Master's Thesis Computing Science

Speaker Similarity for Emotional Speech

Dmitrii Mikhailovskii
s1092973

August 12, 2024

Company:
ReadSpeaker B.V.

Company supervisor:
Aki Kunikoshi, Senior Speech Scientist

University supervisor:
Prof. David A. van Leeuwen

Second assessor:
Dr. Louis F.M. ten Bosch

Radboud University

Abstract

This thesis explores the enhancement of automatic speaker verification (ASV) systems' robustness to emotional speech. Utilizing the emotional datasets such as CREMA-D and RAVDESS, the study focuses on optimizing a state-of-the-art WavLM-based ASV system with ECAPA-TDNN to address performance degradation caused by emotional variability. Traditional similarity measures and embedding space modifications, such as LDA, PLDA, and contrastive learning, offer minimal or no improvements. In contrast, fine-tuning the ECAPA-TDNN system with the incorporation of the modified Barlow Twins objective, cosine loss, CopyPaste augmentation, and dataset extension through pitch shifting led to a significant performance enhancements.

The study also evaluated the model's generalization across different emotional datasets, demonstrating improved robustness to emotional variability at the cost of slight performance reductions in original context, as observed in VoxCeleb1 evaluations. Cross-dataset tests further highlighted the challenges of achieving universal emotion-robustness, underscoring the importance of dataset-specific optimization. An ablation study highlighted the critical role of modified loss functions and augmentations in enhancing system performance.

Contents

Chapter 1

Introduction

Automatic speaker verification (ASV) is an application of speaker recognition technologies that aims to determine whether two speech samples originate from the same speaker. This problem has wide-ranging applications, from security systems that verify individual identities to evaluating text-to-speech models, ensuring that synthesized speech accurately reflects the characteristics of the original speaker.

Developing a robust speaker similarity measure for emotional speech has significant implications across various fields. In security and surveillance, it can enhance forensic analysis and surveillance systems by accurately identifying individuals in emotionally charged situations. In telecommunications, it can improve customer service in call centers and strengthen voice authentication systems. However, given that this thesis is conducted within a company focused on text-to-speech synthesis systems, one of the primary real-world applications is developing a metric to evaluate speaker identity preservation within synthesized emotional speech.

Recent advancements in ASV technology, driven by improvements in machine learning and signal processing techniques, have significantly enhanced speaker verification capabilities. Despite these advancements, accurately measuring speaker similarity, particularly in the presence of emotional speech, remains a challenging problem. Emotional speech introduces variability that can significantly affect the spectral and temporal characteristics of speech signals, thus complicating speaker recognition tasks.

Emotional expressions, such as happiness, anger, sadness, and fear, can alter fundamental speech features like pitch, tone, and prosody. These changes pose substantial challenges for traditional speaker recognition systems, which often rely on features assumed to be relatively invariant across different speaking conditions. Consequently, the variability introduced by emotions can lead to increased error rates in speaker verification systems, undermining their reliability in real-world applications where emotional speech is common.

Therefore, the primary objective of this thesis is to develop a robust

speaker similarity measure that can accurately verify speakers regardless of their emotional state. Specifically, this research aims to create emotion-invariant representations using a state-of-the-art model. By focusing on the development of these representations, the study seeks to enhance the quality and reliability of speaker recognition systems in the presence of emotional variability.

To achieve this, this thesis will:

1. Analyze the impact of various emotional states on speaker similarity.
2. Develop a model that generates emotion-invariant speaker representations.
3. Evaluate the effectiveness of that model in maintaining speaker identity across different emotional states.

By examining the impact of emotions on speaker similarity, this study aims to improve the robustness of speaker recognition systems in diverse real-world environments. Ultimately, the findings from this research will contribute to the advancement of ASV technology, particularly in applications involving emotional speech.

The structure of this thesis is organized into four chapters. Chapter 2 delves into the background knowledge related to speaker verification in general, with a focus on emotional speech and associated challenges to increase ASV robustness. Chapter 3 outlines the experimental setup, describes the specific steps involved in developing the proposed emotion-invariant speaker representations, and presents the results of these experiments, followed by a detailed analysis. Finally, Chapter 4 concludes the thesis by summarizing the key findings, discussing their implications, and providing insights for future research directions.

Chapter 2

Related Work

2.1 Classic Speaker Recognition Approaches

Speaker recognition has significantly evolved, employing various methods to recognize speakers based on their voice characteristics.

Some of the earliest modern speaker verification systems utilized features extracted using the Gaussian Mixture Model (GMM) and compensating speaker and channel variability using Joint Factor Analysis (JFA) [KOD+08]. This approach was built on the success of GMM-Universal Background Model (UBM) systems [RQD00], which employed acoustic features, typically Mel-Frequency Cepstral Coefficients (MFCCs), to model speakers. JFA was introduced to tackle intersession variability, a major challenge in GMM-UBM systems, by separately modeling inter-speaker variability and channel/session variability. In verification tasks, the system decides whether the speakers in utterances are the same or not by computing the likelihood of the test utterance feature vectors against a session-compensated speaker model.

Later research revealed that channel factors in JFA also contained speaker information, leading to the development of i-vectors [DKD+11]. These vectors combine speaker and channel spaces into a single total variability space instead of separating them as in JFA. i-vectors, based on supervector representations derived from GMMs, facilitate efficient matrix-vector operations. Typically, Probabilistic Linear Discriminant Analysis (PLDA) [Iof06] is later employed afterward to obtain similarity scores between the speakers in utterances.

Recent advancements in machine learning have introduced neural networks and deep learning techniques to ASV and other speaker recognition tasks. Initially, these were used to generate fixed-length representations of audio, but they have since evolved to accommodate variable-length inputs as well. These methods, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have significantly improved the accuracy and robustness of speaker verification systems, particularly in short-length

audio scenarios.

One of the most recent and well-known supervised architectures for the ASV is the Emphasized Channel Attention, Propagation, and Aggregation in Time Delay Neural Network (ECAPA-TDNN) [DTD20]. This architecture incorporates multiple enhancements based on recent results in face verification and computer vision, applied to the successful speaker verification architecture TDNN [WHH$^+$89] which, in turn, uses statistical pooling for the projection of variable-length utterances into fixed-length speaker representations. The standard input features of ECAPA-TDNN are 80-dimensional MFCCs extracted from a 25 ms window with a 10 ms frameshift.

