HERTS: Hybrid Prosodic and XLSR Representations for Speech Emotion Recognition

Brandon Wang
Department of Mechanical Engineering & Materials Science
Duke University
Durham, NC, USA
brandon.wang649@duke.edu

***Abstract*—Speech emotion recognition (SER) remains challenging due to speaker variability, limited labeled data, and poor cross-dataset generalization. This work presents HERTS, a hybrid representation framework that integrates interpretable prosodic features with pretrained wav2vec 2.0 XLSR-53 embeddings to improve actor-independent SER. I demonstrate that a lightweight multilayer perceptron (MLP) trained on these hybrid features outperforms models based on either feature type alone, achieving a macro-F1 score of 0.678 on CREMA-D. I further evaluate zero-shot transfer to external datasets and show that hybrid representations provide greater robustness than standalone embedding-based approaches. These findings highlight the value of combining classic signal descriptors with modern self-supervised speech embeddings to enable efficient and generalizable SER.**

***Index Terms*—speech emotion recognition, hybrid features, prosody, wav2vec 2.0, XLSR-53, cross-dataset generalization**

I. Introduction

Speech Emotion Recognition (SER) seeks to infer a speaker's affective state from acoustic characteristics of their voice. Robust SER systems have potential applications in affect-aware human-computer interaction, mental health monitoring, and socially assistive robotics. However, achieving reliable performance in real-world conditions remains challenging due to substantial speaker variability, heterogeneous recording environments, and the limited availability of labeled emotional speech. These factors contribute to poor generalization outside the training distribution, particularly under actor-independent evaluation protocols that require the model to classify emotions from previously unseen speakers.

Early SER research primarily relied on handcrafted prosodic and spectral descriptors such as pitch, energy, and Mel-Frequency Cepstral Coefficients (MFCCs) [1]. These features offer interpretability and capture well-known correlates of emotional expression, but they are often insufficient for modeling the complex acoustic variability present in natural speech [2]. Recent advances in self-supervised learning have introduced large-scale representation models, including wav2vec 2.0 XLSR-53 [3], that learn contextualized embeddings directly from raw waveforms without requiring labeled data. These embeddings exhibit strong performance across paralinguistic tasks; however, they may not fully capture fine-grained prosodic cues and can still degrade substantially under strict speaker-disjoint evaluation [4].

This work investigates whether combining interpretable prosodic features with powerful self-supervised speech embeddings can yield more robust and generalizable SER models [5]. I introduce **HERTS** (Human Emotion Recognition Through Speech), a hybrid representation framework that concatenates classic prosodic descriptors with pretrained XLSR-53 embeddings and applies Principal Component Analysis (PCA) to construct a compact feature space suitable for lightweight classifiers. A MLP is trained on these hybrid representations to assess whether complementary information from prosody and self-supervised embeddings improves recognition performance under actor-independent conditions. To better isolate the contribution of each feature family, I first conduct an ablation study comparing prosodic-only, embedding-only, and hybrid configurations. Experiments on the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [6] dataset are performed using a strict speaker-disjoint protocol. In addition, I perform zero-shot evaluations on external datasets to characterize the extent to which hybrid representations transfer across corpora.

The contributions of this work are as follows:

- I propose a hybrid SER representation that integrates handcrafted prosodic features with pretrained XLSR-53 embeddings to capture complementary emotional information.
- I conduct an ablation study that quantifies the individual contributions of prosodic features and XLSR-53 embeddings to overall SER performance.
- I demonstrate that an MLP classifier trained on the hybrid representation outperforms models based solely on prosody or self-supervised embeddings under actor-independent evaluation, achieving a macro-F1 score of 0.678 on CREMA-D.
- I evaluate zero-shot cross-dataset transfer and characterize the generalization behavior of hybrid features relative to embedding-only approaches.

Overall, this work shows that combining interpretable signal-level descriptors with modern self-supervised speech embeddings provides an effective and computationally efficient strategy for improving both speaker-independent and cross-dataset SER performance.

