HYFuse: Aligning Heterogeneous Speech Pre-Trained Representations in Hyperbolic Space for Speech Emotion Recognition

Orchid Chetia Phukan[*1], *Girish*[* 1,2], *Mohd Mujtaba Akhtar*[*1,3], *Swarup Ranjan Behera*[4], *Pailla Balakrishna Reddy*[5], *Arun Balaji Buduru*[1], *Rajesh Sharma*[6,7]

[1]IIIT-Delhi, India, [2]UPES, India, [3]V.B.S.P.U, India, [4]Independent Researcher, India, [5]Reliance AI, India, [6]University of Tartu, Estonia, [7]Plaksha University, India

Correspondence: orchidp@iiitd.ac.in

Abstract

Compression-based representations (CBRs) from neural audio codecs such as EnCodec capture intricate acoustic features like pitch and timbre, while representation-learning-based representations (RLRs) from pre-trained models trained for speech representation learning such as WavLM encode high-level semantic and prosodic information. Previous research on Speech Emotion Recognition (SER) has explored both, however, fusion of CBRs and RLRs haven't been explored yet. In this study, we solve this gap and investigate the fusion of RLRs and CBRs and hypothesize they will be more effective by providing complementary information. To this end, we propose, **HYFuse**, a novel framework that fuses the representations by transforming them to hyperbolic space. With **HYFuse**, through fusion of x-vector (RLR) and Soundstream (CBR), we achieve the top performance in comparison to individual representations as well as the homogeneous fusion of RLRs and CBRs and report SOTA.

Index Terms: Speech Emotion Recognition, Pre-Trained Models, Neural Audio Codec, Representations

1. Introduction

Speech Emotion Recognition (SER) plays a pivotal role in human-computer interaction, enabling systems to identify and understand the emotional nuances expressed in speech [1]. By analyzing vocal attributes such as pitch, tone, and rhythm, SER systems uncover the intricate nuances of human affect. SER has far-reaching implications, from enhancing mental health monitoring in healthcare [2] to transforming educational tools by gauging student engagement [3]. Initial research in SER mostly focused on the usage of spectral features such MFCC with classical ML models such as GMM [4], SVM [5]. This was succeeded by the use of deep learning models such as LSTM [6], CNN, CNN-LSTM [7], etc.

Recent strides in SER research have seen the usage of representations from state-of-the-art (SOTA) Pre-trained models (PTMs). These representations have provided substantial performance benefit and has led to sufficient development in SER. These representations can be primarily categorized into two types: Representation-learning based representations (RLRs) derived from speech PTMs such as Wav2vec2[8], WavLM [9], XLS-R [10], etc. and compression-based representations (CBRs) extracted from neural audio codecs such as EnCodec [11], DAC [12], and Soundstream [13]. PTMs for RLRs are generally trained for speech representation learning and it can be both for a particular language or multilingual, however, neural audio codecs (NACs) are trained for compression of input data following a encoder-decoder modeling architecture. Researchers have explored various RLRs such as Wav2vec2 [14], HuBERT [15], etc. for SER. Also, usage of compression-based representations (CBRs) from NACs for SER has gained recent traction in the community. Wu et al. [16] gave a initial exploration of CBRs for SER by investigating different NACs such as Encodec, DAC, Speech Tokenizer and so on. However, they only focused on English SER. Ren et al. [17] extended to chinese SER and gave a much more comprehensive analysis of various SOTA CBRs with the inclusion of more NACs. Furthermore, Mousavi et al. [18] presented the first comparative study of CBRs and RLRs for SER. Further. Wu et al. [19] also explored the fusion of RLRs such as Wav2vec2, WavLM, Unispeech-SAT, and so on for more due to existence of complementary behavior of such representations for more improved SER. Such improvement due to the combination of PTM representations can also be seen across various related speech processing tasks such as speech recognition [20], synthetic speech detection [21].