For the training of the speaker representations, an extra classification layer with a Softmax is used. An key enhancement in the training of ECAPA-TDNN is Additive Angular Margin Softmax (AAM-Softmax) [DGXZ19], whose formula is given in 2.1. This advanced version of traditional Softmax normalizes the weights $W \in \mathbb{R}^{d\times n}$ and features $x_i \in \mathbb{R}^d$, making the loss dependent on the angle between them only. Additionally, it introduces an additive angular margin m to enhance intra-class compactness and inter-class separation.

$$L_{\text{AAM}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos(\angle(W_{y_i},x_i)+m)}}{e^{s\cos(\angle(W_{y_i},x_i)+m)}+\sum_{j=1,j\neq y_i}^{N}e^{s\cos\angle(W_j,x_i)}} \tag{2.1}$$

As a result of AAM-Softmax, modern ASV systems no longer require PLDA for scoring, as simple cosine similarity becomes sufficiently effective due to the properties of AAM-Softmax.

Recent developments in self-supervised learning have also influenced speech processing tasks. HuBERT [HBT$^+$21] and especially its later advancement WavLM [CWC$^+$22] are pre-trained on large amounts of unlabeled data, with WavLM being particularly adaptable across various tasks.

So one of the state-of-the-art ASV systems now uses WavLM hidden states as an alternative to MFCCs, feeding them into ECAPA-TDNN to generate speaker representations.

2.2 Impact of Emotional Variation

Although emotion recognition in speech has been extensively studied, its impact on speaker verification remains underexplored, particularly regarding the robustness of ASV systems when encountering emotional speech. Emotional speech tends to increase the variance in speaker representations, leading to higher false positive and false negative rates.

Recently, some approaches have been developed to generate emotion-invariant speaker representations. In speaker identification based on the *i*-vector principle [SD20], both neutral and emotional speech are used as

inputs, but only neutral speech of the same speaker is used as a target for transformation. This technique aligns emotional speaker representations more closely with neutral representations, thereby enhancing the system's robustness to emotional variations.

In training emotion-invariant ASV system based on ResNet34-TDSP [THX24], augmentation techniques such as emotion-aware masking and "CopyPaste" are employed. Emotion-aware masking uses the root mean square energy of the speech signal to mask parts of the signal where emotion expression is most significant. CopyPaste augmentation creates a new utterance by splicing segments from different utterances of the speaker, possibly with different emotions. It introduces more textual and emotional diversity in the training samples, improving robustness and making the system less sensitive to emotional variation. These augmentations are applied to copies of the same sample and then fed into the model, optimized for both samples using AAM-Softmax loss. To further enforce convergence of the same-speaker representations, the cosine similarity between the speaker representations of these samples is maximized. This architecture is visualized in figure 2.1.

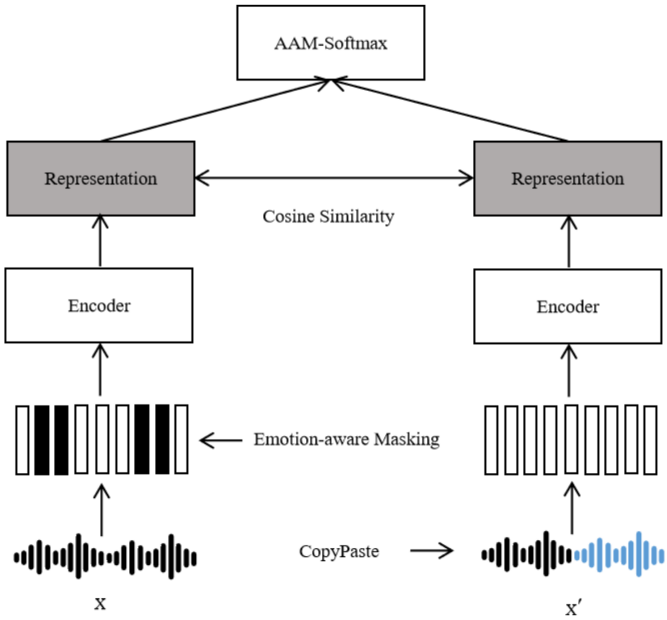


Figure 2.1: Architecture of the emotion-invariant ASV system [THX24]

2.3 Robustness of ASV

The emotional dependency of ASV systems can be viewed more broadly as a robustness problem and the challenge of handling out-of-domain data. While this issue has been more thoroughly studied in other contexts, it

remains under-researched, particularly in real-world, emotionally charged scenarios. Beyond the standard techniques to improve ASV robustness like data augmentations, regularization, and the use of robust architectures in general that improve robustness by preventing overfitting, there are relatively few other methods.

One notable method addressing robustness is the Barlow Twins objective [ZJM+21], which learns distortion-invariant representations and disentangles generated features from each other. Originally developed for self-supervised learning in computer vision, this method (as illustrated in figure 2.2) creates pairs of independently distorted batches, generates embeddings using shared encoder and projector networks, and computes the empirical cross-correlation of these embeddings. The goal is to make the features in these embeddings independent, with the target cross-correlation matrix $\mathcal{C}$ being the identity matrix:

$$L_{\mathrm{BT}} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2.$$

For the downstream tasks, only the encoder is used to obtain the representations without applying distortions.

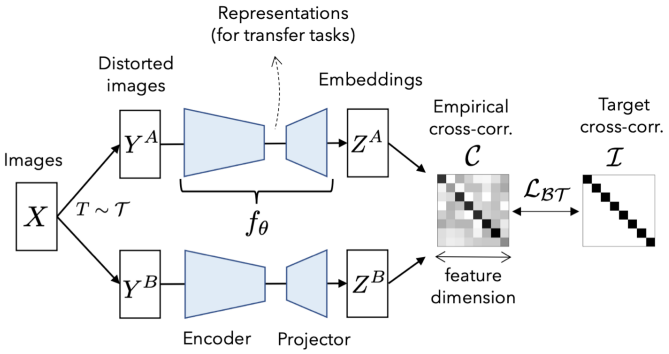


Figure 2.2: Visualization of the Barlow Twins objective [ZJM+21]

Recently, this approach has been adapted for the distillation of self-supervised speech models [RGHN+23], with modifications to handle outputs from different layers of neural networks and avoid averaging over the time dimension. Unlike the original approach, this adaptation does not employ a projector network during training, instead using layer outputs directly.

Chapter 3

Experimental Setup and Results

This chapter presents the datasets, methods, outcomes, and analysis of our experiments in developing an emotion-robust ASV system.

3.1 Datasets

The primary dataset used for developing the emotion-robust ASV system is CREMA-D, which comprises 7442 speech samples in English from 91 actors expressing six different emotions: neutral, anger, sadness, happiness, fear, and disgust. The test set for the system includes audio samples from 10 randomly selected speakers who are excluded during training. Although the emotions in this dataset are acted, which may simplify the task compared to real-world scenarios, the primary application of this research is in text-to-speech synthesis, which often uses acted emotions during training. Additionally, in real emotions, the emotional tone can change throughout the utterance, whereas in synthesized speech a single emotion is typically applied to the specified part of utterance.

