



Multimodal Emotion Diarization: Frame-Wise Integration of Text and Audio Representations

Ziv Tamir¹, Thomas Thebaud², Jesus Villalba², Najim Dehak², Oren Kurland¹

¹Faculty of Data and Decision Sciences, Technion - Israel Institute of Technology, Haifa, Israel

²Department of Electrical and Computer Engineering, Johns Hopkins University, Maryland, USA

ziv.tamir@campus.technion.ac.il, tthebaul@jhu.edu, jvillal17@jhu.edu, ndehak3@jhu.edu, kurland@technion.ac.il

Abstract

Speech emotion diarization (SED) is the task of segmenting an audio stream into time-continuous emotional states, akin to speaker diarization but for emotions. While traditional speech emotion recognition (SER) assigns a single emotion label to a given utterance, real-world conversations exhibit dynamic emotional transitions that require a more granular approach. In this work, we propose a novel multimodal SED framework that uses frame-wise integration of text and audio embeddings using temporal synchronization and direct concatenation, followed by a context-aware sliding window smoothing mechanism. Audio representations are extracted using WavLM, and EmoBERTa generates text embeddings aligned to spoken words. We evaluate our approach using Emotion Diarization Error Rate (EDER), a metric designed for SED. Experimental results show that our proposed method significantly improves diarization performance, with respect to late fusion and cross-attention methods yielding an EDER of 25%.

Index Terms: speech emotion recognition, speech emotion diarization, multimodal fusion

1. Introduction

Emotion recognition from speech has gained significant attention in recent years, driven by its applications in human-computer interaction [1, 2], healthcare [3], and affective computing [4]. Traditional approaches to emotion classification typically assign a single emotion label to an entire utterance. However, human emotions are dynamic and can vary significantly within a conversation [2]. This necessitates the development of speech emotion diarization (SED), a process analogous to speaker diarization, which aims to detect the boundaries between emotional states over time.

Despite the growing interest in speech emotion recognition [5], SED has not attracted yet much research attention, and accordingly, evaluation datasets are scarce. While many speech emotion recognition datasets propose a time-continuous annotation [2], it is often for emotional attributes, such as Valence, Dominance, and Arousal, and rarely for categories, such as Angry, Sad, or Happy. To the best of our knowledge, the only dataset of English speech annotated with frame-wise boundaries between emotional categories is the Zion Emotion Dataset (ZED) [6], which includes four categories, annotated over a total time of 17 minutes of speech. This is sufficient for testing SED models, but other datasets are needed for training. Our selection of training datasets includes IEMOCAP [7], RAVDESS [8], EmoV-DB [9], ESD [10], and JL-CORPUS [11].

Although speech is considered a more informative means of communication than text, intents and emotions are expressed in written text as well [12]. In speech emotion recognition, a wide

range of recent systems now fuse transcripts and speech inputs to improve the accuracy of their predictions [13, 14, 15, 16]. However, in the case of SED, one cannot predict frame-wise or letter-wise segmentation of emotions based on transcripts, as emotions are encoded across entire sentences. To address this problem, the two main approaches in the speech emotion recognition literature are cross-attention early fusion [16, 17] and late fusion of the predictions [18, 15].

In this article, we propose a new early fusion technique: the concatenation of text and audio representations using temporal alignment, followed by a sliding window technique. Our main contributions can be stated as: a new fusion technique using direct concatenation of text and speech embeddings, using alignments to match the speech to the words; as well as the addition of a sliding window to smooth the emotion predictions, as intuitively emotions can not change within the range of milliseconds.

This paper is structured as follows: Section 2 details the model we proposed, followed by Section 3 which details the experimental setup, which includes the metrics, datasets and training details. Section 4 presents experimental results and ablation studies, then Section 5 concludes with a discussion on future research directions.

2. Proposed Model

Figure 1 presents an overview of the proposed architecture, which integrates audio embeddings (A_t) with aligned text embeddings (T_t) to enhance emotion recognition and diarization. Audio embeddings are initially derived from a self-supervised speech model on a frame-by-frame basis. Simultaneously, transcripts are aligned with their respective timestamps, and text embeddings are generated and temporally matched to the corresponding *audio frames*. An audio frame is the time interval corresponding to one audio embedding. By fusing these complementary representations, the system obtains a unified multimodal embedding (X_t), which is subsequently employed for both emotion recognition and diarization tasks.

Emotion recognition is performed using a fully connected network applied to the average of the multimodal embeddings of the entire utterance, while emotion diarization is based on aggregation over a sliding window to obtain frame-level emotion predictions. The sliding window mechanism ensures that each frame's prediction incorporates contextual information from surrounding frames, improving the reliability of diarization.