However, no focus on the fusion of heterogeneous representations i.e. RLRs and CBRs have been given, despite extensive research into SER with PTM representations. In this work, for the first time, to the best of our knowledge, we explore such fusion of heterogenenous representations (RLRs and CBRs). *We hypothesize that fusion of RLRs and CBRs will lead to further improvement in SER performance by the exploitation of complementary information of RLRs and CBRs. CBRs captures the low-level features like pitch, timbre and RLRs encodes higher-level prosodic patterns.* To aid in effective fusion, we propose a novel framework, **HYFuse** (Fusion in Hyperbolic Space) that transforms the representations from euclidean space to hyperbolic space and performs fusion through mobius addition. As far as we know, this is the first study to investigate the usage of hyperbolic space for fusion of representations in the context of SER. The fusion of CBRs and RLRs in hyperbolic space allows for the preservation of their hierarchical relationships and complementary features, ensuring that both low-level acoustic details and high-level prosodic patterns are effectively aligned and integrated.

The key contributions of this work are:

- We propose, **HYFuse** (Figure 1), a novel framework for fusing RLRs and CBRs by transforming them into hyperbolic space and leveraging the strengths of hyperbolic geometry for effective fusion of RLRs and CBRs.
- Using **HYFuse**, with the fusion of x-vector (RLR) and SoundStream (CBR) representations, we achieve superior performance compared to individual representations and homogeneous fusions of RLRs and CBRs. Our framework sets new SOTA results on the CREMA-D and Emo-DB benchmark datasets, establishing the efficacy of combining RLRs and CBRs for SER.

* Contributed equally as a first authors.

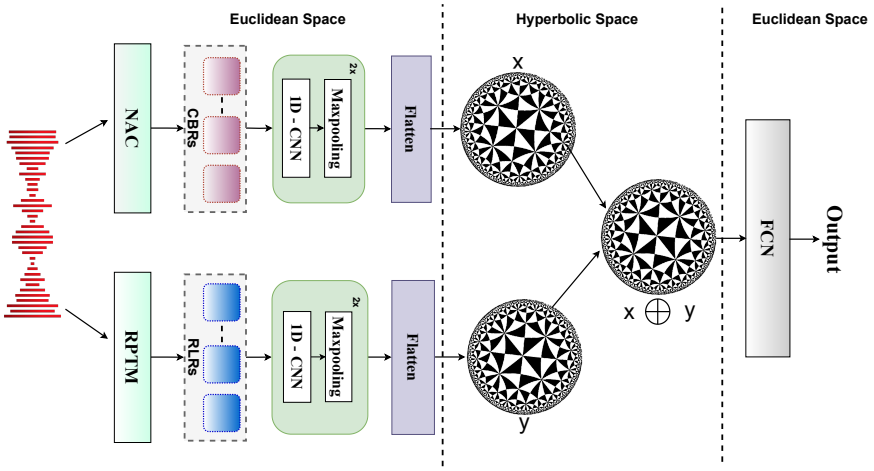


Figure 1: **HYFuse**; $x \oplus y$ *represent mobius addition*

We have released the code and models from this work at: `https://github.com/Helix-IIIT-Delhi/HYFuse-SER`

2. Pre-Trained Representations

In this section, we give a brief overview of the PTMs and NACs behind RLRs and CBRs.

2.1. Representation-learning

WavLM[1] [9] has shown SOTA performance in multiple speech processing tasks within SUPERB. We adopt its base version, which consists of 94.70 million parameters and is pre-trained on 960 hours of Librispeech. Similarly, Wav2Vec2[2] [8] is included in our study as a contrastive learning-based representation-learning based PTM. We use its base variant, which has 95.04 million parameters and is also pre-trained on 960 hours of Librispeech. We also incorporate HuBERT[3] [22], a speech PTM inspired by the BERT architecture, pre-trained on 960 hours of Librispeech. We utilize its base version, which consists of 94.68 million parameters. Additionally, we consider x-vector[4] [23], a time delay neural network designed specifically for speaker recognition, comprising 4.2 million parameters. It is particularly relevant to our research, as its representations have been shown to be effective for SER [24]. All audio recordings are resampled to 16 kHz before passing to the PTMs. The PTMs remain frozen, and we extract RLRs from their final hidden states using average pooling. The resulting feature dimensions are 768 for WavLM, Wav2Vec2, and HuBERT, while x-vector produces 512-dimensional representations.

