



EmoSpeechAuth: Emotion-Aware Speaker Verification

Magdalena Gołębiewska, Piotr Syga

Department of Artificial Intelligence, Wrocław University of Science and Technology, Poland

{magdalena.golebiowska, piotr.syga}@pwr.edu.pl

Abstract

Speaker verification is a process of verifying the identity of a user. Research has shown that emotional variability in speech degrades the performance of speaker verification tasks. Prior approaches were more computationally expensive and did not focus on the state-of-the-art speaker representations. In this paper, we propose a novel framework for constructing emotional speaker embeddings. Our framework utilizes pre-trained state-of-the-art feature extractors for speaker and emotion recognition, including both speaker and emotional information in the final embeddings. We present results of speaker verification on emotional speech datasets. We show that fusing ECAPA2 speaker representations and emotional features from emotion2vec with a cross-attention module improves EER by 8.29 percentage points compared to the baseline.

Index Terms: speaker verification, emotional speech, speaker embeddings

1. Introduction

In this work, we explore speaker verification (SV) with emotional speech. Speaker verification is a process of verifying the identity of a user from the analysis of input speech to decide whether to accept or reject the identity claim of a speaker [1].

Previous studies have proven that emotional variability in speech degrades the performance of speaker recognition tasks [2, 3, 4, 5]. It is an important factor to consider while designing a speaker verification system, as real-life situations are influenced by emotional cues, e.g., emergency calls or frustration caused by an inability to authenticate within a voice verification system. Key challenges include emotional mismatch between the speaker model and test utterances and introduction of intense intra-speaker vocal variability caused by different articulating styles of certain emotions [2], with anger identified as the worst performing emotion in verification [5], which is further elaborated in [4] establishing that the greatest performance drops are caused by extreme values of arousal, valence, and dominance.

Several approaches have been proposed to address this issue. A three-stage system in [6] introduced gender and emotion identification that preceded the final verification of the speaker. Gender, emotion, and speaker models were implemented as HMMs. The system achieved EER of 9.50% and 10.00% on the authors' collected dataset and EPST dataset respectively. Hybrid models combining GMM, HMM, and DNNs [7] achieved the average EER on three datasets (ESD, SUSAS, RAVDESS) of 7.19%, 16.85%, 11.51%, and 11.90%, respectively, for the HMM-DNN, DNN-HMM, DNN-GMM, and GMM-DNN models.

Numerous studies adapt a different strategy to improve

speaker verification, based on emotion-invariant speaker embeddings. Emotion-dependent score normalization (Enorm) was inspired by Hnorm, a method to alleviate the channel effect in cross-channel speaker verification [2]. Enorm outperformed the baseline system by about 2.875% as measured by EER. An encoder-decoder architecture [8] using i-vectors to map embeddings with different emotions to an emotion-invariant space achieved a 2.6% accuracy on the IEMOCAP database. In [9] authors used multiple methods to improve SV, including copy-based augmentation, cosine similarity loss, and emotion-aware masking (EM) based on speech signal energy. The network architecture was based on ResNet34-TSPP. The system achieved an overall decrease of 6.67% in EER compared to baseline.

Recent studies focused on creating embeddings that carry additional emotional information. Although the authors of [10] target speaker identification, their findings may be easily adapted to speaker verification. In the study, a multitask learning system was created that includes two individual Bi-LSTM networks. First, an input of 56-dimensional acoustic features was fed into a pre-trained speech emotion recognition (SER) model. The emotional embeddings obtained from this model acted as an input to speaker identification (SI) model along the original 56-dimensional input. The output of the system was the result of the simultaneous recognition of emotion labels and speaker identification. Compared to baseline, the system's EER improved by 0.40%.

Another way of creating emotion embeddings was proposed in [11]. A DNN was trained for each assumed emotion. These DNNs were then used as embedding extractors in the development phase to extend neutral speech embeddings. Embeddings from all emotional extractors were concatenated and weighted by self-attention mechanism. They achieved EER equal to 8.14% for the speaker verification on CREMA-D.

