard deviation of all samples for each speaker. Sethu *et al.* [10] apply feature warping (i. e., histogram normalisation) to map all features of every speaker to the normal distribution, taking into account both emotions and all samples present in their data set of choice. Schuller *et al.* [11] perform speaker normalisation in the feature space in the context of cross-corpus SER, and compare it to other normalisation strategies. They conclude that speaker normalisation yields the best performance in that context. However, in order to successfully normalise, they require the entire emotional spectrum to be taken into account. Busso *et al.* [12] overcome the need for prerecorded samples of the target speaker by proposing an iterative feature normalisation technique that starts by automatically detecting neutral samples at test time, and adjusts the normalisation parameters based on those. They show improved results in a *neutral* vs *emotional* binary classification problem.

Even though speaker normalisation approaches show improved performance, a major limitation is that they usually require access to the full spectrum of the target speaker's emotions. However, this data is hard to collect in practice, as it would require access to the speaker's true emotional state, unknown at test time. A typical way to mitigate this is to make use of a speaker enrolment phase, which seeks to collect several samples during the system's deployment phase.



(a) *TwoChannel* architecture (b) eResNet architecture (c) A single conditional residual block

Fig. 1: Diagrams of the architectures used in this work. Figure 1a shows the *TwoChannel* architecture. The same architecture was used for the baseline ResNet, but with a single channel input. Figure 1b shows the main architecture on the right, and the embedding sub-network on the left. Finally, Figure 1c shows the inner workings of each single conditional residual block.

Unfortunately, such approaches are highly impractical. In their simplest form, they rely on the user acting their full range of emotion during a preliminary enrolment phase, thus suffering from the discrepancy between acted and naturalistic emotions [13, 14]. Alternatively, they collect and automatically label samples using a universal SER model during the system's initial deployment, and later use that data to compute the normalisation parameters [12]. However, there is no guarantee that the user will maintain the interaction long enough, as e.g. in interactive voice response (IVR) systems [15], nor that the mis-classifications of the universal model will not cause a mis-calibration of the normalisation parameters.

This means that although utilising speaker-specific information can be highly beneficial for SER systems, most approaches that try to incorporate that information end up being impractical for real-life use cases. On the other hand, although methods that focus on removing speaker-specific information are more conducive to practical applications, they fail to capitalise on that information, thus missing out on the potential benefits that come with personalisation.

In contrast, we introduce a new deep neural network (DNN) framework that allows us to adapt to each new speaker using a single neutral utterance, which enables us to take advantage of the benefits associated with speaker adaptation without requiring a lot of data from the target speaker. This makes our approach suitable even for relatively short interactions. Our main contribution lies in introducing a secondary sub-network that processes the neutral sample to generate an embedding, which is then used to condition the intermediate layers of the main classification network. This approach is motivated by recent advances in the use of embeddings for audio tasks [16, 17], but also by the successful use of conditioning embeddings in the image domain, most notably in the recent case of StyleGAN [18]. Our goal is to investigate whether conditioning the intermediate layers of a DNN on speaker information can help it model the target task, in this case SER, more efficiently. In order to judge whether this embedding sub-network is necessary, we also compare it against a straightforward extension of the baseline architecture, which accepts the neutral sample as an additional input. Our results show that the use of the embedding sub-network is necessary for the success of the approach.

The remaining of this paper is organised as follows. In Section 2, we introduce our proposed framework. The data sets we used are described in Section 3. Experimental results are presented in Section 4.

2. ARCHITECTURE

Our architecture, which we will be referring to as *eResNet*, consists of a main residual network (ResNet) [19] that performs the SER task, and an auxiliary sub-network which conditions each residual layer of the main network on the neutral sample of the target speaker. This architecture has been initially proposed for monaural speech denoising [16, 20], where the auxiliary network was trained to condition on recordings of the target noise environment. Liu *et al.* [17] used this architecture for monaural speech separation by conditioning on a sample from either on or both of the target speakers. In both these cases, the network was trained to output an additive denoising (or separation) mask, and its input was fixed-length sequences. In the present contribution, we adapt it for a classification task, and extend it to work for variable length sequences by introducing a temporal average pooling layer.

A detailed diagram is shown in Figure 1b. The main architecture consists of 8 conditional residual blocks, each block consisting of two convolutional layers, batch normalisation [21], and rectified linear unit (ReLU) activations. The structure of a conditional residual block is shown in Figure 1c. The residual blocks are followed by a convolution layer applied only on the frequency axis, and a temporal average pooling layer applied on the time axis. This temporal pooling layer allows us to handle input sequences of variable length. Finally, the network's output consists of two linear layers, intertwined with a ReLU activation.

