REVEALING EMOTIONAL CLUSTERS IN SPEAKER EMBEDDINGS: A CONTRASTIVE LEARNING STRATEGY FOR SPEECH EMOTION RECOGNITION

Ismail Rasim Ulgen[1], *Zongyang Du*[1], *Carlos Busso*[2], *Berrak Sisman*[1]

[1]Speech & Machine Learning (SML) Lab, The University of Texas at Dallas, USA
[2]Multimodal Signal Processing (MSP) Lab, The University of Texas at Dallas, USA

ABSTRACT

Speaker embeddings carry valuable emotion-related information, which makes them a promising resource for enhancing speech emotion recognition (SER), especially with limited labeled data. Traditionally, it has been assumed that emotion information is indirectly embedded within speaker embeddings, leading to their under-utilization. Our study reveals a direct and useful link between emotion and state-of-the-art speaker embeddings in the form of intra-speaker clusters. By conducting a thorough clustering analysis, we demonstrate that emotion information can be readily extracted from speaker embeddings. In order to leverage this information, we introduce a novel contrastive pretraining approach applied to emotion-unlabeled data for speech emotion recognition. The proposed approach involves the sampling of positive and the negative examples based on the intra-speaker clusters of speaker embeddings. The proposed strategy, which leverages extensive emotion-unlabeled data, leads to a significant improvement in SER performance, whether employed as a standalone pretraining task or integrated into a multi-task pretraining setting.

Index Terms— Speech emotion recognition, speaker embeddings, clustering, contrastive learning, multi-task learning

1. INTRODUCTION

Speech emotion recognition remains a challenging task due to its complexity and the subjective nature of emotional expression, compounded by the scarcity of labeled emotional data [1]. These factors significantly hinder the development of effective SER methods, and encourage researchers to leverage auxiliary knowledge from closely related speech tasks, such as speaker verification (SV) [2–5].

In contrast to SER, SV benefits from the availability of sufficient labeled data [6, 7]. Although the tasks of recognizing emotions from speech and verifying speakers differ in their primary objectives, they both revolve around the identification of fundamental voice attributes, including pitch, tone, and phonation patterns. Consequently, speaker verification techniques with robust performance are now being explored as promising tools for enhancing the performance of speech emotion recognition systems [2, 3, 8]

Emotion information within speaker features has been explored in various emotional speech tasks. Studies [9–11] revealed increased equal error rates in speaker verification for non-matching emotional conditions, highlighting the sensitivity of speaker features to emotional states [12]. Research by [13] demonstrated emotion-related information in speaker embeddings via autoencoder-based reconstruction analysis and emotion classification. This finding was confirmed by [8], which also performed reconstruction analysis and used speaker embeddings as SER input features. Recent works [2, 3] employed deep speaker embedding networks to transfer knowledge from speaker verification to speech emotion recognition. However, the potential of recent deep speaker embeddings like d-vector [14] and ECAPA-TDNN [15] in encoding emotional information remains an area that requires comprehensive exploration. Previous studies are limited by the assumption that emotion information is indirectly encoded within speaker embeddings and can be utilized under supervision. In this paper, we aim to explore whether emotion-related information directly resides within the speaker embedding space and find effective ways to leverage this information in SER tasks.

Self-supervised speech models such as wav2vec2.0 [16] can leverage large unlabeled speech datasets to enhance supervised SER frameworks [5, 17, 18]. However, it's important to note that these pre-training objectives were not originally designed for SER, except for [19] which incorporated audio-visual features. Additionally, existing pretraining tasks utilized in SER are frame-level tasks while speech emotion is usually formulated as an uttterance-level task. Consequently, a significant gap exists in the field, particularly in the development of an utterance-level, unsupervised pre-training strategy explicitly tailored to SER, exclusively using speech-related features, which is one of the contributions of this paper.