II. Related Work

Research in SER spans handcrafted acoustic features, deep neural architectures, and self-supervised representation learning. Early approaches extensively utilized prosodic and spectral features including pitch, energy, jitter, shimmer, and MFCCs, often extracted using openSMILE [2]. While these features are interpretable and computationally efficient, large-scale comparative studies have shown that handcrafted descriptors alone are insufficient for modeling speaker- and context-dependent emotional variability, particularly in naturalistic conditions [7], [8].

Deep learning models, including convolutional and recurrent neural networks trained on spectrograms, MFCCs, or raw audio, have demonstrated improved within-dataset emotion classification performance [9], [10]. However, extensive cross-dataset evaluations reveal that such models often overfit to speaker-specific or dataset-specific artifacts, resulting in substantial performance degradation under speaker-disjoint or cross-dataset conditions [5], [11]. This performance gap has motivated research into domain adaptation and adversarial training methods aimed at improving robustness to distribution shifts [12].

Self-supervised learning has recently provided powerful alternatives for speech representation. Models such as wav2vec 2.0 [13], XLSR-53 [3], and HuBERT [14] learn contextualized embeddings from large unlabeled corpora and achieve state-of-the-art performance across several paralinguistic tasks. Despite their strong representational capacity, studies have shown that such embeddings may insufficiently encode fine-grained prosodic cues relevant for emotion discrimination, suggesting the need for complementary prosodic features [4], [13], [15].

Hybrid SER frameworks integrate handcrafted features with learned embeddings to leverage the complementary strengths of each modality. Prior work demonstrates that hybridization can improve robustness to speaker variability and mismatched recording conditions [16], [17]. However, the literature rarely isolates the relative contribution of each feature type through structured ablation, and hybrid effectiveness remains dependent on dataset characteristics and evaluation protocols.

The present work builds on these findings by constructing a hybrid representation that concatenates prosodic descriptors with pretrained XLSR-53 embeddings. In contrast to earlier studies, I perform a structured ablation analysis to quantify the contribution of each component and evaluate zero-shot transfer to external datasets to characterize hybrid generalization behavior under cross-dataset conditions.

III. Methods

The HERTS framework transforms raw audio into a compact hybrid feature representation composed of classic prosodic descriptors and self-supervised speech embeddings. Figure 1 summarizes the full processing pipeline. This section describes only the core components of the feature extraction and classification framework.

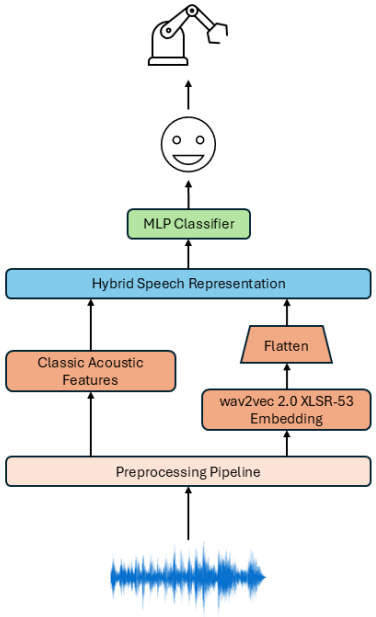


Fig. 1: Overview of the HERTS feature pipeline. Audio is preprocessed, classic acoustic features and XLSR-53 embeddings are extracted in parallel, embeddings are standardized and reduced using PCA, and both streams are concatenated into a hybrid representation for downstream classification.

A. Preprocessing

All audio samples were converted to mono, resampled from their original sample rates to 16 kHz, and trimmed or zero-padded to a standardized duration to preserve pauses carrying emotional content.. Corrupted or extremely short utterances were removed during data preparation. These transformations were applied identically across CREMA-D [6], ESD [18], and EYASE [19]. No dataset-specific preprocessing was performed.

B. Classic Acoustic Features

Classic acoustic descriptors were extracted from the waveform using standard digital signal processing routines. These features included MFCCs, log-mel filterbank energies, short-time energy, and fundamental frequency (pitch) statistics. A 5-dimensional prosodic summary vector was computed for each utterance, consisting of interpretable pitch and energy-related measures. All classic features were standardized using a standard scaler fit on the training set only.

C. Self-Supervised Embeddings

Deep contextual speech embeddings were obtained using the pretrained wav2vec 2.0 XLSR-53 model [3]. For each utterance, the final hidden-state activations were averaged across time to produce a 1024-dimensional embedding vector. Embeddings were normalized using a standard scaler fit on the training embeddings.