The embedding sub-network is made up of 4 standard

residual blocks [19], followed by a convolution layer on the frequency axis, a temporal average pooling layer on the time axis, and a single linear layer that maps the output to the appropriate dimensionality. This sub-network is fed a neutral sample of the target speaker and outputs an embedding vector of fixed dimensionality. This embedding vector is then fed into each residual block of the main architecture, which then maps it to the appropriate dimension and adds it to the output of both convolutional layers.

We apply this architecture to the task of SER for two widely used data sets described in Section 3, and show that using an auxiliary embedding sub-network to condition on the target speaker is beneficial to the downstream task. However, it is not straightforward why that happens. A particular concern is that the auxiliary network is superfluous, and that the baseline network was capable of doing that adaptation on its own, given the appropriate input. To test that hypothesis, we train a variant of the baseline ResNet architecture with the input layer modified to accept two channels. The first channel is used for the sample to be classified, and the second for the neutral sample of the target speaker. With this experiment, we attempt to investigate whether the auxiliary sub-network brings any benefits at all. This architecture, which, with the exception of the number of input channels in the first layer, is identical to the baseline network, is shown in Figure 1a. We will henceforth refer to this architecture as *TwoChannel*.

The input to the networks were log-spectrograms. To compute the short-time Fourier transform (STFT), we used a window size of 0.025 s, and a hop size of 0.010 s. In addition, we peak-normalised the amplitude of the input audio per segment. In our preliminary experiments, we tried out Mel spectrograms but got no performance improvement and therefore report no results for them.

As mentioned, our architecture is capable of handling sequences of variable length. However, during training time, it is beneficial to train on mini-batches to decrease the training time and obtain better convergence. A downside of this approach is that all the samples in the batch must have the same dimensionality. We chose to crop all training samples to a fixed length during training time, and use the entire sequences at test time. This methodology is quite common in the speaker identification literature [22, 23]. FAU Aibo Emotion Corpus (FAU-AIBO) contains a lot of short segments, so we used a 2 s window which is the maximum segment length in the data set. interactive emotional dyadic motion capture database (IEMOCAP) contains longer segments, and so we used the mean segment length which is 4.5 s.

Finally, all networks were trained using stochastic gradient descent (SGD) for a total of 30 epochs with a learning rate of 0.001 and a weight decay of 0.005. We tested three different batch sizes: $8, 16, 32$, and report the best result achieved for each architecture. We used early stopping based on the model's performance on a validation set for each data set as described in Section 4.

Table 1: Data set specifications.

	IEMOCAP		**FAU-AIBO**			
Speakers	10 (5f/5m)		51 (30f/21m)			
Duration	06:22:43		08:50:49			
Samples	**Neutral**	1,747	**A**	1,492	**IDL**	5,823
	Anger	658	**E**	3,601	**NEG**	12,393
	Sadness	1,782	**N**	10,967		
	Happiness	1,074	**P**	889		
			R	1,267		

3. DATA SETS

As outlined, we test our approach in two widely used SER data sets: IEMOCAP [24], and FAU-AIBO [25].

IEMOCAP consists of scripted and improvised dyadic conversations between a total of 10 actors (5 male - 5 female). The data was recorded in 5 sessions, with each session consisting of conversations between a single actor pair (1 male - 1 female). It consists of approximately 12 hours of total data[1]. The corpus contains video, audio, and motion capture data, though in the present work we only make use of the audio modality. Similar to other previous work [26, 27, 28, 29], we only focus on the four basic emotions of *anger*, *happiness*, *neutral*, and *sadness*. We did not fuse *excitement* into *happiness*.

FAU-AIBO was used in the first ComParE challenge. It contains German children emotional speech recorded in a Wizard-of-Oz scenario. The children were giving instructions to a robot which performed predetermined actions irrespective of the instructions given.

The data set has been initially annotated for 11 classes, but we use the formulations defined for the first InterSpeech ComParE challenge, namely, first, a five class problem, by defining the following categories: *angry (A), emphatic (E), neutral (N), positive (P), rest (R)*, and, second, a binary classification problem by adopting the following partitioning: *negative (NEG), idle (IDL)*.