2.1. Audio and text embedding extraction and alignment

Audio embeddings are derived using the large version of the WavLM model¹ [19], a self-supervised framework adept at cap-

¹<https://huggingface.co/microsoft/wavlm-large>

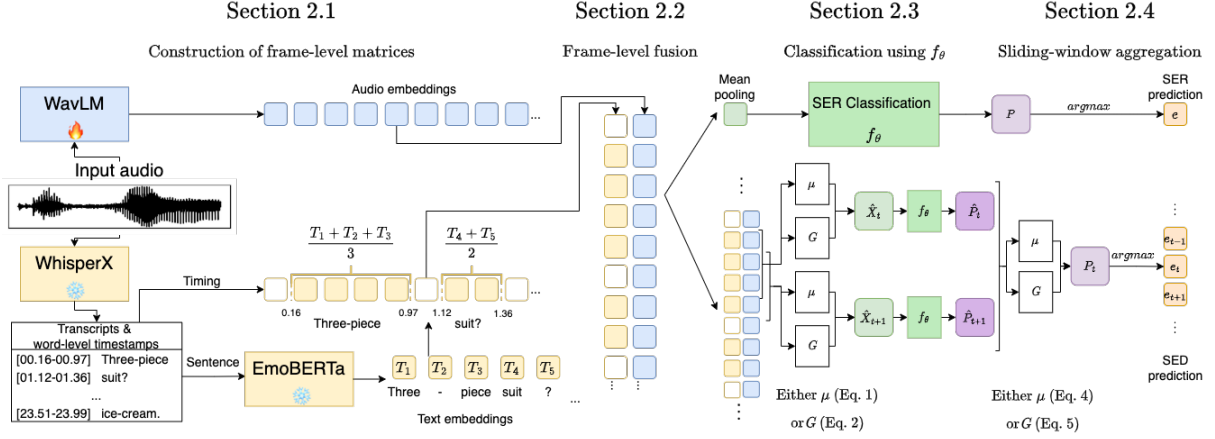


Figure 1: *Schematic of the embedding concatenation technique proposed. WavLM extracts audio embeddings, and EmoBERTa generates text embeddings aligned via WhisperX. The fused representations are used for utterance-level SER and frame-level emotion diarization via the proposed sliding window mechanism.*

turing both local and global acoustic patterns through frame-level representations. The model is fine-tuned to adapt to the specific characteristics of emotion recognition and diarization.

Transcriptions and word timings are obtained using frozen WhisperX² [20, 21], which provides precise temporal alignment of words within the audio signal. Text embeddings are then extracted via the pretrained frozen EmoBERTa large model³ [12], which encodes the emotional context present in the text. For words composed of multiple tokens, we average the token embeddings to obtain a single word-level vector. Each word-level embedding is then assigned to every audio frame that falls within its time interval (as determined by the word timings), thus constructing a frame-level text embedding matrix. Finally, frames without active words are assigned zero vectors to ensure consistent alignment between the audio and text modalities.

2.2. Fusion at the embedding level

Let T be the total number of audio frames in an utterance, and let d_{audio} and d_{text} be the dimensions of the audio and text embeddings, respectively. The audio and text embeddings are concatenated to create a frame-level fusion matrix of dimensions $T \times (d_{\text{audio}} + d_{\text{text}})$. This fusion mechanism enables the model to exploit complementary information from both modalities at a granular level, enriching frame-level representations with emotional context derived from textual data.

To show the effectiveness of the proposed fusion technique, we compare with two existing methods: training a **cross-attention** layer directly between the text embeddings and the audio embeddings (which outputs a sequence of embeddings $\hat{X}_{1 < t < T}$) and processing the text and audio embeddings separately before operating a **late fusion**, at the predictions level.

2.3. Prediction

The constructed fusion matrix, denoted $X \in \mathbb{R}^{T \times (d_{\text{audio}} + d_{\text{text}})}$, serves as input for two tasks: SER and SED.

For emotion recognition, the mean embedding of the entire fusion matrix is passed through a fully connected layer, f_{θ} , to predict the overall emotion of the segment.

For emotion diarization, using the same f_{θ} , a sliding window mechanism is applied to refine frame-level emotion pre-

dictions. This method ensures that each frame’s embedding, X_t (the representation of the t -th frame), is smoothed by incorporating local context from surrounding frames, rather than being predicted in isolation.

2.4. Sliding window aggregation

Emotion diarization is performed using a sliding window mechanism that assigns an emotion probability distribution to each frame. The following steps are applied.