The most popular speaker recognition methods include Gaussian Mixture Models (GMMs) [12], Hidden Markov Models (HMMs) [13], i-vectors [14], x-vectors [15], embeddings extractors based on the ResNet architecture [16, 17, 18] and most recent networks, introducing transformer-based architectures and self-supervised learning [19], such as wav2vec [20], enhanced x-vector architecture known as ECAPA-TDNN [21] and its successor, ECAPA2 [22].

In this article, we propose a novel framework for constructing emotional speaker embeddings. Namely, we experiment with pre-trained speaker recognition and pre-trained speech emotion recognition front-ends to use emotion as an auxiliary data instead of a perturbation. We choose state-of-the-art speaker embeddings extractors, including ECAPA-TDNN, ECAPA2 and ResNet-TDNN, and present results of SV on emotion datasets. Our contributions are as follows.

1. To our knowledge, this is the first study that explored emo-

tional SV with ECAPA-TDNN and ECAPA2, which are state-of-the-art speaker representation extractors.

2. We propose a new framework for emotional speaker embeddings and assess it with different combinations of feature extractors.
3. Our work noticeably improves the verification efficacy of speech authentication systems in emotional environments.

2. System Architecture

In this section, we present details of our EmoSpeechAuth architecture used for emotional speaker verification. We describe pre-trained models utilized for speech emotion recognition and speaker verification.

ECAPA-TDNN was initially presented as an improvement to the x-vector architecture in [21] using 1D Res2Net modules with skip connections, squeeze-and-excitation modules, multi-layer feature aggregation and attentive statistical pooling. ECAPA-TDNN was found to focus overly on local feature extraction [23, 24]. We use the pre-trained version trained in VoxCeleb1 and VoxCeleb2 from SpeechBrain Toolkit¹², achieving EER of 0.80% on Voxceleb1-test set.

ECAPA2 [22] is a recent hybrid neural network architecture that combines 1D and 2D convolutional operations, featuring Local Feature Extractor Blocks, a TDNN-based Global Feature Extractor, and channel-dependent statistics pooling. In our framework, we adapt the pre-trained model³ from the original paper trained on the development part of the VoxCeleb2 dataset. The model achieved EER equal to 0.34%, 0.52%, and 0.99% on Vox1-O, Vox1-E, and Vox1-H, respectively.

In our experiments we use ResNet-TDNN⁴ with basic ResNet blocks, squeeze-and-excitation modules and an attention layer. The utilized version from the SpeechBrain Toolkit⁵ was trained on VoxCeleb1 and VoxCeleb2 training data. On VoxCeleb1 test set it achieved 1.05% EER.

Wav2vec2.0 was proposed in [25] as a system for self-supervised learning of speech representations through a multi-layer convolutional feature encoder and transformer via contrastive learning. We use the version published on HuggingFace⁶ trained on Acted Emotional Speech Dynamic Database (AESDD) for speech emotion recognition.

Emotion2vec [26], designed for speech emotion recognition, uses a 1-D convolutional network as a feature extractor and data2vec as a backbone network. The model was pre-trained with self-supervised student-teacher strategy and later fine-tuned with supervised learning. Our system used embeddings produced by the base version published by the original authors on HuggingFace⁷.