As shown in Table 1, both data sets exhibit a class imbalance, which is more pronounced for FAU-AIBO. Previous work [30] has shown that simply balancing the training set by random sub-sampling can lead to increased performance for a number of different algorithms. We avoid sub-sampling our data set since DNN typically need a lot of data to train, and use a weighted non-negative likelihood loss, weighting each sample by the inverse of the frequency of its class in the loss. The loss function then becomes:

$$l = \sum_{n=1}^{N} \sum_{k=1}^{K} -\frac{1}{w_k} \mathbb{1}(y_n == k) log p_k^n, \quad (1)$$

[1] In Table 1 we report the duration of the data we used for our experiments.

Table 2: Unweighted average recall (UAR)% results for FAU-AIBO using the official train/test splits and emotion classes of the IS2009 ComParE challenge.

Method	2-class	5-class
SVM	66.8	26.3
SVM-NORM	**71.2**	37.5
ResNet	64.1	31.5
ResNet-NORM	67.7	35.1
TwoChannel	63.5	32.4
eResNet	67.3	**41.3**

where K is the number of classes, N the number of samples in a mini-batch, y_i the one-hot representation of the labels, and $\mathbb{1}$ being the indicator function p_k^i the network outputs for the i^{th} element in the batch corresponding to the k^{th} class, and w_k is defined as:

$$w_k = \frac{N}{\sum_{n=1}^{N} \mathbb{1}(y_i == k)}, \tag{2}$$

and computed once over the entire training set.

4. EXPERIMENTS

For both data sets, we perform the following set of experiments:

1. baseline ResNet architecture without the embedding sub-network
2. ResNet-NORM which is the baseline architecture but trained and evaluated on speaker-normalised features
3. *TwoChannel*. This is the variant of the model modified to accept a two-channel input, with the second input being the neutral sample for the target speaker.
4. our eResNet architecture

In addition, we report results with a standard baseline, namely support vector machines (SVMs) [31] trained on utterance-level acoustic features. As features, we use the extended Geneva minimalistic acoustic predictors (eGeMAPS) [32] computed with openSMILE [33]. The features are normalised either on the entire training set (referred to as SVM in the result tables), or independently for each speaker (referred to as SVM-NORM).

For IEMOCAP, we report leave-one-speaker-out (LOSO) cross-validation (CV) results for all our experiments. When testing on one speaker, we use the other speaker of their session as our validation set for early stopping. We report UAR results in Table 3. Results have been averaged over the 10 folds. We also report results by recent state-of-the-art methods that focused on learning speaker invariant representations [7, 9]. However, we caution that the results reported in those works are not directly comparable to ours. Specifically, Tu *et al.* [7] used different test conditions (5-fold CV) than we did. Li *et al.* [9] used the same folds, but merged the *excitement* class into *happy*, and furthermore extended the training set through data augmentation. Therefore, when comparing to these works we primarily focus on relative improvements.

For our experiments on FAU-AIBO, we use official partitions of the InterSpeech 2009 ComParE challenge [6]. Specifically, we use one school for training (*Ohm*), and one for testing (*Mont*) following the protocol of the challenge. In all our experiments, we used one female and one male subject from our training set as validation. We picked subjects with the following ids: *31, 32*. We report UAR results for both the binary and the multi-class classification problems in Table 2. During training, we use random neutral samples from each target speaker. During testing, we always use the first one for that particular speaker in our dataset for our preliminary experiments, and perform an ablation study on the choice of neutral sample in Table 4.

4.1. Results

Overall, the eResNet architecture performs consistently better than both the baseline and the *TwoChannel* network. Specifically, we obtain a 8.3% relative improvement for IEMOCAP, which compares favourably to 7.4% and 4.3% reported by Tu *et al.* [7] and Li *et al.* [9], respectively. For FAU-AIBO, our gains are 5.0% and 30.9% for the 2-class and 5-class problems.

Results in Table 2 show that speaker normalisation is beneficial for both the SVM baseline and our ResNet architecture, a finding consistent with previous literature results. The eResNet architecture yields marginally worse performance for the 2-class case over its baseline, and a big improvement for the 5-class case compared to normalising the features of each speaker over the entire set. In comparison, the *TwoChannel* architecture does poorly with respect to both other normalisation approaches, and performs almost on-par with the baseline.