2.2. Compression

Soundstream[5] [13] is an efficient NAC designed for low-bitrate compression, utilizing an encoder-decoder architecture with Residual Vector Quantization (RVQ) and multi-scale STFT discriminators to maintain a balance between compression and audio quality. It supports bitrates ranging from 3 kbps to 18 kbps. Descript Audio Codec (DAC)[6] [12] offers a universal approach to audio compression, achieving an impressive 90x compression rate at 8 kbps for 44.1 kHz audio. It is designed to handle a wide range of audio signals while maintaining high fidelity. Speech Tokenizer[7] [25] is a unified tokenizer for speech language models (SLMs) that employs RVQ to generate hierarchical representations capturing both linguistic and acoustic features. It demonstrates speech reconstruction quality comparable to EnCodec. EnCodec[8] [11] is a high-fidelity NAC that features a streaming encoder-decoder architecture combined with RVQ for efficient audio compression. It is designed to preserve fine-grained audio details while achieving effective compression. All input audios are resampled to 16 kHz before being processed by DAC, Soundstream, and Speech Tokenizer, while EnCodec processes audio at 24 kHz. We extract CBRs from the frozen encoders of these codecs using average pooling, resulting in feature dimensions of 256 for Soundstream, 251 for DAC, 250 for Speech Tokenizer, and 375 for EnCodec.

3. Modeling

In this section, we detail the modeling approaches with individual RLRs and CBRs as well as the proposed framework, **HYFuse** for fusion of heterogenous RLRs and CBRs. For modeling individual representations, we use fully connected network (FCN) and CNN. The CNN has two 1D convolutional layers with 64 and 128 filters and a kernel size of 3. ReLU is the activation function used in the convolutional layers. The output is then flattened and passed through a FCN block with a dense layer of 128 neurons, followed by output ayer for classification which uses softmax as activation function. For the FCN model, we use same the modeling as the FCN block in the CNN.

3.1. HYFuse

We propose, **HYFuse** for the effective fusion of RLRs and CBRs. The architecture is presented in Figure 1. **HYFuse** leverages hyperbolic geometry to fuse CBRs and RLRs while preserving their hierarchical relationships. The fusion in hyperbolic space complements both fine-grained acoustic details and high-level

[1] `https://huggingface.co/microsoft/wavlm-base`
[2] `https://huggingface.co/facebook/wav2vec2-base`
[3] `https://huggingface.co/facebook/hubert-base-ls960`
[4] `https://huggingface.co/speechbrain/spkrec-xvect-voxceleb`
[5] `https://github.com/haydenshively/SoundStream`
[6] `https://huggingface.co/descript/dac_16khz`
[7] `https://github.com/ZhangXInFD/SpeechTokenizer.git`
[8] `https://huggingface.co/facebook/encodec_24khz`

Rep	CREMA-D				Emo-DB			
	FCN		CNN		FCN		CNN	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
RLRs								
W2	58.66	55.14	65.16	65.08	88.21	86.42	**91.51**	**90.65**
W	64.71	61.63	68.81	68.64	87.23	85.21	89.72	89.52
XE	63.69	61.20	68.77	68.67	85.65	85.25	81.31	80.61
H	67.85	66.25	70.63	70.45	86.25	85.39	88.81	87.85
CBRs								
E	47.85	46.98	48.48	42.56	47.65	45.08	48.04	40.20
D	42.52	41.86	43.84	35.74	39.78	38.52	40.56	34.10
ST	41.61	40.14	48.51	45.92	45.65	43.71	49.91	37.20
SS	54.20	53.94	55.19	53.03	59.12	57.96	**61.36**	**58.96**