In addition to CREMA-D, the RAVDESS emotional dataset is utilized. RAVDESS, similar to CREMA-D in terms that it features acted emotions, includes 1440 utterances from 24 speakers. It add calm and surprised emotions to the previously mentioned set and serves to evaluate the performance on out-of-domain data. Evaluation is conducted on 5 randomly selected speakers, while the remaining 19 speakers are used in training.

3.2 Model architecture

For our experiments, we utilize one of the current state-of-the-art ASV systems that is based on WavLM Large [CWC+22] architecture, which is pre-trained for diverse speech-related tasks. This model consists of a

convolutional feature encoder and 24 transformer encoder layers with gated relative position bias as shown in figure 3.1.

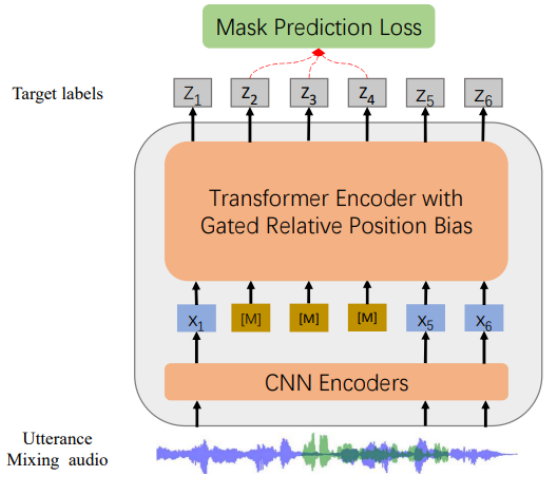


Figure 3.1: Visualization of WavLM architecture [CWC+22]. This image shows the pre-training process, in our application we use it in inference mode without mixing utterances and masking

For the ASV task, ECAPA-TDNN (small) is used as a downstream model. This model uses frozen pre-trained WavLM hidden states summed with trainable weights as input. The ASV model is pre-trained on the VoxCeleb1 [NCZ17] dataset using AAM-Softmax loss.

3.3 Metrics

The standard metric for ASV is the equal error rate (EER). It is calculated based on a set of trials where the system is presented with voice samples and must decide whether the speaker's claimed identity is target or not. EER is derived from the detection error trade-off (DET) curve of the trials, which plots the false negative rate against the false positive rate at various threshold settings. The EER is the point on this curve where the rates of false positives and false negatives are identical, providing a clear and concise single-number measure of a system's discrimination capability without dependence on the specific threshold settings used in the ROC and DET analysis, as shown in figure 3.2.

However, our goal is not only to improve the overall EER on emotional speech, but also to bridge the gap between the performance on neutral speech and (cross)-emotional utterances. By selecting specific samples in the trials list, we can analyze the impact of emotions on the system's performance.

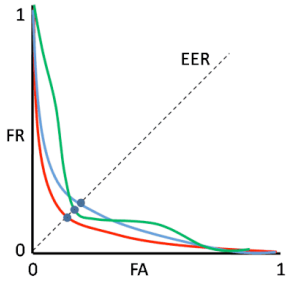


Figure 3.2: Visualization of the EER calculation [GPNFRS12]

For this reason, in addition to calculating the total EER across all test trials, regardless of emotional labels of utterances, we also calculate the EERs for trials consisting of utterance pairs with specified emotions. These results are visualized in a matrix format, where columns represent the emotion of the first utterance and the rows represent the emotion of the second utterance. We will further refer to this as the *EERs matrix*. Additionally, we measure the performance gap by calculating the difference between maximum and minimum EERs in this matrix, denoted as Δ^{EER}. Our aim is not only to reduce the EERs, but also to minimize this gap between neutral and emotional speech. It is important to note that these metrics are effective proxies for evaluating the maintenance of speaker identity in emotional settings only when considered together. Focusing solely on improving EER could lead to significant variance in emotion-specific EERs, resulting in unreliable outcomes. Conversely, only minimizing Δ^{EER} could reduce emotion-specific variance at the expense of accurate speaker verification, potentially leading to a system that is ineffective in practical applications.

3.4 Methods

First, we validate that WavLM-based ASV performance declines with emotional speech. There are several approaches to improve robustness against emotional variability:

- Similarity Metric Substitution. Experimenting with metrics like PLDA instead of cosine similarity.
- Embedding Space Mapping. Transforming speaker representations into a space less influenced by emotional variability.
- Embedding Enhancement via Fine-Tuning. Fine-tune model parameters (ECAPA-TDNN layers) to better handle emotional speech.

In this thesis, we investigate each enhancement approach above as follows:

- We examine the impact of LDA and PLDA as standard methods applied to scoring, and Spherical PLDA [SKLC23], a recent modification of PLDA, on system performance to see if changing the scoring function can better capture the speaker information.
- We apply contrastive learning to refine the embeddings, aiming to create a more emotion-agnostic embedding space.
- Finally, we fine-tune the ECAPA-TDNN on WavLM features using various strategies including:
  - Standard augmentations: MUSAN music, noise and speech dataset [SCP15], Room Impulse Response and Noise Database.
  - Modified Barlow Twins objective to reduce redundancy in speaker representations.
  - CopyPaste augmentation for reduction of text and emotion dependency.
  - Cosine loss to bring speaker representations of the same speaker closer to each other.

3.5 Results

3.5.1 Preliminary Results

We begin by examining the CREMA-D EERs matrix across different emotion pairs using pre-trained speaker representations. As shown in figure 3.3, there is a significant imbalance in the performance of the system across different emotions.

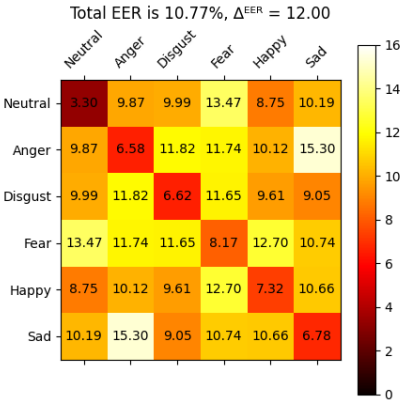


Figure 3.3: EERs matrix of the speaker representations pre-trained on VoxCeleb1; here and in the next matrices, emotions in each cell are selected for both target and non-target speakers

Next, we examine the distribution of these representations using t-SNE [vdMH08]. We fit and transform the same set of representations without speaker labels using this method.