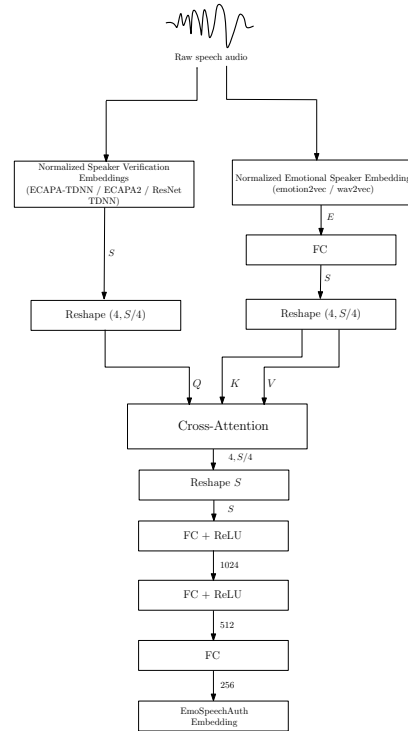


Figure 1: The network architecture of the proposed speaker verification approach.

2.1. EmoSpeechAuth System

To improve emotional speaker verification, we suggest an approach that uses two feature extractors at the same time, one for providing emotional embeddings and the other to supply speaker representations. The network architecture is shown in Fig. 1. Initially, a pre-trained model is used to convert raw voice data into speaker and emotion embeddings. We test every possible pairing of the selected extractors, that is, every SV model is coupled with every speech emotion recognition (SER) model.

Embeddings obtained from the pre-trained models are normalized and emotional speaker embedding is projected onto dimensions of the speaker verification embedding S with a linear layer. We reshape both embeddings to dimensions $(4, S/4)$ to mimic a sequential input. Embeddings prepared in such a way are fed into a one-headed cross-attention block. Speaker features act as queries to explore how emotion influences speaker identity. Output is flattened back to the original shape and passed through two hidden fully connected layers with a ReLU activation, downsampling the size to 1024 and 512 consecutively. The final embeddings are produced by a linear projection to 256 dimensions.

3. Experimental Setup

3.1. Datasets

We trained and tested our system on five datasets in English language with emotional speech: IEMOCAP [27], CREMA-D [28], RAVDESS [29], SAVEE [30] and TESS [31].

IEMOCAP includes speech data from 10 speakers recorded during 5 sessions. One female and one male actor were recorded during each session on a pre-defined topic. Each sentence is categorized into one of 8 emotions (angry, happy,

¹M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J. Chou, S. Yeh, S. Fu, C. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. De Mori, Y. Bengio, “SpeechBrain: A General Purpose Speech Toolkit”, 2021, arXiv:2106.04624.

²<https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb>

³<https://huggingface.co/Jenthe/ECAPA2>

⁴X. Qin, N. Li, Y. Lin, Y. Ding, C. Weng, D. Su, M. Li, “The DKU-Tencent System for the VoxCeleb Speaker Recognition Challenge 2022”, 2022, 10.48550/arXiv.2210.05092.

⁵<https://huggingface.co/speechbrain/spkrec-resnet-voxceleb>

⁶<https://huggingface.co/harshit345/xlsr-wav2vec-speech-emotion-recognition>

⁷https://huggingface.co/emotion2vec/emotion2vec_base

neutral, sad, disgust, fear, frustration, surprise and excited). CREMA-D is an audiovisual dataset collected from 91 actors of various ethnic backgrounds. The actors recorded 12 defined sentences in 6 target emotions (happy, sad, anger, fear, disgust, and neutral). RAVDESS consists of recordings of 12 female and 12 male actors. The actors said two sentences, each at two levels of emotional intensity, per emotion. The speech dataset includes 8 emotions (calm, happy, sad, angry, fearful, surprise, disgust, and neutral). SAVEE database consists of speech samples from 4 male actors in 7 different emotions (happy, sad, angry, fear, surprise, disgust, and neutral), resulting in 480 English utterances in total. TESS is performed by two female actresses. A predefined list of 200 words in English was recorded in each of 7 target emotions (happiness, sadness, angry, fear, pleasant surprise, disgust, and neutral).

We decided to merge these datasets in order to increase the number of speakers. In total, for training and validation purposes, our combined data set had 131 speakers with more than 18 hours of emotional speech recordings. The resulting combined dataset was not balanced as the original datasets varied across included emotion categories, that is calm, excited, surprised, and frustrated were not shared in all five data sets. We used the 70-15-15 training-validation-testing split, dividing the dataset by subjects.