Table 1: *Performance Evaluation of models trained with various RLRs and CBRs; Scores are in % and average of five folds; Acc and F1 stands for accuracy and marco-average F1 score; The abbreviations given are: Wav2vec2 (W2), WavLM (W), x-vector (XE), HuBERT (H), EnCodec (E), DAC (D), Speech Tokenizer (ST), Soundstream (SS); The abbreviations used herre are kept same for Table 2*

prosodic structures from the RLRs and CBRs respectively, resulting in more expressive and structured feature representations. Unlike euclidean fusion methods, which may distort the intrinsic organization of representations, hyperbolic fusion maintains relative distances and ensures that complementary features are optimally integrated. Detailed walkthrough of **HYFuse** is given as follows. The representations are first passed through 1D convolutional layers with the same architecture as used for modeling individual RLRs and CBRs. The features are then flattened and ready to be transformed to hyberbolic space. The transformation from euclidean to hyperbolic space is achieved using the exponential map:

$$\exp_0(x) = \begin{cases} \tanh(\kappa\|x\|)\frac{x}{\|x\|}, & \text{if } \|x\| > 0, \\ 0, & \text{if } \|x\| = 0, \end{cases} \quad (1)$$

where $\kappa > 0$ denotes the curvature of the hyperbolic space, and $\|x\|$ represents the Euclidean norm of input feature. Now, the transformed features of RLRs and CBRs will be represented as x_1 and y_1. They are then fused through the Möbius addition operation. Möbius addition between the two hyperbolic points x_1 and y_1 is given by:

$$x_1 \oplus y_1 = \frac{(1 + 2\langle x_1, y_1\rangle + \|y_1\|^2)x_1 + (1 - \|x_1\|^2)y_1}{1 + 2\langle x_1, y_1\rangle + \|x_1\|^2\|y_1\|^2}, \quad (2)$$

where $\langle x_1, y_1\rangle$ denotes the Euclidean dot product, and $\|\cdot\|^2$ represents the squared Euclidean norm. Once fused, the resultant representation y is mapped back to Euclidean space using the logarithmic map:

$$\log_0(y) = \begin{cases} 2 \cdot \text{arctanh}(\|y\|)\frac{y}{\|y\|}, & \text{if } \|y\| < 1, \\ 0, & \text{if } \|y\| = 0. \end{cases} \quad (3)$$

Ensuring that $\|y\| < 1$ maintains numerical stability within the Poincaré ball. FCN block with a dense layer is attached on top of the final fused representation followed by the output layer with softmax activation which outputs probabilities of the emotion classes. **HYFuse** trainable parameters for different input representations are from 8 to 13 millions.

4. Experiments

4.1. Dataset

CREMA-D [26]: It contains 7,442 samples from 91 actors, representing a diverse range of racial and ethnic backgrounds, including Caucasian, African American, Hispanic, and Asian participants. The dataset features 48 male and 43 female actors, aged between 20 and 74 (average age: 36), offering a broad demographic coverage. Each actor delivers 12 sentences and spans over six emotions—happy, sad, anger, fear, disgust, and neutral.

Emo-DB [27]: It comprises approximately 800 utterances recorded by 10 actors (5 male, 5 female), each performing a set of 10 carefully selected sentences expressing seven emotions: neutral, anger, fear, joy, sad, disgust, and boredom. It is a german SER corpus. Due to differences in audio duration, the NACs will produce different length representations. So as a initial preprocessing step, we pad the audios to the length of the maximum duration audio in the respective dataset for all our experiments.

Training Details: The models are trained using cross-entropy as the loss function and Adam as the optimizer. We set the batch size as 32, learning rate as 1e-5, and epochs as 50. We leverage dropout and early stopping to mitigate overfitting. We follow 5-fold cross-validation for training and validating our models where 4 folds are used as training set and 1 fold as test set.