This paper marks the first attempt to investigate the direct accessibility of emotion-related information within state-of-the-art deep speaker embeddings. Our analysis reveals distinct intra-speaker clusters that reflect emotional states, suggesting a strong link between speaker and emotion recognition. To utilize this information, we propose a novel pretraining strategy using large-scale, emotion-unlabeled data. This approach employs contrastive learning, forming positive-negative pairs based on speaker embedding clusters, without the need for emotion labels. We apply this strategy both as the primary objective of pretraining and as an additional task for the existing pretraining methods in a multi-task

Table 1: Intra-speaker clustering evaluation for emotion classification.

	d-vector				ECAPA-TDNN			
Dataset	**NMI [0,1]** ↑	**ARI [0,1]** ↑	**Purity [0,1]** ↑	**Silhoutte [-1,1]** ↑	**NMI [0,1]** ↑	**ARI [0,1]** ↑	**Purity [0,1]** ↑	**Silhoutte [-1,1]** ↑
ESD	0.76	0.72	0.89	0.14	0.89	0.91	0.97	0.13
IEMOCAP	0.29	0.21	0.66	0.01	0.31	0.25	0.67	0.01
CREMA-D	0.43	0.39	0.63	0.07	0.36	0.27	0.57	0.04
RAVDESS	0.59	0.38	0.67	0.14	0.51	0.28	0.62	0.05

setting. Our contributions can be summarized as follows: 1) We reveal readily available emotion information within speaker embeddings; 2) We introduce a unique, utterance-level contrastive learning approach for SER, without relying on emotion labels; 3) We demonstrate that combination of pretraining tasks in a multi-task setting can further improve SER performance; and 4) Through our proposed training strategy, we enhance a very strong framework, wav2vec2.0, in terms of emotion recognition performance.

2. REVEALING EMOTION CLUSTERS IN SPEAKER EMBEDDINGS

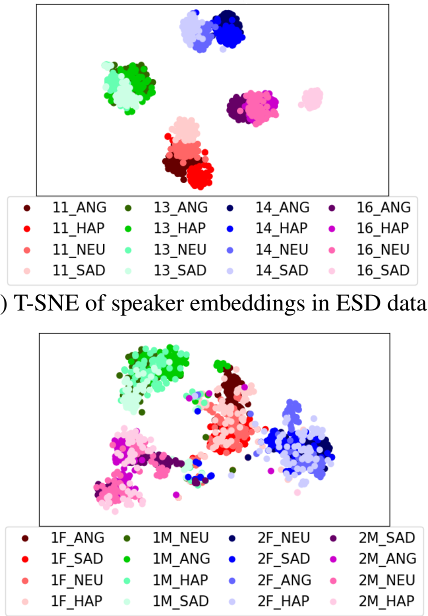
In this section, we conduct clustering analysis on speaker embeddings to explore emotion discrimination within the speaker embedding space, aiming to establish a direct link between intra-speaker clusters of embeddings and emotional categories. This connection holds significant potential for various SER applications, particularly in harnessing extensive, emotion-unlabeled data. Our analysis is driven by the hypothesis that speaker embeddings, designed to capture voice characteristics, are sensitive to variations in a speaker's voice across different emotional states [8–10, 13], drawing inspiration from studies indicating distinct speaker patterns in different emotional contexts [9–11].

2.1. Dataset, speaker embeddings and evaluation metrics

We applied k-means clustering to length-normalized speaker embeddings of using a maximum of 320 different utterances for each speaker within a dataset, using a fixed number of clusters to align with the four categorical emotions: neutral, happiness, sadness, and anger. We selected four widely used labeled emotion datasets: IEMOCAP [20], ESD [21], CREMA-D [22], and RAVDESS [23]. Our choice of deep speaker embedding networks includes d-vector [14] and ECAPA-TDNN [15], both trained with metric-based objectives like generalized end-to-end loss and angular margin softmax loss on the voxceleb2 dataset [7]. We evaluated the alignment between intra-speaker cluster labels and emotion categories using metrics such as Normalized Mutual Information (NMI) [24], Adjusted Rand Index (ARI) [24], Purity Score [25], and Silhouette Score [26], averaged over speakers and larger values indicate a stronger alignment.