We present evaluation of our framework with the Equal Error Rate (EER) metric commonly used for speaker verification tasks, which specifies equality of missclassification of positive and negative data, aiming to be the lowest as possible for the best performance. We present EER as percentages. We also assess the performance of the models with histogram intersection of pair-wise distances between positive pairs (where both samples in the pair originate from the same speaker) and negative pairs (where the samples in the pair are from different speakers).

3.2. Environment and Training

We reran our experiments 5 times to estimate model skill and stability, averaging the results. Each experiment was run on a single NVIDIA TESLA P100 GPU (16GB VRAM). Code base for this paper is available in a Github repository⁸.

Audio samples were preprocessed as required by the front-ends, that is, conversion to mono-channel and resampling to 16 kHz. To conserve resources, embeddings were produced and saved to files prior to training. The feature extractors were not optimized, that is their weights were frozen. We performed basic augmentation on the embeddings with probability $1/2$, including applying Gaussian noise ($\mu = 0$, $\sigma = 0.1$), scaling the embedding between 0.8 and 1.2, and masking (zeroing) the elements from the embedding with probability 0.5.

The system was trained via unsupervised learning using a siamese network with AdamW optimizer (learning rate of 10^{-4} , weight decay of 10^{-4}) and cosine contrastive loss for 100 epochs, batch size 8. If validation EER stagnated for 10 epochs, learning rate was reduced by 0.1. Input pairs were dynamically generated with a $1/2$ probability of forming either a positive pair or a negative pair for each sample.

4. Results

We tested our framework on the testing dataset. The results are presented in Tab. 1. The system demonstrates a notable enhancement when utilizing both speaker and emotional embeddings in contrast to the baseline that used only speaker em-

Table 1: *The comparison of EER scores on the testing dataset. The table lists averages and standard deviation from 5 reruns. Including emotional features in the speaker verification task led to a significant improvements for all models included in the experiment. The best score is presented in bold.*

Speaker embd Emo embd	ECAPA-TDNN	ECAPA2	ResNet-TDNN
None	21.19% ± 0.59%	18.72% ± 0.59%	18.82% ± 0.43%
emotion2vec	12.81% ± 0.26%	11.36% ± 0.10%	13.55% ± 0.32%
wav2vec	13.214% ± 0.057%	12.919% ± 0.061%	12.881% ± 0.064%

Table 2: *The comparison of histogram overlap of positive and negative pair-wise distances on the testing dataset. The table lists averages and standard deviation from 5 reruns. Adding emotional information pulled away distributions of distances for positive and negative sample pairs.*

Speaker embd Emo embd	ECAPA-TDNN	ECAPA2	ResNet-TDNN
None	0.4330 ± 0.0035	0.334 ± 0.017	0.349 ± 0.014
emotion2vec	0.2171 ± 0.0111	0.1898 ± 0.0044	0.217 ± 0.011
wav2vec	0.2147 ± 0.0035	0.2129 ± 0.0026	0.2124 ± 0.0043

beddings (shown by *None* in the emotion embedding column; EER 21.19%, 18.72%, 18.82% for ECAPA-TDNN, ECAPA2, ResNet-TDNN, respectively). Combining ECAPA-TDNN with different emotional front-ends led to a relative improvement of 24.77% and 12.64%, for emotion2vec and wav2vec, respectively. Using ECAPA2 speaker embeddings with emotion2vec resulted in a relative enhancement of 36.98% and 14.05% with wav2vec. Lastly, the application of ResNet-TDNN embeddings in conjunction with emotion2vec yielded a relative improvement of 34.91% and 19.17% with wav2vec. In general, the greatest enhancement was observed in ECAPA2 with emotion2vec, while the smallest change was observed for ECAPA-TDNN with wav2vec. Combining speaker embeddings with emotion2vec embeddings gave superior results compared to wav2vec. The lowest EER was achieved by already mentioned ECAPA2 with emotion2vec.