As shown in figure 3.4, the clear clustering of speakers is observed, suggesting the potential for improved scoring functions or embedding mappings. However, t-SNE's complex transformation may limit prediction quality when it is applied to individual samples rather than a set of representations.

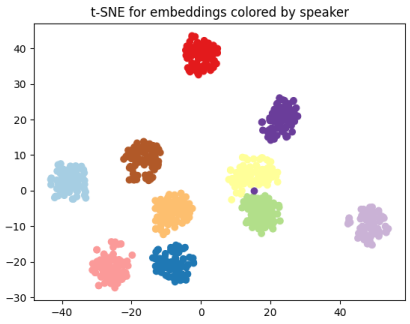


Figure 3.4: 2D visualization of the pre-trained speaker representations using the same data for fitting and visualization

To explore if such a transformation can be applied to different speaker representations, we train a UMAP [MHSG18] transformation on the training part of the dataset and use this mapping for the test representations used above. Since the t-SNE implementation from scikit-learn package [PVG+11] does not allow the application of trained transformation to new data points, UMAP is used as an alternative. As shown in figure 3.5, the resulting transformation does not generalize the desired properties to the new speakers as well as we would need for the speaker separability with multiple overlaps and less dense clusters.

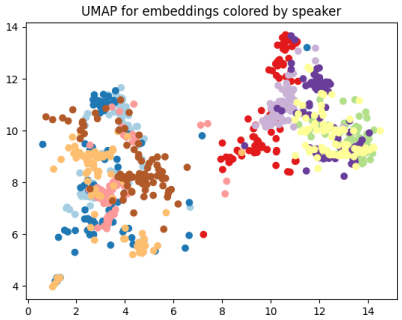


Figure 3.5: 2D visualization of the pre-trained speaker representations using different data for fitting and visualization

Hence, it cannot be guaranteed that a good scoring function exists that improves the performance without modifying speaker representations.

3.5.2 Embedding Space

One of the most convenient ways to skew embedding space to make it less dependent on emotion variability is to apply a transformation on top of the generated representations. This transformation can be trained using various losses. However, since we already have relatively good representations, the goal is to bring representations of the same speakers closer to each other and push representations of different speakers further apart. For this reason, we use contrastive learning, which attracts representations of the same speaker to each other and pushes representations of different speakers away.

For this task, we use contrastive loss [KTW+21] with a projector consisting of a couple of linear layers that do not change dimensions and the first one is followed by batch normalization and ReLU for 10 epochs with a learning rate of 10^{-3}. The parameters are the same as in [TLD+22].

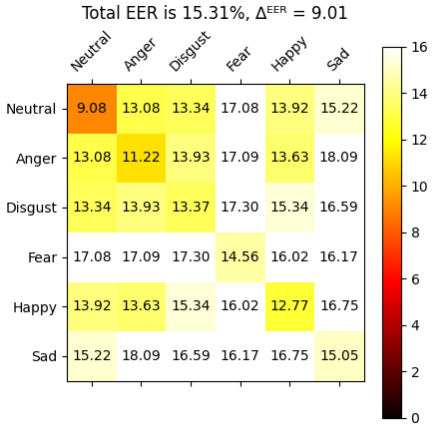


Figure 3.6: EERs matrix of the speaker representations mapped with contrastive learning

The contrastive loss is conceptually similar to the AAM-Softmax loss, but instead of calculating cosine similarity between weights and features to classify speakers, it calculates cosine similarity between embeddings of utterances of the same speakers to bring these embeddings closer together. A batch of size $N = 300$ is randomly sampled, and the loss is calculated for each positive pair against all the negative pairs [CKNH20], averaging them

obtaining the loss function:

$$l_{i,j} = -\log \frac{\exp(\cos(f(x_i), f(x_j)))}{\sum_{k=1}^{N} \mathbb{1}_{s_i \neq s_j} \exp(\cos(f(x_i), f(x_j)))}, \tag{3.1}$$

$$L_{\text{cl}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j, s_i = s_j} l_{i,j}, \tag{3.2}$$

where f is the projector mapping, s_i is the speaker ID of i-th utterance. However, this method did not preform well, as shown in figure 3.6. This outcome may be due to contrastive learning typically being used for pre-training and requiring more data to obtain reliable results.

3.5.3 Scoring Function

For the scoring function, we revisit the following traditional ASV scoring functions that were used for the previous generations of ASV systems, but have been gradually replaced by cosine similarity due to the rise of AAM-Softmax loss:

- LDA, used as an embedding mapping and then fed to cosine similarity.
- PLDA, a standard scoring function for previous ASV generation.
- Spherical PLDA, a recent modification of PLDA.

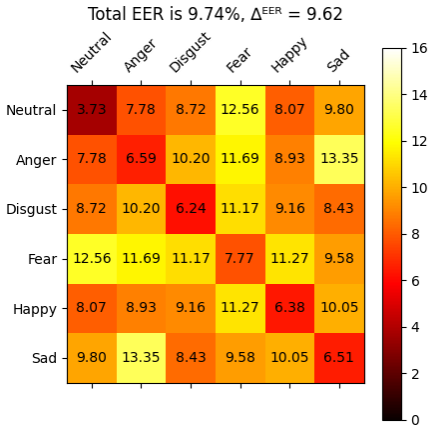


Figure 3.7: EERs matrix of the speaker representations obtained by LDA

In the training, we have 81 speakers, so we use the maximum possible number of LDA components (80) and an eigenvalue decomposition solver with automatic shrinkage. For the LDA-based EERs matrix, see figure 3.7.

We observe a slight improvement in the overall EER as well as a reduction in the difference between the pairwise emotional EERs.

For the applicability of the PLDA (using speechbrain implementation [RPP+21]), we employ LDA to remove colinearity in the embedding features. Additionally, PLDA uses speaker labels during the training, and the rank of the between-class covariance matrix is set to 80. Although PLDA yields slightly better EERs than cosine similarity (in figure 3.8), the result is worse than for the LDA in terms of both total EER and Δ^{EER}.