D. Hybrid Representation

To reduce redundancy and stabilize downstream learning, PCA was applied to the standardized XLSR-53 embeddings, reducing the dimensionality to 256 components. PCA was trained exclusively on the training split. The final hybrid representation concatenates the 256-dimensional reduced embedding with the 5-dimensional prosodic vector, yielding a 261-dimensional feature vector for classification.

E. Classification Models

Three classifiers were evaluated:

- **Logistic Regression:** A multinomial logistic regression classifier with elastic-net regularization was trained as a linear baseline. The model was optimized using the Stochastic Average Gradient Adaptive (SAGA) solver and evaluated on classic-only and hybrid feature configurations to assess linear separability.
- **Random Forest:** A nonlinear baseline capable of modeling decision boundaries shaped by interactions among features. Model configurations included up to 500 trees with a maximum depth of 20, tuned based on validation set performance on the hybrid feature representation.
- **Multilayer Perceptron (MLP):** A feed-forward network with two fully connected hidden layers (256 and 128 units), ReLU activations, dropout, and a softmax output over the six canonical emotions. Training used the Adaptive Moment Estimation (Adam) optimizer with a learning rate of 1×10^{-3} and early stopping based on validation macro-F1, terminating when no improvement was observed for ten consecutive epochs.

IV. Experimental Setup

This section describes the datasets used in this study, the speaker-disjoint evaluation protocol, the metrics adopted for model assessment, and the experimental configurations used for actor-independent, ablation, and zero-shot evaluation.

A. Datasets

1) Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D): The CREMA-D dataset contains 7,442 audio clips produced by 91 actors spanning diverse demographic groups, with utterances labeled as one of six emotion categories: *anger*, *disgust*, *fear*, *happy*, *neutral*, and *sad*. Only the audio modality is used in this work. Additionally, audio samples are distributed evenly across the six emotion classes, ensuring a balanced class distribution suitable for supervised learning. CREMA-D serves as the sole supervised training corpus for all experiments.

2) Emotional Speech Dataset (ESD): The ESD dataset comprises 17,500 professionally recorded utterances in English and Mandarin Chinese. ESD includes five emotional categories: *anger*, *happy*, *neutral*, *sad*, and *surprise*. To align with CREMA-D's taxonomy, *surprise* was mapped to *happy* for evaluation. The Mandarin subset of ESD is used exclusively for zero-shot generalization experiments.

3) Egyptian Arabic Speech Emotion (EYASE): The EYASE dataset contains 579 utterances of Egyptian Arabic emotional speech recorded by native speakers. Its labels mirror those of CREMA-D and were mapped consistently. EYASE enables evaluation in a non-Indo-European language setting and is also used only for zero-shot evaluation.

B. Actor-Disjoint Protocol

CREMA-D was partitioned into non-overlapping training (70%), validation (15%), and testing (15%) sets, stratified by actor. All preprocessing, scaling, and PCA fitting were performed using only the training speakers to prevent leakage. This protocol ensures evaluation reflects generalization to entirely unseen speakers.

C. Evaluation Metrics

Performance was assessed using accuracy and macro-averaged F1. Macro-F1 was emphasized because it weighs classes uniformly and is more sensitive to performance differences across emotion categories. Confusion matrices were generated to analyze systematic error patterns and misclassifications.

D. Ablation: The Effect of Feature Families

To isolate the contribution of each feature family, classifiers were trained under three feature configurations: (1) classic prosodic features only on logistic regression, (2) XLSR-53 embeddings only on MLP, and (3) the full hybrid representation on MLP. All training and validation procedures were held constant across conditions and models to ensure that performance differences reflect representational differences rather than training effects.

E. Zero-Shot Generalization to non-English Languages

To assess cross-dataset and cross-lingual robustness, identical preprocessing transforms and classifiers trained on CREMA-D were applied directly to ESD and EYASE without fine-tuning. This zero-shot setup evaluates the extent to which hybrid representations learned from English emotional speech transfer to Mandarin and Arabic datasets exhibiting different phonetic, stylistic, and recording characteristics.