Furthermore, we assessed our framework by measuring the intersection of pairwise distance histograms for positive and negative sample pairs listed in Tab. 2. The histogram overlap for the baseline models was included in the row denoted by *None*. The pairwise distance distributions for positive and negative pairs were found to be farther apart when emotional embed-

⁸Code base is available at <https://github.com/mgraves236/EmoSpeechAuth>

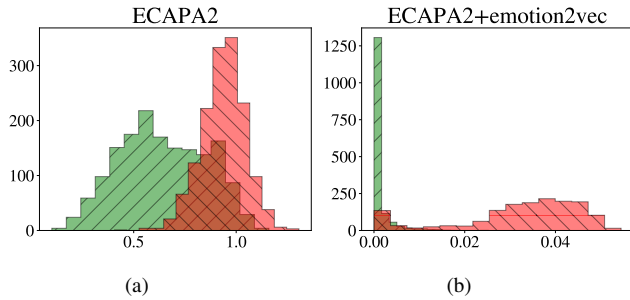


Figure 2: Example of pair-wise distance histograms for positive (green) and negative (red) samples from the test set, generated dynamically with a 0.5 chance of each for (a) baseline model ECAPA2 and (b) our framework using ECAPA2 and emotion2vec embeddings, which reduces intersection and increases separation between different-speaker embeddings.

dings were added to speaker verification embeddings than when the baseline systems were used (cf. Fig. 2). ECAPA-TDNN with emotion2vec showed the largest relative change. Additionally, ECAPA2 was identified as the best performing speaker representation with emotional speech (cf. Tab. 1).

4.1. Ablation Study

In addition to the system architecture presented, we evaluated two other slightly modified designs. First, we tried concatenating the normalized embeddings and passing them through the same sequence of fully connected layers as in the original system. Comparing results from Tab. 3 to the proposed Emo-SpeechAuth architecture (Tab. 1), it was observed that the use of a more complicated attention block was superior to a simple concatenation. The greatest EER difference between the two architectures, reaching 5.30 percentage points, occurred when ECAPA-TDNN and wav2vec were combined. However, EER improved by 1.30 percentage points for ResNet-TDNN and emotion2vec concatenation. We also tested multi-head attention with 4, 6 and 8 heads. However, the input sequences are not long (48) and adding more heads does not improve the performance of the entire system.

Another modification analyzed was the change of ReLu to the Mish [32] activation function in the architecture with cross-attention. As seen in Tab. 4 the difference in the results compared to our original architecture was not as high as previously; the biggest drop of 1.17 percentage points was registered for ECAPA2 and emotion2vec. Choosing Mish was advantageous for the combination of ECAPA-TDNN + emotion2vec and ResNet-TDNN + emotion2vec, decreasing EER by 0.080 percentage points and 1.098 percentage points, respectively. We demonstrated that there might be some alterations that would help individual front-end configurations. However, the majority of extractors worked the best with the cross-attention fusion, which is proposed as the final architecture.

5. Conclusion

In this paper, we show that combining speaker verification and emotion recognition embeddings from pre-trained models improves the performance of speaker verification systems in emotional environments. Specifically, fusing ECAPA2 and emotion2vec embeddings with cross attention improves EER rela-

Table 3: The comparison of EER scores on the testing dataset for the modified architecture with embedding concatenation. The table lists averages and standard deviation from 5 reruns.

Speaker embd \ Emo embd	ECAPA-TDNN	ECAPA2	ResNet-TDNN
emotion2vec	15.946% ± 0.032%	11.799% ± 0.061%	12.251% ± 0.024%
wav2vec	18.51% ± 0.11%	16.092% ± 0.098%	15.213% ± 0.039%

Table 4: The comparison of EER scores on the testing dataset for the modified architecture with the Mish activation function. The table lists averages and standard deviation from 5 reruns.