2.2. Clustering and Evaluations

The clustering results are reported in Table 1. Notably, the ESD dataset consistently demonstrates exceptionally high metrics, indicating a direct alignment between intra-speaker clusters and emotion categories in specific conditions where the utterances are very clean, linguistic content is normalized over emotion categories and emotion intensity tends to be high. While the metrics for other datasets are not as high as in ESD, a meaningful correlation exists across all datasets. The IEMOCAP dataset, with challenges like reverberation and overlapping speech, exhibits the lowest metrics, possibly due to variance introduced into speaker embeddings.



(a) T-SNE of speaker embeddings in ESD dataset

(b) T-SNE of speaker embeddings in IEMOCAP dataset

Fig. 1: Visualization of intra-speaker clusters in two datasets, the colors represent {speaker id}_{emotion}.

The distribution of embeddings can be observed in the t-SNE plots in Figure 1, showing clear separation in the ESD dataset and some distinction in the IEMOCAP dataset. We've plotted t-SNE plots only for ESD and IEMOCAP due to similar trends in other databases. NMI values tend to be higher than ARI values, indicating uneven clustering errors. Higher purity values, compared to lower ARI values, suggest overlaps between specific emotion pairs, hinting at unique relationships between emotion categories. Low silhouette scores are expected due to closely spaced embeddings, aligning with their original goal of grouping speaker utterances together.

In general, the clustering results validate that speaker embeddings tend to group together for different emotional states in the embedding space due to distinct vocal characteristics for each emotion. The correspondence between emotion categories and intra-speaker clusters is limited in non-ideal conditions possibly due to other factors affecting the speech signal. The results show that even clusters with limited accuracy can serve as effective learning tasks [27–29]. Inspired by these findings, we propose a contrastive learning strategy based on the trend of intra-speaker clustering of emotion categories.

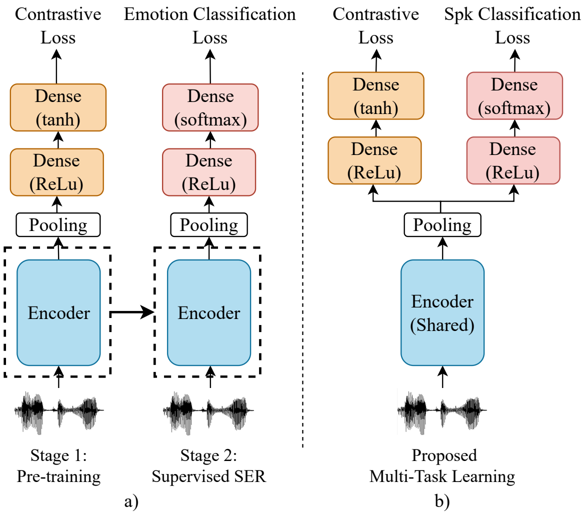


Fig. 2: a) Proposed contrastive pre-training and SER training, b) Proposed multi-task learning framework.

3. CONTRASTIVE LEARNING FOR SER

In this study, we introduce a novel contrastive pretraining strategy without emotion labels, which capitalizes on emotion-related information present in the form of intra-speaker clusters within speaker embeddings. Our approach is based on contrastive learning, a technique well-known for its efficacy across various tasks [30, 31]. The learning objective tries to maximize the similarity between positive pairs while minimize it for negative pairs. In our approach, positive pairs consist of utterances sampled from the same intra-speaker cluster, likely sharing the same emotion category. In contrast, negative examples are created from different intra-speaker clusters of the same speaker, likely to have different emotion categories given our analysis in Section 2. This setup inherently fosters an utterance-level emotion classification.

3.1. Contrastive Pretraining

In the pretraining stage, we obtain intra-speaker clusters of speaker embeddings in a separate process similar to the experiments in Section 2.1, where the only difference is in the number of clusters N since we don't have a prior about categories on emotion-unlabeled data. A variant NT-Xent [30] loss is used as an objective in the training:

$$l = -log\frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{N/2} exp(sim(z_i, z_k)/\tau)_{[k\neq i]}} \qquad (1)$$

where z_i, z_j is the positive pair and z_i, z_k are the negative pairs for a given utterance. The similarity function $sim(x, y) = x^T y/||x||.||y||$ calculates the cosine similarity and τ denotes the temperature parameter.