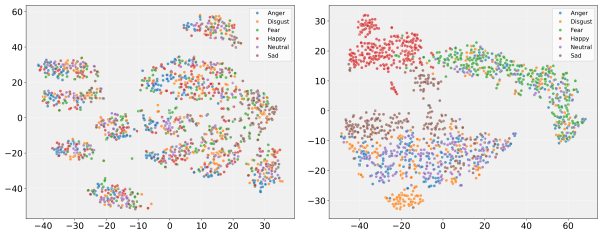
4.2. Experimental Results

We present the evaluation of downstream models trained with individual RLRs and CBRs in Table 1. We see that RLRs outperform CBRs across both datasets, indicating that CBRs struggle to capture the speech characteristics necessary for better SER. This performance is also observed across previous research evaluating RLRs and CBRs for SER [18]. Among CBRs, Soundstream shows the strongest performance in both the datasets. However, even the best-performing CBRs still lags behind the RLRs. Among RLRs, HuBERT report that top performance with CNN in CREMA-D and Wav2vec2 with CNN in Emo-DB. This mixed performance points towards the effect of downstream data distribution on the performance of the models trained with representations. Overall, we see that CNN based models shows better performance than FCN models. These scores will be considered as baselines for experiments with combinations of different representations.

Table 2 presents the results of homogeneous (RLRs + RLRs, CBRs + CBRs) and heterogeneous (RLRs + CBRs) fusions of representations. We use concatenation (Concat) based fusion as the baseline fusion technique. We follow the same architecture as **HYFuse** up to feature flattening and subsequently applying a FCN with the same modeling details as **HYFuse**. We also keep the training details same as **HYFuse** for fair comparison. Our findings reveal that **HYFuse** consistently outperforms both individual representations and concatenation-based fusion across CREMA-D and Emo-DB, reinforcing the strength of hyperbolic transformation in aligning representations. When examining homogeneous fusion, we observe that while combining two RLRs, such as Wav2vec2 and HuBERT, yields improvements over its individual performances. This improvements in performances after combinations shows some emergence of complementary behavior in the representations. Also, fusing CBRs with CBRs yields lower performance than fusion of RLRs with RLRs and this is due to inherently lower individual performance of CBRs. However, we observe a surprising criterion as the fusion of some high performing RLRs and CBRs, shows far better performance

Pairs	CREMA-D				Emo-DB			
	Concat		HYFuse		Concat		HYFuse	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
RLRs + RLRs								
W2 + W	60.98	59.65	64.58	63.38	88.63	87.36	93.36	92.27
W2 + XE	58.78	57.73	66.88	65.53	87.69	86.64	91.45	90.03
W2 + H	71.18	70.36	76.61	75.52	87.25	86.13	94.63	94.48
W + XE	69.33	68.18	74.49	73.38	88.24	87.34	92.08	91.36
W + H	65.97	65.82	77.25	76.69	89.64	88.61	94.25	93.37
XE + H	68.52	67.79	73.64	72.28	88.33	87.79	93.62	93.57
CBRs + CBRs								
E + D	58.97	47.61	64.68	63.26	58.96	52.45	64.85	63.28
E + ST	58.97	45.08	66.67	65.59	58.68	56.51	64.14	63.68
E + SS	57.10	40.07	66.48	65.52	55.69	52.38	68.73	67.78
D + ST	55.23	51.44	63.43	62.52	55.96	54.09	61.19	60.05
D + SS	61.78	60.06	66.64	65.53	52.85	41.98	65.06	58.54
ST + SS	59.63	58.46	67.64	66.68	58.61	53.28	66.62	65.53
RLRs + CBRs								
W2 + E	75.13	75.07	60.24	60.58	78.55	78.50	90.77	89.72
W2 + D	77.70	77.70	78.26	78.19	85.33	85.05	89.77	89.72
W2 + ST	72.41	71.89	74.62	74.39	84.15	83.79	91.84	91.52
W2 + SS	76.15	76.15	**79.29**	**79.10**	82.57	82.24	**95.33**	**95.18**
W + E	66.89	66.79	77.68	76.82	80.37	79.49	95.20	94.11
W + D	60.91	60.55	66.13	65.75	85.05	84.71	85.98	85.98
W + ST	76.02	72.61	75.35	75.14	83.18	83.02	95.05	95.05
W + SS	65.33	65.28	66.89	66.99	64.96	63.01	87.31	85.98
XE + E	63.64	63.47	67.85	66.28	86.03	85.98	94.39	94.14
XE + D	65.16	65.01	69.63	68.32	86.02	85.05	91.59	91.59
XE + ST	65.28	65.14	71.52	70.89	87.30	86.92	92.52	92.20
XE + SS	63.65	63.53	69.88	68.13	69.31	65.14	93.01	92.52
H + E	67.02	66.51	69.18	69.04	85.56	85.05	88.35	87.85
H + D	64.14	63.99	68.23	68.47	82.62	82.24	90.14	88.79
H + ST	68.03	68.03	67.29	67.22	83.18	82.92	86.17	85.98
H + SS	67.02	66.82	68.83	68.64	65.96	63.21	89.31	88.79