Speaker embd \ Emo embd	ECAPA-TDNN	ECAPA2	ResNet-TDNN
emotion2vec	12.72% ± 0.12%	12.53% ± 0.11%	12.46% ± 0.15%
wav2vec	13.79% ± 0.11%	13.140% ± 0.053%	13.74% ± 0.13%

tively by 36.98% (baseline EER 21.10% vs. ours EER 12.81%). Compared to previous approaches, our framework requires a little computational cost because our system only needs to train a few fully linked layers and the cross-attention block. We identified a few limitations of our study. The major problem is the unavailability of emotional speech data. We utilized five different datasets that altogether included 131 speakers. Speaker embeddings systems are usually trained on datasets with a few thousand speakers [21, 22]. We consider it crucial to construct publicly available unbiased [33] emotional speech datasets with a larger number of speakers to truly advance the quality of speaker verification with emotional speech. Additionally, the performance of our system depends on the quality of the front-ends used. In future work, we would like to obtain a model that is capable of generalizing well on data from different distributions, which requires bigger emotional speech dataset. Continuous speaker verification is also worth exploring and demands a dataset with longer emotional speech segments. Additionally, more work has to be done to analyze a multilingual setup, to verify if the emotional information adapt in crosslingual setting. Furthermore, we are interested in exploring more combinations of speaker verification end emotion recognition front-ends.

6. Acknowledgements

This work has been partially funded by Department of Artificial Intelligence, Wrocław University of Science and Technology.

7. References

- [1] B. H. Juang, M. Sondhi, and L. R. Rabiner, "Digital speech processing," in *Encyclopedia of Physical Science and Technology (Third Edition)*, third edition ed., R. A. Meyers, Ed. New York:

- Academic Press, 2003, pp. 485–500. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B0122274105001782>
- [2] W. Wu, F. Zheng, M. Xu, and H. Bao, “Study on speaker verification on emotional speech,” in *Proceedings of Interspeech*, 09 2006.
 - [3] S. Parthasarathy and C. Busso, “Predicting speaker recognition reliability by considering emotional content,” in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 434–439.
 - [4] S. Parthasarathy, C. Zhang, J. H. Hansen, and C. Busso, “A study of speaker verification performance with expressive speech,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5540–5544.
 - [5] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, “X-vectors meet emotions: A study on dependencies between emotion and speaker recognition,” *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7169–7173, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:211083028>
 - [6] I. Shahin, “Speaker verification in emotional talking environments based on three-stage framework,” in *2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, 2017, pp. 1–5.
 - [7] I. Shahin, A. B. Nassif, N. Nemmour, A. Elnagar, A. Alhudaif, and K. Polat, “Novel hybrid dnn approaches for speaker verification in emotional and stressful talking environments,” *Neural Computing and Applications*, vol. 33, no. 23, p. 16033–16055, Jun. 2021. [Online]. Available: <http://dx.doi.org/10.1007/s00521-021-06226-w>
 - [8] B. D. Sarma and R. K. Das, “Emotion invariant speaker embeddings for speaker identification with emotional speech,” in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 610–615.
 - [9] J. Tian, X. Hu, and X. Xu, “Learning emotion-invariant speaker representations for speaker verification,” in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 611–10 615.
 - [10] K. Noh and H. Jeong, “Emotion-aware speaker identification with transfer learning,” *IEEE Access*, vol. 11, pp. 77 292–77 306, 2023.
 - [11] D. Li, Z. Yang, J. Liu, H. Yang, and Z. Wang, “Emotion embedding framework with emotional self-attention mechanism for speaker recognition,” *Expert Systems with Applications*, vol. 238, p. 122244, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S095741742302746X>
 - [12] D. A. Reynolds, “Speaker identification and verification using gaussian mixture speaker models,” *Speech Communication*, vol. 17, no. 1, pp. 91–108, 1995. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/016763939500009D>
 - [13] L. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
 - [14] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
 - [15] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.
 - [16] Y. Khokhlov, A. Zatorvitskiy, I. Medennikov, I. Sorokin, T. Prisyach, A. Romanenko, A. Mitrofanov, V. Bataev, A. Andrusenko, M. Korenevskaya, and O. Petrov, “R-vectors: New technique for adaptation to room acoustics,” in *Interspeech 2019*, 2019, pp. 1243–1247.
 - [17] X. Qin, N. Li, Y. Lin, Y. Ding, C. Weng, D. Su, and M. Li, “The dku-tencent system for the voxceleb speaker recognition challenge 2022,” 10 2022.
 - [18] J. Thienpondt, B. Desplanques, and K. Demuynck, “Integrating frequency translational invariance in tdnn and frequency positional information in 2d resnets to enhance speaker verification,” 08 2021, pp. 2302–2306.
 - [19] V. Pankov, V. Pronina, A. Kuzmin, M. Borisov, N. Usoltsev, X. Zeng, A. Golubkov, N. Ermolenko, A. Shirshova, and Y. Matveeva, “Dino-vits: Data-efficient zero-shot tts with self-supervised speaker verification loss for noise robustness,” in *Interspeech 2024*. ISCA, Sep. 2024, p. 697–701. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2024-549>
 - [20] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” *CoRR*, vol. abs/1904.05862, 2019. [Online]. Available: <http://arxiv.org/abs/1904.05862>
 - [21] B. Desplanques, J. Thienpondt, and K. Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” 10 2020.
 - [22] J. Thienpondt and K. Demuynck, “Ecapa2: A hybrid neural network architecture and training strategy for robust speaker embeddings,” *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:267027744>
 - [23] J. Yao, C. Liang, Z. Peng, B. Zhang, and X.-L. Zhang, “Branch-ecapa-tdnn: A parallel branch architecture to capture local and global features for speaker verification,” in *INTERSPEECH 2023*. ISCA, Aug. 2023, p. 1943–1947. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2023-402>
 - [24] S.-H. Liou, P.-C. Chan, C.-P. Chen, T.-C. Lin, C.-L. Lu, Y.-H. Cheng, H.-F. Chuang, and W.-Y. Chen, “Enhancing ecapa-tdnn with feature processing module and attention mechanism for speaker verification,” in *Interspeech 2024*. ISCA, Sep. 2024, p. 2120–2124. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2024-601>
 - [25] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” in *Interspeech*, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:266840788>
 - [26] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, “emotion2vec: Self-supervised pre-training for speech emotion representation,” *Proc. ACL 2024 Findings*, 2024.
 - [27] 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. 335–359, Nov. 2008. [Online]. Available: <http://dx.doi.org/10.1007/s10579-008-9076-6>
 - [28] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, “Crema-d: Crowd-sourced emotional multimodal actors dataset,” *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.
 - [29] S. R. Livingstone and F. A. Russo, “The ryerson audio-visual database of emotional speech and song (ravdess),” 2018. [Online]. Available: <https://zenodo.org/doi/10.5281/zenodo.1188976>
 - [30] P. Jackson and S. ul haq, “Surrey audio-visual expressed emotion (savee) database,” 04 2011.
 - [31] M. K. Pichora-Fuller and K. Dupuis, “Toronto emotional speech set (TESS),” 2020. [Online]. Available: <https://doi.org/10.5683/SP2/E8H2MF>
 - [32] D. Misra, “Mish: A self regularized non-monotonic activation function,” in *British Machine Vision Conference*, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:221113156>
 - [33] J. Pahl, I. Rieger, A. Möller, T. Wittenberg, and U. Schmid, “Female, white, 27? bias evaluation on data and algorithms for affect recognition in faces,” in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 973–987. [Online]. Available: <https://doi.org/10.1145/3531146.3533159>