Soft-sampling: For each utterance, we select one positive and $N/2$ negative utterances based on intra-speaker cluster labels. Due to rough clustering, when sampling the negative examples, we employ a soft-sampling strategy, selecting one negative sample from each of the $N/2$ intra-speaker clusters that are farthest from the positive cluster center. The model architecture consists of an encoder followed by a contrastive learning head, as shown in Fig.2(a) and Fig. 3.

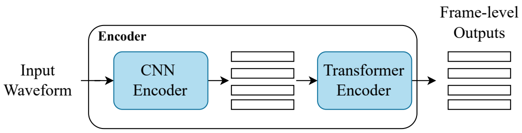


Fig. 3: The encoder architecture utilized in the networks.

3.2. Contrastive Pretraining for Multi-Task Design

Given the success of the transfer learning from speaker recognition to SER due to their connection, we also propose a multi-task learning (MTL) strategy to utilize available speaker labels. The proposed multi-task framework includes shared encoder layers along with two separate heads: contrastive learning and speaker classification head which can be seen in Figure 2(b). The contrastive learning head is trained with the proposed objective in Section 3.1; while the speaker classification head is trained with the cross-entropy loss with speaker labels. Along with the multi-task framework, speaker adversarial setting is also experimented, by including a Gradient Reversal Layer (GRL) just before the speaker classification head.

3.3. Speech Emotion Recognition

After pretraining on a large-scale, emotion-unlabeled dataset, the model is trained in a supervised manner on a smaller dataset with categorical emotion labels. During supervised training, we introduce a freshly initialized classification head on top of pre-trained encoder layers. This classification head comprises an average pooling layer, a dense projection layer with rectified linear unit (ReLU) activation, and a dense output layer with softmax activation. In this stage, we fine-tune the pre-trained layers in conjunction with the classification head, utilizing cross-entropy loss and emotion labels. The diagram can be seen in the Figure 2(a).

4. EXPERIMENTS

In this section, we report the effect of our proposed pretraining strategies with only contrastive loss and multi-task learning on SER performance when dealing with a limited amount of labeled data. We have evaluated our strategies independently and in conjunction with wav2vec2.0 to clearly discern their effect on emotion recognition performance.

4.1. Experimental setup

Datasets: During pretraining, we utilize voxceleb2 [7] as an emotion-unlabeled dataset, known for its diverse emotional contexts [13], aligning with our intra-speaker clustering approach. In supervised SER training, we separately employ two labeled emotion datasets, IEMOCAP and CREMA-D. We focus exclusively on *Anger, Happiness, Neutral, Sadness*, establishing a speaker-independent emotion recognition scenario. For the IEMOCAP corpus, we only use improvised utterances and create 5-fold training and test splits following the leave-one-session-out rule described in [17] and [2], excluding a small subset from one of the test speakers for validation. For CREMA-D, we use training data from 64 speakers, with 8 for validation and 19 for testing.

Baselines: In our basic SER experiments, we establish three baselines: *No-pretraining*, which involves initializing the model randomly before supervised SER training, without any pretraining; *No-pretraining (small)*, which has a smaller architecture with only 2 transformer layers to assess the impact of overfitting; and *Pretraining w/ spk classification* which employs pretraining the model with encoder followed by only speaker classification head and loss, similar to the methodology in [2]. For SER experiments based on wav2vec2.0, we utilize a smaller version of the original *wav2vec2.0* as the baseline pretraining method.

Model Architecture & Training: In our pretraining and basic SER experiments, our proposed methods and baseline models, have the same encoder architecture, which is based on wav2vec2.0 [16]. This encoder architecture includes a feature extractor and a transformer encoder, similar to wav2vec2.0, but with a more compact design featuring only 6 transformer layers. The contrastive learning head includes an average pooling layer for frame-level outputs, followed by two dense layers featuring ReLU and tanh activation functions shown in Figure 2, respectively. The speaker and emotion classification heads have a similar structure as the contrastive head but use softmax activation at the output layer, shown in Figure 2. In the speaker adversarial setting, we introduce an additional GRL layer after pooling and before the speaker classification head. All the proposed models take the raw waveform of an utterance as input.