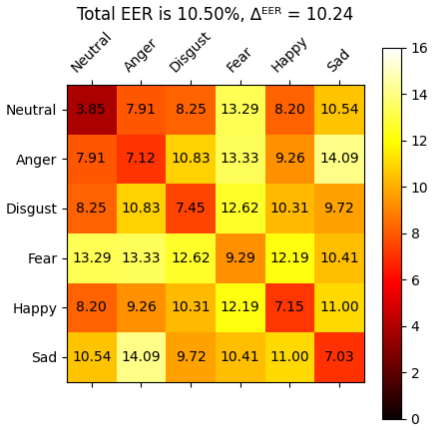


Figure 3.8: EERs matrix of the pre-trained speaker representations with PLDA-based scoring function

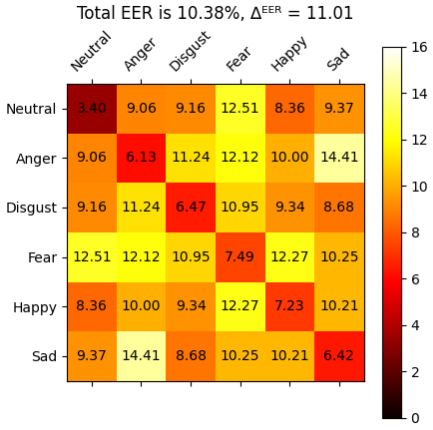


Figure 3.9: EERs matrix of the pre-trained speaker representations with SphPLDA-based scoring function

Finally, we use Spherical PLDA, a PLDA constrained by making between- and within-covariance matrices spherical (i.e., proportional to the identity matrix). This method is more stable than the cosine distance and is claimed to improve performance in a multi-enrollment setup. As shown in figure 3.9, this method yields results similar to both cosine similarity, with slightly reduced variance in cross-emotion comparisons, and standard PLDA, with slightly better total EER.

3.5.4 Fine-tuning

Next, we fine-tune the ECAPA-TDNN layers on CREMA-D.

Initially, we fine-tuned the ECAPA-TDNN layers on the CREMA-D dataset using hyper-parameters similar to those in the final training described in [CWC+22], but without employing the large-margin fine-tuning strategy [TDD21]. Specifically, we used an AAM margin of 0.2 and a scale of 30, training for 2 epochs with a batch size of 128 and a constant learning rate of 5×10^{-5}. Before fine-tuning, we initialized the weights of the AAM-Softmax for 81 target speakers by averaging the speaker representations of each speaker in the training set. To ensure stability, we performed a warm-up phase by freezing the ECAPA-TDNN layers and training only the last layer for 3 epochs. As shown in figure 3.10, this method significantly improved performance compared to all previous approaches.

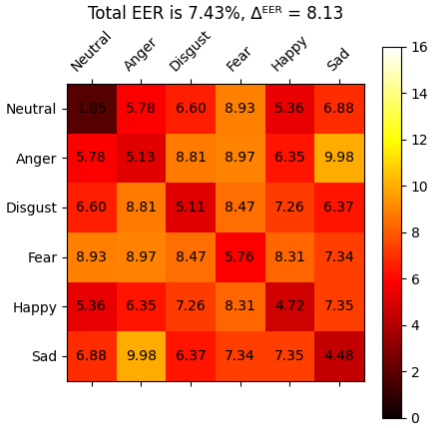


Figure 3.10: EERs matrix of the speaker representations fine-tuned with L_{AAM}

The first enhancement approach is to add Barlow Twins (BT) objective. For this task, we require a concurrent batch of the utterances in addition to simple batches, used to calculate the cross-correlation matrix. We apply AAM-Softmax loss L_{AAM} to both batches. In the original work [ZJM+21],

an extra projector network is employed, but in our case we use a pre-trained encoder, raising the challenge of initializing the projector and selecting a proper size for it. Given this, we do not use a projector network and calculate cross-correlation directly on encoder outputs, similar to the approach in [RGHN+23]. Selecting the second batch remains an open task as it can be selected in multiple ways. In our study, we evaluate the following options:

(a) The same utterances with different augmentations as in the original implementation (figure 3.11a).

(b) A random utterance of the same speaker (figure 3.11b).

(c) A random neutral utterance of the same speaker (figure 3.11c).

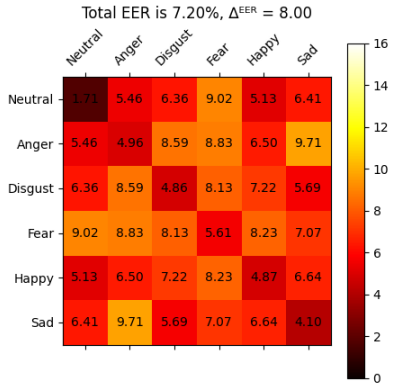
(d) A neutral utterance of the same speaker with the same spoken text, thanks to the structure of the CREMA-D dataset (see figure 3.11d).

It is important to note that the primary expectation is not a direct improvement in the ASV performance. Rather, we anticipate that it will remove redundancy in the dimensions of the representations and help prevent overfitting. However, all of the approaches still improve the ASV performance with the best results achieved by selecting the neutral version of the same utterance by the same speaker, as shown in figure 3.11.

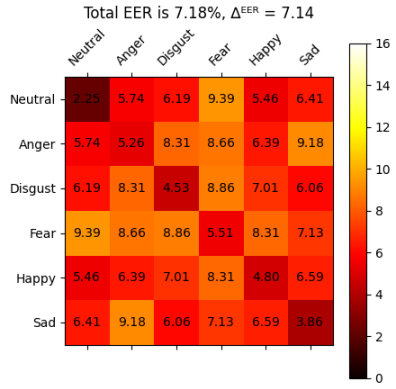
Next, inspired by [THX24], we add cosine loss between the representations of the previously used batches:

$$L_{\cos} = -\sum_{i=1}^{N} \frac{f(x_i^1) \cdot f(x_i^2)}{||f(x_i^1)|| \cdot ||f(x_i^2)||},$$

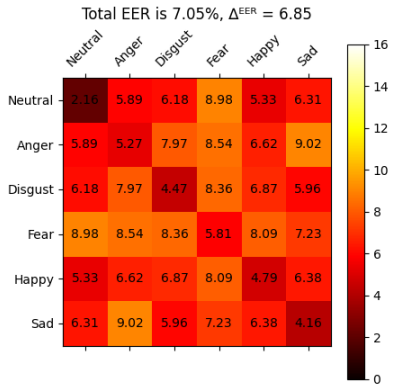
where x_i^j is the i-th element of j-th batch from Barlow Twins and f is the encoder used to obtain speaker representations (WavLM + ECAPA-TDNN). This loss explicitly targets the convergence of emotional and neutral representations. Additionally, as the CREMA-D dataset contains a limited number of spoken sentences, we employ CopyPaste augmentation, which randomly selects an audio sample of the same speaker (possibly with a different emotion) and appends it either to the start or end of the original audio. This augmentation reduces potential text dependency in the trained system and further encourages emotion-independent representations by mixing different emotions within the same utterance. After adding these improvements, we obtain the results shown in figure 3.12.