Table 2: *Performance evaluation of model trained on combination of various RLRs and CBRs: The scores are presented in % and average of five-folds*

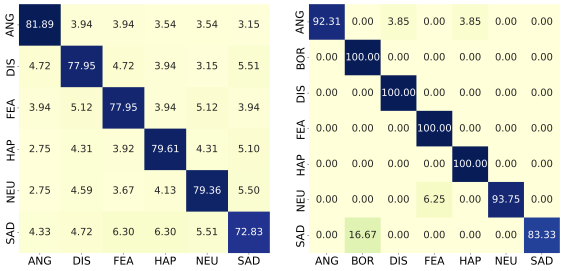


(a) (b)

Figure 2: *t-SNE visualizations: (a) CNN (HuBERT) (b)* **HYFuse (Wav2vec2 + Soundstream)**

than the homogeneous fusion of RLRs and CBRs, despite the performance of individual CBRs and fusion of CBRs with CBRs is quite low. Such fusion leads to improved performance by leveraging the complementary strengths of RLRs and CBRs, where RLRs provide rich prosodic information while CBRs capture fine-grained acoustic characteristics.

This behavior is observed across both baseline concatenation-based fusion technique and fusion with **HYFuse**. However, the fusion through **HYFuse** brings out the complementary behavior more effectively by aligning the representations in hyperbolic space, thereby preserving their hierarchical relationships and minimizing distortion. Notably, the fusion of Wav2vec2 and Soundstream with **HYFuse** emerges as the best-performing combination, surpassing all individual representations, homogeneous representations fusion, and the baseline concatenation-based fusion technique. With this performance, we achieve SOTA performance as the individual RLRs were SOTA for SER [28]. Further, this highlights the advantage of integrating heterogeneous representations for improved. This validates *our hypothesis that heterogeneous fusion of RLRs and CBRs will be most effective for SER.* Additionally, not all the combinations of RLRs with CBRs leads to improved performance over homogeneous fusion of RLRs, however, we believe that fusion of strong individual CBRs with strong individual RLRs brings out the best in them. Aside from the best-performing combination of Wav2vec2 and Soundstream, we observe that the fusion of WavLM and EnCodec also shows competitive performance, though it does not surpass the top combination. This suggests that while the integration of complementary representations enhances performance, the degree of improvement depends on the specific characteristics of the fused representations. Overall, our findings highlight the significance of selecting the right combination of RLRs and CBRs, as well as the effectiveness of **HYFuse** in leveraging their strengths to further enhance SER. We present the t-SNE plot visualization of downstream trained on HuBERT (best performing representations for CREMA-D) and **HYFuse** with the fusion of Wav2vec2 and Soundstream in Figure 2 for CREMA-D. We extract the representations from the penultimate layer of the downstream models. We observe far better clustering of the emotion classes with **HYFuse**, thus showing its effectiveness for better SER. Additionally, we also plot the confusion matrices of **HYFuse** with Wav2vec2 and Soundstream for both the datasets in Figure 3.



(a) (b)

Figure 3: *Confusion matrices: (a) CREMA-D, (b) Emo-DB; y-axis indicates the True Values, whereas the x-axis represents the Predicted Values.*

5. Conclusion

In this study, we explored the fusion of RLRs and CBRs for SER, addressing a previously unexplored research gap. While CBRs such as EnCodec capture intricate acoustic features like pitch and timbre, RLRs from pre-trained models like WavLM encode high-level prosodic information. We hypothesized that their fusion would enhance SER performance by leveraging their complementary strengths. To this end, we proposed **HYFuse**, a novel framework that transforms RLRs and CBRs into hyperbolic space for effective fusion. Through the integration of x-vector (RLR) and SoundStream (CBR), **HYFuse** outperforms individual representations and homogeneous fusion approaches, achieving SOTA performance. These findings highlight the potential of heterogeneous representation fusion in advancing SER and also as a reference for future research exploring such heterogeneous fusion.