During pretraining, we segment the input utterances into 4-second intervals and perform offline intra-speaker clustering with $N = 20$. The models are pretrained for 250k steps using the AdamW optimizer with a batch size of 8. In the supervised SER training that follows, the model undergoes 30 epochs of training with a learning rate of 1e-5, stopping based on the validation accuracy. We repeat each supervised SER training 5 times with different initialization seeds and measure unweighted average recall (UAR) during the evaluation. For the SER experiments based on wav2vec2.0 reported in Table 3, the baseline *wav2vec2.0*[1] with 6 transformer layers, is pretrained for 400k steps on voxceleb2. We then fine-tune this model with our strategies on voxceleb2 for an extra 50k steps. The feature extractor and transformer layers of fine-tuned wav2vec2.0 are utilized in the supervised SER training.

4.2. Results and Discussion

According to the results in Table 2, our proposed contrastive strategy, denoted as *Pretraining w/ proposed contrastive*, demonstrate a significant improvement in SER compared to cases with no pretraining in both datasets. We note that pretraining with supervised speaker classification also leads to substantial improvements in both datasets, consistent with findings in [2]. The proposed multi-task learning approach, denoted as *Pretraining w/ proposed MTL*, leverages the inherent connection between speaker and emotion recognition, while simultaneously considering intra-speaker variations and obtains the best performance in the IEMOCAP corpus. We observe that speaker adversarial network degrades performance, indicating that trying to remove speaker information has a negative impact and supports the connection between speaker and emotion recognition. In CREMA-D, the speaker classification baseline performs exceptionally well, possibly due to the presence of normalized linguistic content, creating ideal conditions for discriminating emotions through speaker embeddings, as discussed in Section 2. Overall, these results underscore the effectiveness of our multi-task learning method and highlight the strong relationship between emotion and speaker recognition.

Table 2: SER results, in terms of mean UAR.

Pre-Training	**IEMOCAP (UAR)**	**CREMA- D (UAR)**
No pretraining (small)	*58.37*	*65.12*
No pretraining	*58.54*	*64.15*
Pretraining w/ spk classification	*67.57*	***75.23***
Pretraining w/ proposed contrastive	65.50	70.44
Pretraining w/ proposed spk ADV	64.48	66.58
Pretraining w/ proposed MTL	**69.16**	73.80

Table 3: SER results with wav2vec2.0, mean UAR.

Pre-Training	**IEMOCAP (UAR)**	**CREMA- D (UAR)**
wav2vec2.0 [16]	*72.14*	*80.78*
FT wav2vec2.0 w/ proposed contrastive	72.78	81.72
FT wav2vec2.0 w/ proposed MTL	**73.80**	**83.01**

In Table 3, baseline *wav2vec2.0* model, pretrained with voxceleb2, performs impressively well as a pretraining method for SER, underscoring its effectiveness. Fine-tuning this baseline with our contrastive learning strategy, *FT wav2vec2.0 w/ proposed contrastive*, seems leading to minor improvements in both datasets. Fine-tuning wav2vec2.0 with our proposed multi-task setting, *FT wav2vec2.0 w/ proposed MTL*, yields substantial enhancement, highlighting the effectiveness of our approach.

5. CONCLUSION

Our research reveals the potential of speaker embeddings for enhancing SER task, even with limited labeled data. Our study establishes a direct link between emotions and state-of-the-art speaker embeddings through intra-speaker clusters. Our novel contrastive pretraining approach on emotion-unlabeled datasets, based on these clusters, significantly improves SER performance, whether used alone or in multi-task settings. Our findings not only advance our understanding of speaker embeddings and emotions but also provide practical solutions for data scarcity in SER. As a future work, we intend to extend the analysis of emotion information in speaker embeddings, analyzing other factors which potentially affect the appearance of that information.