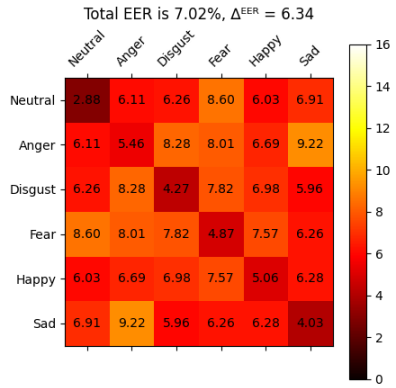
(a) Same batch



(b) Random utterances of the same speakers



(c) Random neutral utterances of the same speakers



(d) Neutral utterances of the same speakers with the same sentences

Figure 3.11: EERs matrices of the fine-tuned speaker representations with the Barlow Twins objective with a different selection of the batch for cross-correlation

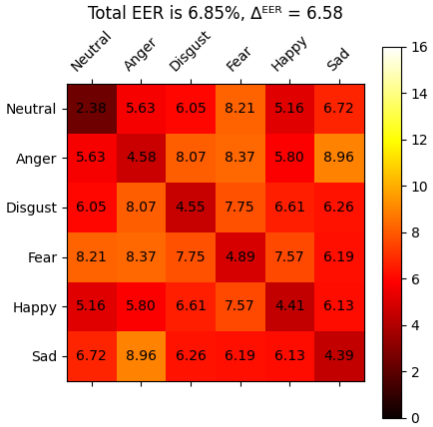


Figure 3.12: EERs matrix of fine-tuned speaker representations with the Barlow Twins objective, cosine loss, and CopyPaste augmentation

Finally, in order to address the limited number of speakers compared to traditional speaker verification datasets, we extend the dataset by augmenting new speakers with a pitch shift of 6 semitones. This approach yields the results shown in figure 3.13.

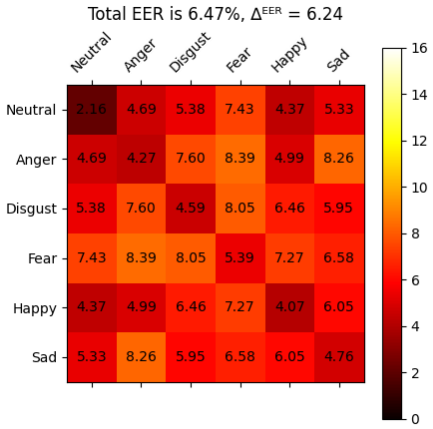


Figure 3.13: EERs matrix of fine-tuned speaker representations with the BT objective, cosine loss, CopyPaste augmentation, and pitch shift extension

3.5.5 VoxCeleb1 Performance

The experiments above show the performance of the systems trained on the CREMA-D dataset. However, it is also important to look at the performance on the test set of the original dataset used during the initial training of

ECAPA-TDNN – VoxCeleb1 [NCZ17] – to assess how much the original performance degrades on this data. The performance of the fine-tuned systems on this data is presented in table 3.1.

System	VoxCeleb1 EER
Original	0.63%
Simple Fine-tune	0.96%
Barlow Twins	1.12%
BT + Cosine + CopyPaste	1.10%
Previous + Pitch shift	1.54%

Table 3.1: EERs on VoxCeleb1 test for each model trained on CREMA-D

3.5.6 Ablation Study

To isolate the contribution of each method, we performed an ablation study on the final model trained on CREMA-D using the modified Barlow Twins objective, CopyPaste augmentation, cosine loss, and pitch shift dataset extension.

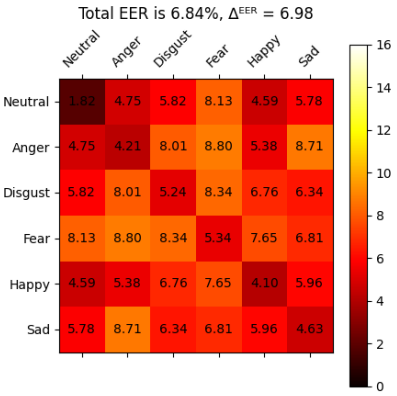
The EERs matrices for ablated features above are presented in figure 3.14.

As shown in these matrices, all of the applied methods significantly influence system performance. Loss functions have the most impact, while augmentations contribute less on the test performance, although, adding to the robustness of the trained system.

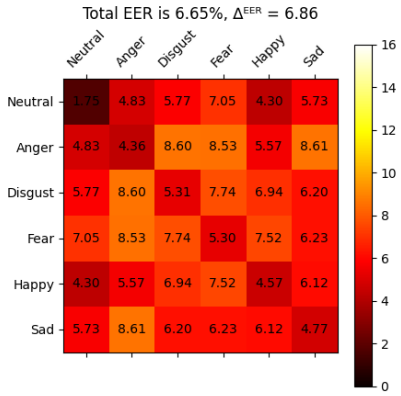
3.5.7 Out-Of-Domain Evaluation

To assess the generalization ability of the trained model, evaluate it on the RAVDESS dataset.

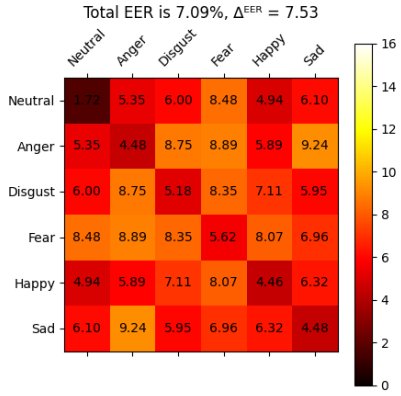
First, the EERs matrix on pre-trained WavLM + ECAPA-TDNN, shown in figure 3.15, indicated that pre-trained speaker representation perform slightly better on RAVDESS compared to CREMA-D, especially for neutral speech. However, it it also exhibits a greater disparity in emotional mismatch than the CREMA-D.



(a) Ablated Barlow Twins objective



(b) Ablated CopyPaste augmentation



(c) Ablated cosine loss

Figure 3.14: EERs matrices of the ablation study

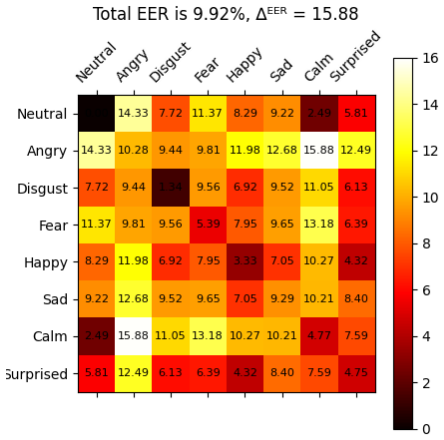


Figure 3.15: EERs matrix of the pre-trained RAVDESS speaker representations

Next, we evaluate speaker representations generated from the fine-tuned system on CREMA-D (figure 3.16). The performance improves in this case compared to the pre-trained system, although the improvement is less significant than for CREMA-D itself.