6. References

[1] M. Tahon and L. Devillers, "Towards a small set of robust acoustic features for emotion recognition: Challenges," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 16–28, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:8157773

[2] G. Sougancioulu, O. Verkholyak, H. Kaya, D. Fedotov, T. Cadee, A. A. Salah, and A. Karpov, "Is everything fine, grandma? acoustic and linguistic modeling for robust elderly speech emotion recognition," in *Interspeech*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221535008

[3] D. Tanko, S. Dogan, F. B. Demir, M. Baygin, S. E. Sahin, and T. Tuncer, "Shoelace pattern-based speech emotion recognition of the lecturers in distance education: Shoepat23," *Applied Acoustics*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:246425020

[4] K. Krishna Kishore and P. Krishna Satish, "Emotion recognition in speech using mfcc and wavelet features," in *2013 3rd IEEE International Advance Computing Conference (IACC)*, 2013, pp. 842–847.

[5] A. Milton, S. S. Roy, and S. T. Selvi, "Svm scheme for speech emotion recognition using mfcc feature," *International Journal of Computer Applications*, vol. 69, no. 9, 2013.

[6] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech emotion recognition with dual-sequence lstm architecture," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6474–6478.

[7] S. K. Hazra, R. R. Ema, S. M. Galib, S. Kabir, and N. Adnan, "Emotion recognition of human speech using deep learning method and mfcc features," *Radioelectronic and Computer Systems*, no. 4, pp. 161–172, 2022.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[9] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[10] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech 2022*, 2022, pp. 2278–2282.

[11] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[12] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[13] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[14] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Interspeech 2021*, 2021, pp. 3400–3404.

[15] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6922–6926.

[16] H. Wu, H.-L. Chung, Y.-C. Lin, Y.-K. Wu, X. Chen, Y.-C. Pai, H.-H. Wang, K.-W. Chang, A. Liu, and H.-y. Lee, "Codec-SUPERB: An in-depth analysis of sound codec models," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 10 330–10 348. [Online]. Available: https://aclanthology.org/2024.findings-acl.616/

[17] W. Ren, Y.-C. Lin, H.-C. Chou, H. Wu, Y.-C. Wu, C.-C. Lee, H.-y. Lee, and Y. Tsao, "Emo-codec: An in-depth look at emotion preservation capacity of legacy and neural codec models with subjective and objective evaluations," *arXiv preprint arXiv:2407.15458*, 2024.

[18] P. Mousavi, L. Della Libera, J. Duret, A. Ploujnikov, C. Subakan, and M. Ravanelli, "Dasb–discrete audio and speech benchmark," *arXiv preprint arXiv:2406.14294*, 2024.

[19] Y. Wu, P. Yue, C. Cheng, and T. Li, "Investigation of ensemble of self-supervised models for speech emotion recognition," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, pp. 988–995.

[20] A. Arunkumar, V. Nileshkumar Sukhadia, and S. Umesh, "Investigation of ensemble features of self-supervised pretrained models for automatic speech recognition," in *Interspeech 2022*, 2022, pp. 5145–5149.

[21] D. Combei, A. Stan, D. Oneata, and H. Cucu, "Wavlm model ensemble for audio deepfake detection," *arXiv preprint arXiv:2408.07414*, 2024.

[22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. 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.

[23] 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)*. IEEE, 2018, pp. 5329–5333.

[24] O. Chetia Phukan, A. Balaji Buduru, and R. Sharma, "Transforming the embeddings: A lightweight technique for speech emotion recognition tasks," in *Interspeech 2023*, 2023, pp. 1903–1907.

[25] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "Speechtokenizer: Unified speech tokenizer for speech language models," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=AF9Q8Vip84

[26] 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.

[27] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[28] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "Superb: Speech processing universal performance benchmark," in *Interspeech 2021*, 2021, pp. 1194–1198.