V. Results

The results reveal substantial differences in how classic prosodic features, XLSR-53 embeddings, and hybrid representations contribute to emotion classification. Table I summarizes test accuracy and macro-F1 under the actor-disjoint CREMA-D split. Classic features alone yield weak performance: logistic regression achieves only 34.6% accuracy and a macro-F1 of 0.323, demonstrating that low-dimensional prosodic descriptors are insufficient for capturing the complexity of emotional speech when generalizing to unseen speakers. Incorporating pretrained XLSR-53 embeddings leads to a dramatic improvement. The embedding-only MLP reaches 67.2% accuracy and a macro-F1 of 0.672, nearly doubling the classic-only baseline. The highest performance is obtained with the hybrid representation, where concatenating the prosodic summary

TABLE I: Model performance across datasets on test splits.

Dataset	Model	Feature Set	Accuracy	Macro-F1
CREMA-D	Logistic Regression	Classic	0.3458	0.3235
		Hybrid	0.6507	0.6498
	Random Forest	Hybrid	0.5209	0.4969
	MLP	XLSR	0.6725	0.6722
		Hybrid	**0.6794**	**0.6781**
ESD	MLP	Hybrid	0.3819	0.3149
EYASE	MLP	Hybrid	0.2712	0.1521

with PCA-reduced embeddings yields 67.9% accuracy and a macro-F1 of 0.678. Although the gain over embeddings alone is modest, it is consistent and indicates that prosodic cues provide complementary information beyond what is captured by self-supervised representations.

A. Model Performance on Hybrid Features

Performance comparisons were conducted by training all three classifiers on the identical hybrid feature representation, ensuring that differences in accuracy and error patterns reflect the modeling capacity of each architecture rather than variability in input features.

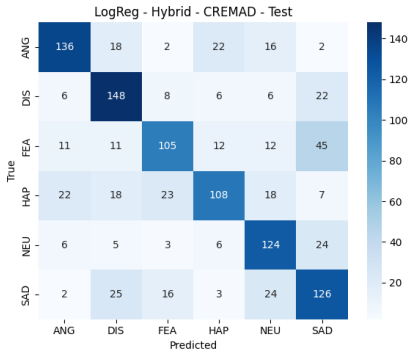
The logistic regression baseline attains 65.1% accuracy and a macro-F1 of 0.650. The confusion matrix in Figure 2a indicates that the hybrid representation provides sufficient separability for a linear model to form meaningful decision boundaries. Nevertheless, notable confusions persist, particularly between low-arousal classes such as *neutral* and *sad*, reflecting the limitations of linear decision surfaces in capturing subtle emotional distinctions.

The random forest model achieves 52.1% accuracy and a macro-F1 of 0.497. As shown in Figure 2b, performance is inconsistent across classes. High-arousal emotions such as *anger* are recognized more reliably, whereas *fear* and *neutral* remain difficult to classify. This behavior suggests that tree ensembles struggle to represent continuous, high-dimensional embedding features, often fragmenting the feature space rather than forming smooth decision boundaries.

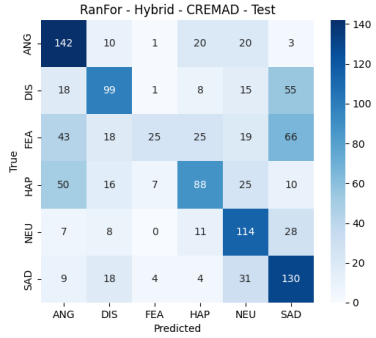
The MLP achieves the strongest performance, reaching 67.9% accuracy and a macro-F1 of 0.678. Beyond its ability to model nonlinear relationships within the hybrid feature space, the model also benefits from early stopping. Validation loss begins to increase while training loss continues to decline, indicating overfitting; early stopping based on validation macro-F1 selects a checkpoint that balances performance across classes. The confusion matrix in Figure 2c exhibits clearer diagonal structure and reduced cross-class leakage compared to the other models. Notably, the MLP better separates *neutral* from *sad* and improves recognition of *fear*, demonstrating that it most effectively exploits the complementary information encoded in the hybrid representation.