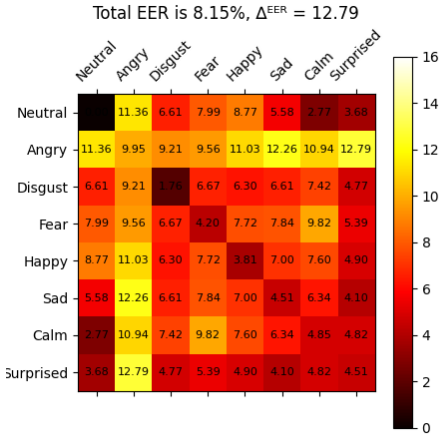


Figure 3.16: EERs matrix of the RAVDESS speaker representations by the fine-tuned system on CREMA-D

To gain a complete understanding of the cross-dataset performance, we also fine-tune the system on the RAVDESS dataset only and on a combined emotional dataset consisting of both CREMA-D and RAVDESS datasets.

For the first experiment, training on RAVDESS only, as shown in figures 3.17 and 3.18, yields a symmetric situation to the previous one: performance

on RAVDESS improves, however, CREMA-D performance worsens.

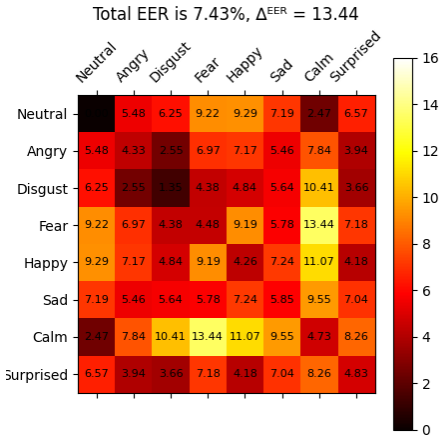


Figure 3.17: EERs matrix of the RAVDESS speaker representations by the fine-tuned system on RAVDESS

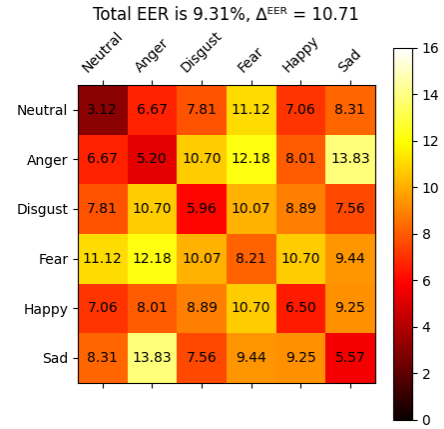


Figure 3.18: EERs matrix of the CREMA-D speaker representations by the fine-tuned system on RAVDESS

Finally, we fine-tune the system on both datasets. As shown in figures 3.19 and 3.20, cross-dataset performance improves, but does not match the results of dataset-specific training.

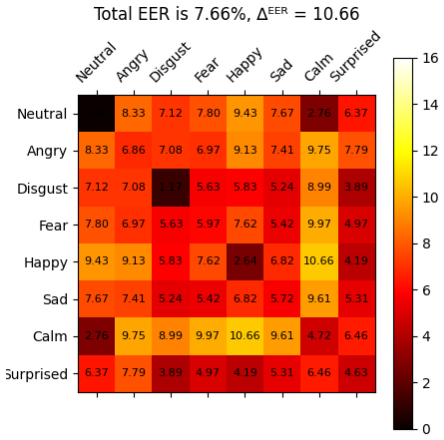


Figure 3.19: EERs matrix of the RAVDESS speaker representations by the fine-tuned system on both datasets

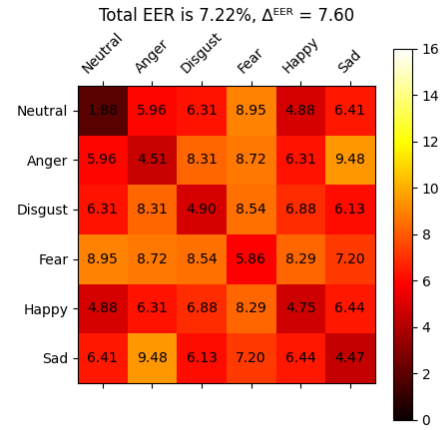


Figure 3.20: EERs matrix of the CREMA-D speaker representations by the fine-tuned system on both datasets

3.6 Analysis

Here we will discuss the results obtained in previous experiments.

Metric-based methods and contrastive learning did not sufficiently improve the performance of the pre-trained ASV system. This may be due to the pre-trained speaker representations not containing enough information to disentangle speakers.

On the other hand, fine-tuning-based methods improved the ASV performance significantly. As mentioned earlier, the most important metrics

System	Δ^{EER}	Total EER
Original	12.00	10.77%
Simple Fine-tune	8.13	7.43%
Barlow Twins	6.34	7.02%
BT + Cosine + CopyPaste	6.58	6.85%
Previous + Pitch shift	**6.24**	**6.47%**
Ablation BT	6.98	6.84%
Ablation CopyPaste	6.86	6.65%
Ablation Cosine	7.53	7.09%

Table 3.2: Δ^{EER} for each model fine-tuned on CREMA-D

for our research question are total EER over emotion-independent trials, individual EERs per pair of emotions, and the difference between the maximum and minimum of these EERs (Δ^{EER}). The total EER and Δ^{EER} for fine-tuning-based methods are summarized in table 3.2.

These results show that all modifications to the fine-tuning procedure positively influence the overall system performance. All the studied features including (modified) Barlow Twins, Cosine loss, CopyPaste augmentation, and pitch shift dataset extension improve both total EER and Δ^{EER}. The ablation study results indicate that the loss function modifications give the largest improvement: cosine loss improves Δ^{EER} from 7.53 to 6.24 and total EER from 7.09% to 6.47% giving a relative improvement of 17.1% and 8.7% respectively; Barlow Twins objective improves the Δ^{EER} from 6.98 to 6.24 and total EER from 6.84% to 6.47% with a relative improvement of 10.6% and 5.4% respectively. On the other hand, data modifications introduce less significant improvements: CopyPaste augmentation gives a relative improvement of 9% in Δ^{EER} and 2.7% and pitch shift dataset extension gives a relative improvement of 5.2% and 5.5% for Δ^{EER} and total EER, respectively. However, this does not pose a problem as CopyPaste is mainly expected to increase robustness to emotions and prevent overfitting, with its improvement in Δ^{EER} is comparable to that of loss function modifications. Pitch shift, alternatively, is expected to increase robustness in general as it introduces more speakers in the training set, yielding an improvement in the total EER similar to that of the Barlow Twins.