On the IEMOCAP data set, we observe a counter-intuitive drop in performance when doing speaker normalisation over the features of the entire set for the ResNet architecture, and small gains for the SVM baseline. We were not able to find a satisfactory explanation for this phenomenon, but an examination of the data shows that there is some overlap between two speakers of the same session, which may be responsible for a miscalibration of the normalisation process. The eResNet architecture, on the other hand, yields a clear improvement over the baseline, that compares favourably to previous work, and does not suffer from this miscalibration issue.

4.2. Discussion

Relative to recent state-of-the-art systems in SER [34, 35, 36], our method yields subpar performance. However, our focus on this work was to investigate whether the introduction of

Table 3: Average UAR% results for IEMOCAP on 4-class emotion classification using leave-one-speaker-out cross-validation. Target classes where angry, happy, neutral, sad.

Method	SVM	SVM-NORM	ResNet	ResNet-NORM	*TwoChannel*	eResNet	Li *et al.* [9]*	Tu *et al.* [7]*
UAR	48.9	50.3	52.3	47.1	52.0	**56.5**	59.9	57.3

* Performance reported by original authors. Settings not directly comparable to ours. See text for more details.

auxiliary embedding sub-networks can act as a new paradigm for speaker adaptation in contemporary deep learning (DL) systems. This is demonstrated by the relative improvements obtained by our methodology, which compare favourably to our feature normalization baselines, and recent DL methods for obtaining speaker-invariant representations [7, 9].

Our results further indicate that simply passing the neutral sample as a second input to an architecture is not enough for it to learn a mapping between a speaker's neutral voice and their current emotional state. Rather, conditioning each individual residual block on learnt embeddings seems to yield substantial performance improvements, which highlights the importance of the auxiliary sub-network.

We have not yet established a sufficient cause for that boost in performance. As previously stated, there is a connection between our work and *style transfer* approaches such as StyleGAN [18, 37]. There, the embeddings (referred to as *latent codes* by Karras *et al.* [18]) serve the purpose of locally conditioning each convolution block to learn a particular affine transformation that corresponds to a particular "style" (e. g., a person in an image wearing glasses or not). The goal of the embedding sub-network (referred to as *mapping network*) is then to generate latent codes that are able to disentangle the different styles, which then help the main generator apply those styles locally. Likewise, we expect our embedding sub-network to learn some particular speaker-specific characteristics, which are then used by the residual blocks of the main network in a similar manner to process specific parts of the input space (in our case log-spectrograms) while taking into account the speaker 'style'. This connection will be further explored in follow-up work.

Another important observation is that all networks were given the same computational budget (30 epochs). However, the eResNet architecture was bigger than the other two networks used in this work, which could have led to decreased performance compared to the smaller baseline architectures. This acts as further justification that this particular training formulation could allow architectures to learn the SER task in a more data-efficient manner.

4.3. Impact of neutral sample selection

To study the effects that the neutral sample selection has at test time, we evaluated the eResNet and *TwoChannel* architectures 10 times, each with a different neutral sample of the target subject. For our ablation experiments, we focus on the FAU-AIBO data set which has a well-defined test set that includes multiple speakers, since that allows us to get a better understanding of how different enrollment samples impact performance across different speakers for one particular model. We used the first 10 samples for each subject in the order they appear in the data set, and report mean and standard deviation of performance over those 10 runs in Table 4.

Table 4: Mean (standard deviation) of UAR% for FAU-AIBO when varying the neutral sample at test time

Classifier	**2-class**	**5-class**
TwoChannel	60.7 (0.6)	32.4 (0.7)
eResNet	67.3 (0.0)	40.5 (0.6)

This experiment shows that both architectures are fairly robust to the selection of the neutral sample for each speaker, as we observe very small differences in performance when varying our choice. In addition, we measure the statistical significance of the difference in the performance of the architectures using a two-sided t-test. For both the 2-class ($p < 0.0001$) and the 5-class ($p < 0.0001$) tasks, we observe statistically important differences at the 0.05 level.

5. CONCLUSION

In this work, we introduce a novel DNN framework for speaker adaptation, utilizing an additional sub-network that conditions the primary classification network on a single reference sample. We were able to show that our approach gives better performance than a baseline network without any adaptation, and that using an auxiliary embedding sub-network is beneficial to simply providing the network with the reference at its input.

From a practical perspective, results demonstrate that we can get performance improvements even compared to speaker normalisation approaches that have access to the full range of a speaker's emotions, while adapting to the target speaker in a data-efficient manner suitable for practical applications. Although our approach still lacks behind the current state-of-the-art for SER, the performance gains achieved for this particular setting show great promise. In the future, we intend to combine our method with recent approaches in SER [34, 35, 36] in order to bridge that performance gap.