B. Ablation Study

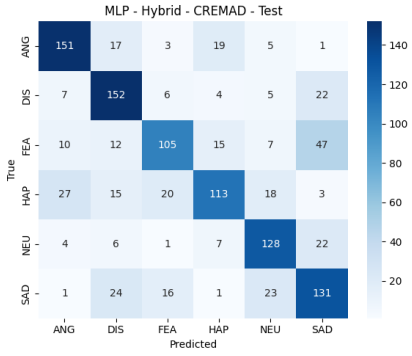
The contribution of each feature type was evaluated by training models on classic prosodic features alone, XLSR-53 embeddings alone, and the full hybrid representation. Classic



(a) Logistic Regression



(b) Random Forest



(c) MLP

Fig. 2: Confusion matrices for logistic regression, random forest, and MLP classifiers using hybrid features on the CREMA-D test set. Panels (a)–(c) correspond to the listed models.

features perform poorly in isolation: logistic regression attains only 34.6% accuracy and a macro-F1 of 0.323, indicating that low-dimensional prosodic cues are insufficient for robust actor-independent emotion classification (Figure 3a). Using frozen XLSR-53 embeddings yields a substantial improvement; the embedding-only MLP achieves 67.2% accuracy and a macro-F1 of 0.672, demonstrating that self-supervised multilingual representations capture rich emotional structure. The hybrid representation achieves the highest overall performance, with the hybrid MLP reaching 67.9% accuracy and a macro-F1 of 0.678. Although the gain over embeddings alone is modest, it consistently reduces key confusions, particularly between

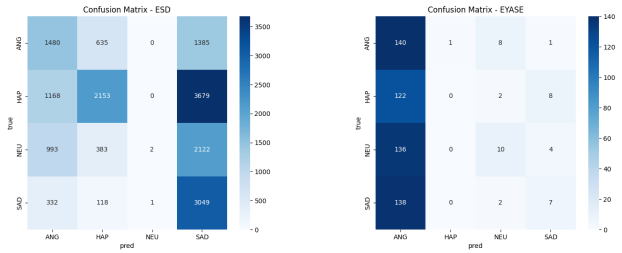
neutral and *sad*, indicating that prosodic cues provide complementary information beyond that encoded in the pretrained embeddings (Figure 3b).

(a) Logistic Regression - Classic (b) MLP - Embedding

Fig. 3: Confusion matrices for models using single feature types on the CREMA-D test set.

C. Zero-Shot Generalization

Zero-shot evaluation on the external ESD and EYASE datasets assessed the extent to which representations learned from CREMA-D transfer to new linguistic and acoustic domains. On ESD (Figure 4a), the hybrid MLP achieved 38.2% accuracy and a macro-F1 of 0.315, with reasonable recall for *sad* but markedly low performance on *neutral*. Performance declined further on EYASE (Figure 4b), where accuracy dropped to 27.1% and macro-F1 to 0.152. Predictions were heavily biased toward *anger* and *sad*, while *happy* and *neutral* were seldom identified. These results indicate that, although the hybrid representation performs well under speaker-independent evaluation within CREMA-D, it generalizes poorly across languages and recording conditions. Likely contributing factors include phonetic mismatch, cultural differences in emotional expression, and substantial variation in dataset scale. Overall, the zero-shot findings highlight persistent challenges in cross-lingual SER and suggest that techniques such as domain adaptation or limited supervised fine-tuning may be required for robust performance in new linguistic domains.



(a) ESD (b) EYASE

Fig. 4: Confusion matrices for the MLP using hybrid features on external datasets. Panels (a) and (b) show performance on ESD and EYASE, respectively.

VI. Discussion

The experimental results highlight both the strengths and limitations of the proposed HERTS framework. The hybrid representation demonstrates strong speaker-independent performance on CREMA-D, with the hybrid MLP achieving the highest accuracy and macro-F1 while requiring only a frozen XLSR-53 encoder and a lightweight classifier. These findings confirm that combining prosodic cues with deep self-supervised embeddings yields a more expressive feature space than either modality alone. The consistent reduction in confusions, especially between low-arousal categories such as *neutral* and *sad*, further supports the complementary nature of prosodic information within the hybrid representation.