After conducting a listening test of the incorrect predictions in trials of the best model (with a threshold at the EER point), it is evident that a significant amount of the ASV errors are made in cases when there is an extreme display of emotion, making it difficult even for a human to determine whether it is the same speaker or not.

It is important to note that while the performance of the systems improved

on CREMA-D, its performance dropped on the original VoxCeleb1 test set (table 3.1). The most significant decrease in performance is seen in the best-performing system on CREMA-D which includes pitch shift dataset extension. However, this performance drop is by less than 1% which is still acceptable and is less significant than the improvement in the emotional data.

In out-of-domain experiments, the systems performed as expected. Although, the cross-dataset evaluations yielded worse results than matching datasets, the performance was better than that of the pre-trained system.

Chapter 4

Conclusions and Future Work

4.1 Conclusions

This research delved into enhancing the emotion-robustness of ASV systems using state-of-the-art models and various datasets. By leveraging the CREMA-D and RAVDESS datasets, we were able to extensively fine-tune and test the WavLM-Large-based ASV system, employing ECAPA-TDNN for generating speaker representations.

The initial evaluation highlighted the challenge posed by emotional speech, which degrades the ASV system's performance. Through a series of experiments, we explored multiple strategies to mitigate this issue, including altering similarity metrics, modifying embedding space, and implementing various fine-tuning techniques.

The key findings are:

1. Similarity measures and embedding space.
    - Substituting cosine similarity with LDA, PLDA, and spherical PLDA resulted in insignificant improvement.
    - Contrastive learning for embedding space modification did not yield any improvements.
2. Fine-tuning.
    - Simple fine-tuning of ECAPA-TDNN on CREMA-D substantially enhanced performance.
    - Incorporating a modification of the Barlow Twins objective and cosine loss further reduced embedding redundancy and improved robustness against emotional variability.
    - Augmenting training with CopyPaste and pitch shift techniques provided additional resilience, though with a lesser impact compared to loss functions.

3. VoxCeleb1 Evaluation. Although the methods above improved the performance on emotional speech, they also introduced a minor decrease in performance on the original VoxCeleb1 dataset, highlighting a trade-off between emotion robustness and general performance.

4. Out-of-Domain Evaluation.

    - Testing on the RAVDESS dataset confirmed the improved robustness of the fine-tuned model, though the cross-dataset performance was naturally lower compared to within-dataset evaluations.
    - Training on a combined dataset of CREMA-D and RAVDESS yielded better generalization, though the optimal performance remained dataset-specific.

5. Ablation Study.

    - The ablation study underscored the critical role of the modified Barlow Twins objective and cosine loss in enhancing system performance.
    - CopyPaste and pitch shift augmentations, while beneficial, had a more moderate impact compared to the loss functions.

In summary, the study demonstrates that fine-tuning ASV systems with targeted objectives and augmentations can significantly improve their resilience to emotional speech. While there is still room for improvement, particularly in cross-dataset scenarios, the methodologies explored provide a solid foundation for developing more emotion-robust speaker verification systems.

4.2 Future work

Several paths can be explored to further enhance the robustness and generalization of ASV systems in the presence of emotional variability:

- Expanding the dataset to include a broader range of speakers and including real-world emotional speech data, rather than acted emotions, could provide a more realistic training scenario and improve system performance in practical applications.
- Employing multi-task learning where the system simultaneously learns to perform speaker verification and emotion recognition. By jointly optimizing these tasks, the model may better disentangle speaker identity from emotional content, leading to improved robustness.
- Developing scoring functions that are emotion-aware, where the scoring mechanism adjusts based on detected emotions, could involve dynamically weighting embeddings or scores based on the emotional state of the speaker.

- Extending the evaluation to include cross-language and cross-cultural datasets to understand how emotional expressions in different languages and cultures affect ASV performance, aiding in building more universally robust systems.

By pursuing these directions, future research can contribute to the development of ASV systems that are more robust, accurate, and capable of performing well in diverse and emotionally rich real-world scenarios.

Bibliography

[CKNH20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020.

[CWC+22] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[DGXZ19] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019.

[DKD+11] Najim Dehak, Patrick Kenny, R. Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19:788 – 798, 06 2011.

[DTD20] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proc. Interspeech 2020*, pages 3830–3834, 2020.

[GPNFRS12] Leibny Paola García-Perera, Juan Arturo Nolazco-Flores, Bhiksha Raj, and Richard M. Stern. Optimization of the det curve in speaker verification. *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 318–323, 2012.

[HBT+21] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed.

Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021.

[Iof06] Sergey Ioffe. Probabilistic linear discriminant analysis. In Aleš Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision – ECCV 2006*, pages 531–542, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[KOD+08] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel. A study of interspeaker variability in speaker verification. *Trans. Audio, Speech and Lang. Proc.*, 16(5):980–988, jul 2008.

[KTW+21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021.

[MHSG18] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

[NCZ17] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[RGHN+23] Fabian Ritter-Gutierrez, Kuan-Po Huang, Dianwen Ng, Jeremy H. M. Wong, Hung yi Lee, Eng Siong Chng, and Nancy F. Chen. Noise robust distillation of self-supervised speech models via correlation metrics, 2023.

[RPP+21] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. Speechbrain: A general-purpose speech toolkit, 2021.

[RQD00] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1):19–41, 2000.

[SCP15] David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus, 2015.

[SD20] Biswajit Dev Sarma and Rohan Kumar 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)*, pages 610–615, 2020.

[SKLC23] Alexey Sholokhov, Nikita Kuzmin, Kong Aik Lee, and Eng Siong Chng. Probabilistic back-ends for online speaker recognition and clustering, 2023.

[TDD21] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck. The idlab voxsrc-20 submission: Large margin fine-tuning and quality-aware score calibration in dnn based speaker verification. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, June 2021.

[THX24] Jingguang Tian, Xinhui Hu, and Xinkang Xu. Learning emotion-invariant speaker representations for speaker verification. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10611–10615, 2024.

[TLD+22] Ruijie Tao, Kong Aik Lee, Rohan Kumar Das, Ville Hautamäki, and Haizhou Li. Self-supervised speaker recognition with loss-gated learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6142–6146. IEEE, 2022.

[vdMH08] Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

[WHH+89] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339, 1989.

[ZJM+21] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021.