[1] https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec

6. REFERENCES

[1] Vidhyasaharan Sethu, Emily Mower Provost, Julien Epps, Carlos Busso, Nicholas Cummins, and Shrikanth S. Narayanan, "The ambiguous world of emotion representation," *ArXiv*, vol. abs/1909.00360, 2019.

[2] R. Pappagari, Tianzi Wang, Jesús Villalba, Nanxin Chen, and Najim 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.

[3] Sarala Padi, Seyed Omid Sadjadi, Dinesh Manocha, and Ram D. Sriram, "Improved speech emotion recognition using transfer learning and spectrogram augmentation," *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021.

[4] Sitong Zhou and Homayoon S. M. Beigi, "A transfer learning method for speech emotion recognition from automatic speech recognition," *ArXiv*, vol. abs/2008.02863, 2020.

[5] Leonardo Pepino, Pablo Riera, and Luciana Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404.

[6] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech*, 2017.

[7] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech*, 2018.

[8] Zakaria Aldeneh and Emily Mower Provost, "You're not you when you're angry: Robust emotion features emerge by recognizing speakers," *IEEE Transactions on Affective Computing*, vol. 14, pp. 1351–1362, 2023.

[9] Srinivas Parthasarathy, Chunlei Zhang, John H. L. Hansen, and Carlos Busso, "A study of speaker verification performance with expressive speech," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5540–5544, 2017.

[10] Srinivas Parthasarathy and Carlos Busso, "Predicting speaker recognition reliability by considering emotional content," *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 434–439, 2017.

[11] Michelle I Bancroft, Reza Lotfian, John H. L. Hansen, and Carlos Busso, "Exploring the intersection between speaker verification and emotion recognition," *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 337–342, 2019.

[12] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 788–798, 2011.

[13] Jennifer Williams and Simon King, "Disentangling style factors from speaker representations," in *Interspeech*, 2019.

[14] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez-Moreno, "Generalized end-to-end loss for speaker verification," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4879–4883, 2017.

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

[16] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020, NIPS'20, Curran Associates Inc.

[17] Edmilson da Silva Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz, "Speech emotion recognition using self-supervised features," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6922–6926, 2022.

[18] Mao Li, Bo Yang, Joshua Levy, Andreas Stolcke, Viktor Rozgic, Spyros Matsoukas, Constantinos Papayiannis, Daniel Bone, and Chao Wang, "Contrastive unsupervised learning for speech emotion recognition," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6329–6333, 2021.

[19] Lucas Goncalves and Carlos Busso, "Improving Speech Emotion Recognition Using Self-Supervised Learning with Domain-Specific Audiovisual Tasks," in *Proc. Interspeech 2022*, 2022, pp. 1168–1172.

[20] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[21] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[22] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[23] Steven R Livingstone and Frank A Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, pp. e0196391, 2018.

[24] Xuan Vinh Nguyen, Julien Epps, and James Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, 2010.

[25] Eréndira Rendón, Itzel M. Abundez, Citlalih Gutierrez, Sergio Díaz Zagal, Alejandra Arizmendi, Elvia M. Quiroz, and H. Elsa Arzate, "A comparison of internal and external cluster validation indexes," in *Proceedings of the 2011 American Conference on Applied Mathematics and the 5th WSEAS International Conference on Computer Engineering and Applications*, Stevens Point, Wisconsin, USA, 2011, AMERICAN-MATH'11/CEA'11, p. 158–163, World Scientific and Engineering Academy and Society (WSEAS).

[26] Peter J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[27] 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," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[28] Bing Han, Zhengyang Chen, and Yanmin Qian, "Self-supervised learning with cluster-aware-dino for high-performance robust speaker verification," *ArXiv*, vol. abs/2304.05754, 2023.

[29] Ruijie Tao, Kong-Aik Lee, Rohan Kumar Das, Ville Hautamaki, and Haizhou Li, "Self-supervised speaker recognition with loss-gated learning," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6142–6146, 2021.

[30] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*. 2020, ICML'20, JMLR.org.

[31] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick, "Momentum contrast for unsupervised visual representation learning," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2019.