At the same time, the system exhibits notable weaknesses, particularly in zero-shot generalization. Accuracy and macro-F1 decrease sharply on ESD and EYASE despite the multilingual nature of XLSR-53. This suggests that, while XLSR-53 effectively models phonetic and linguistic variation, emotional expression appears far more domain-dependent, shaped by language-specific prosody and recording conditions. The confusion patterns on non-English datasets where predictions skew toward *anger* and *sad* indicate that cross-lingual SER requires more than phonetic generalization. These observations are consistent with prior work showing persistent challenges in cross-corpus SER, even with state-of-the-art self-supervised representations.

The model's performance also highlights difficulty with low-arousal classes more generally, which is common in SER due to subtle acoustic cues and subjective distinctions. Addressing these shortcomings may require integrating temporal modeling, leveraging attention over longer contexts, or incorporating additional paralinguistic features.

Overall, the results show that the hybrid HERTS framework is effective for in-domain, speaker-independent SER but insufficient for reliable cross-dataset or cross-lingual transfer. Overcoming these limitations will require domain adaptation, multilingual fine-tuning, or exposure to more diverse emotional speech data.

VII. Conclusion

This work introduced HERTS, a hybrid SER framework that combines prosodic features with wav2vec 2.0 XLSR-53 embeddings. The hybrid MLP achieved the best actor-independent performance on CREMA-D, showing that prosodic cues complement self-supervised representations. However, zero-shot tests on ESD and EYASE revealed substantial drops in accuracy, highlighting the strong domain and language dependence of emotional expression. Improving cross-lingual robustness remains an open challenge, motivating future work in domain adaptation, multilingual fine-tuning, and broader training data. Owing to its lightweight architecture, the hybrid model is well suited for real-time human-robot interaction and provides a promising foundation for practical SER systems. The link to the GitHub repository is here.

REFERENCES

[1] B. Schuller and A. Batliner, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[2] F. Eyben and B. Schuller, "opensmile:): the munich open-source large-scale multimedia feature extractor," *SIGMultimedia Rec.*, vol. 6, no. 4, p. 4–13, Jan. 2015. [Online]. Available: https://doi.org/10.1145/2729095.2729097

[3] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[4] T. Li, L. Qu, C. Weber, T. Pekarek-Rosin, F. Ren, and S. Wermter, "Disentangling prosody representations with unsupervised speech reconstruction," *arXiv preprint arXiv:2212.06972*, 2023.

[5] J. Gideon, S. Khorram, F. Aldeneh, D. Dimitriadis, and S. Narayanan, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain adaptation," *Interspeech*, pp. 1118–1122, 2017.

[6] 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.

[7] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Interspeech*, 2009, pp. 312–315.

[8] B. Schuller *et al.*, "Cross-corpus acoustic emotion recognition: Variances and strategies," *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction*, pp. 552–561, 2010.

[9] A. Satt, R. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep convolutional neural networks," in *Interspeech*, 2017, pp. 1089–1093.

[10] X. Zhang *et al.*, "Speech emotion recognition using deep convolutional neural networks," *IEEE Access*, vol. 5, pp. 10 652–10 661, 2017.

[11] H. Stuhlsatz *et al.*, "Deep neural networks for acoustic emotion recognition," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 80–93, 2011.

[12] A. Hassan and R. Damper, "A comparative study of machine learning and deep learning models for speech emotion recognition," *IEEE Transactions on Affective Computing*, 2016.

[13] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, 2020, arXiv:2006.11477.

[14] 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.

[15] S. Latif *et al.*, "Deep representation learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, 2020.

[16] E. Lakomkin *et al.*, "Emotional speech recognition using multimodal deep learning," in *IJCNN*, 2018, pp. 1–7.

[17] P. Manocha *et al.*, "Multimodal speech emotion recognition using hybrid features," *IEEE Access*, vol. 11, pp. 45 112–45 125, 2023.

[18] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 920–924.

[19] M. E. Seknedy and S. A. Fawzi, "Emotion recognition system for arabic speech: Case study egyptian accent," in *Model and Data Engineering (MEDI 2022)*, ser. Lecture Notes in Computer Science, P. Fournier-Viger, A. Hassan, and L. Bellatreche, Eds., vol. 13761. Springer, Cham, 2023, pp. 89–100.