6. References

[1] A. Batliner, J. Buckow, H. Niemann, E. Nöth, and V. Warnke, "The prosody module," in *Verbmobil: foundations of speech-to-speech translation*, Springer, 2000, pp. 106–121.

[2] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing," in *Proc. Affective Computing and Intelligent Interaction*, 2007, pp. 139–147.

[3] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech communication*, vol. 53, no. 5, pp. 768–785, 2011.

[4] C. L. Gohm and G. L. Clore, "Individual differences in emotional experience: Mapping available scales to processes," *Personality and Social Psychology Bulletin*, vol. 26, no. 6, pp. 679–697, 2000.

[5] J. J. Gross and O. P. John, "Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being.," *Journal of personality and social psychology*, vol. 85, no. 2, p. 348, 2003.

[6] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. Interspeech*, Brighton, UK, 2009, pp. 321–315.

[7] M. Tu, Y. Tang, J. Huang, X. He, and B. Zhou, "Towards adversarial learning of speaker-invariant representation for speech emotion recognition," *arXiv preprint arXiv:1903.09606*, 2019.

[8] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[9] H. Li, M. Tu, J. Huang, S. Narayanan, and P. Georgiou, "Speaker-invariant affective representation learning via adversarial training," *Proc. ICASSP*, 2020.

[10] V. Sethu, E. Ambikairajah, and J. Epps, "Speaker normalisation for speech-based emotion detection," in *Proc. IEEE conference on digital signal processing*, 2007, pp. 611–614.

[11] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[12] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, "Iterative feature normalization scheme for automatic emotion detection from speech," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 386–397, 2013.

[13] J. Wilting, E. Krahmer, and M. Swerts, "Real vs. acted emotional speech," in *Proc. Interspeech*, Pittsburgh, USA, 2006.

[14] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Proc. IEEE Conference on Multimedia and Expo*, IEEE, 2005, pp. 474–477.

[15] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to find trouble in communication," *Speech communication*, vol. 40, no. 1-2, pp. 117–143, 2003.

[16] G. Keren, J. Han, and B. Schuller, "Scaling speech enhancement in unseen environments with noise embeddings," pp. 25–29, 2018.

[17] S. Liu, G. Keren, and B. Schuller, "Single-channel speech separation with auxiliary speaker embeddings," *arXiv preprint arXiv:1906.09997*, 2019.

[18] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. CVPR*, 2019, pp. 4401–4410.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] S. Liu, G. Keren, and B. Schuller, "N-hans: Introducing the augsburg neuro-holistic audio-enhancement system," *arXiv preprint arXiv:1911.07062*, 2019.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *Proc. ICML*, 2015.

[22] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. Interspeech 2017*, pp. 2616–2620, 2017.

[23] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP*, Calgary, Canada, 2018, pp. 5329–5333.

[24] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[25] S. Steidl, *Automatic classification of emotion related user states in spontaneous children's speech*. University of Erlangen-Nuremberg Erlangen, Germany, 2009.

[26] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.

[27] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[28] E. Tzinis and A. Potamianos, "Segment-based speech emotion recognition using recurrent neural networks," in *Proc. Affective Computing and Intelligent Interaction*, 2017, pp. 190–195.

[29] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. ICASSP*, IEEE, 2017, pp. 2227–2231.

[30] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proc. ACM Multimedia*, Mountain View, CA, 2017, pp. 478–484.

[31] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[32] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[33] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia*, 2010, pp. 1459–1462.

[34] Z. Zhao, Y. Zhao, Z. Bao, H. Wang, Z. Zhang, and C. Li, "Deep spectrum feature representations for speech emotion recognition," in *Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data*, 2018, pp. 27–33.

[35] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, "Attention-enhanced connectionist temporal classification for discrete speech emotion recognition," *Proc. Interspeech*, pp. 206–210, 2019.

[36] G. Ramet, P. N. Garner, M. Baeriswyl, and A. Lazaridis, "Context-aware attention mechanism for speech emotion recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2018, pp. 126–131.

[37] G. Ghiasi, H. Lee, M. Kudlur, V. Dumoulin, and J. Shlens, "Exploring the structure of a real-time, arbitrary neural artistic stylization network," in *Proc. British Machine Vision Conference*, London, UK, 2017.