2.4.1. Embeddings aggregation

Each window contains a subset of the frame sequence, capturing local emotional context. Given a fixed-length window W , centered on the frame t , the aggregated embedding \hat{X}_t is computed using one of the following techniques:

- **Mean Embedding Aggregation (μ):** Compute the mean of all embeddings in the window:

$$\hat{X}_t = \frac{1}{|W|} \sum_{i=-\lfloor W/2 \rfloor}^{\lfloor W/2 \rfloor} X_{t+i}. \quad (1)$$

- **Gaussian-Weighted Embedding Aggregation (G):** Assign higher importance to the frames near the center of the window using Gaussian weights:

$$\hat{X}_t = \sum_{i=-\lfloor W/2 \rfloor}^{\lfloor W/2 \rfloor} w_i X_{t+i}, \quad (2)$$

where

$$w_i = \frac{\exp\left(-\frac{i^2}{2\sigma^2}\right)}{\sum_{j=-\lfloor W/2 \rfloor}^{\lfloor W/2 \rfloor} \exp\left(-\frac{j^2}{2\sigma^2}\right)}. \quad (3)$$

The parameter σ controls the spread of the Gaussian distribution. We set $\sigma = W/4$, ensuring that the weighting function smoothly distributes importance across the entire window. With this choice, approximately 95% of the total weight falls within $\pm 2\sigma$, covering the entire window range. This design ensures that even the frames at the edges of the window receive non-negligible weights, preventing sharp discontinuities in the smoothing.

²<https://github.com/m-bain/whisperX>

³<https://huggingface.co/tae898/emoberta-large>

Once the aggregated embedding, \hat{X}_t , is computed, it is passed through the same fully connected layer, f_θ , followed by a softmax activation function, to predict the probability distribution over emotion classes, \hat{P}_t . Here, \hat{P}_t represents the window-level prediction for the window centered on frame t .

2.4.2. Probability aggregation

Since frames appear in multiple overlapping windows, each frame is assigned multiple window-level probability distributions \hat{P}_j , each corresponding to a different window j that includes frame t . The final frame-level probability distribution P_t is obtained by aggregating all window-level predictions for the windows the frame t was part of. The aggregation options are the same as for the embeddings:

- **Mean Probability Aggregation (μ):** The mean of all probability distributions assigned to frame t :

$$P_t = \frac{1}{N} \sum_{j \in \mathcal{W}(t)} \hat{P}_j, \quad (4)$$

where $\mathcal{W}(t)$ is the set of all windows containing frame t , and $N = |\mathcal{W}(t)|$ is the number of these windows.

- **Gaussian-Weighted Probability Aggregation (G):** Assign higher importance to predictions from windows where the frame appears near the center:

$$P_t = \sum_{j \in \mathcal{W}(t)} w_j \hat{P}_j. \quad (5)$$

The weights w_j are computed using the same Gaussian formulation as in Eq. 3, assigning higher weight to predictions for windows in which the target frame is near the center. This method emphasizes local context in the final frame-wise probability distribution.

2.4.3. Final aggregation methods

By combining different embedding and probability aggregation techniques, we define four distinct methods:

- **REG (Regular Aggregation):** Uses mean aggregation for both embeddings and probabilities. (Eq. 1 and Eq. 4)
- **GE (Gaussian Embeddings):** Uses Gaussian-weighted averaging for embeddings and mean aggregation for probabilities. (Eq. 2 and Eq. 4)
- **GP (Gaussian Probabilities):** Uses mean aggregation for embeddings and Gaussian-weighted averaging for probabilities. (Eq. 1 and Eq. 5)
- **GEP (Gaussian Embeddings and Probabilities):** Uses Gaussian-weighted averaging for both embeddings and probabilities. (Eq. 2 and Eq. 5)

3. Experiments

3.1. Evaluation metric

The Emotion Diarization Error Rate (EDER) [6] is an extension of the Diarization Error Rate (DER) to handle emotion diarization. It aggregates false alarms (FA), missed emotions (ME), confusion (CF), and overlap (OL) into a single metric:

$$EDER = \frac{FA + ME + CF + OL}{Utterance\ Duration},$$

providing a measure of how accurately predicted emotion segments align with ground-truth intervals. We compare our

results against a baseline of 29.2% EDER and confirm the statistical significance of improvements using a two-sided paired t -test at $p < 0.05$.

3.2. Datasets

To our knowledge, the ZED dataset is the only available corpus with frame-level emotion labels [6]. Collected from various YouTube videos, it captures authentic emotional expressions with fine-grained annotations. Each utterance contains a single emotional event, allowing precise analysis of emotional boundaries and transitions. However, with just 180 recordings (17 minutes in total), it is used solely for testing.

To train and validate our models, we will use utterance-level labeled datasets. To generate pseudo-frame-level labels, we assume that each frame shares the same label as the entire utterance. It is important to note that this assumption does not perfectly reflect reality. Previous studies have observed that when an emotionally neutral speech segment is followed by an emotional segment with emotion E, human listeners often classify the entire sequence as emotion E [22].

To create multi-emotional utterances, we apply CopyPaste augmentation [23], concatenating neutral and emotional segments. Frame-level labels for these new samples are obtained by merging the original segment labels.

Our training and validation data come from five utterance-level labeled datasets: IEMOCAP [7], RAVDESS [8], EmoVDB [9], ESD [10], and JL-CORPUS [11] — only including utterances labeled as *happy*, *sad*, *angry*, or *neutral*. Table 1 shows the total duration (in minutes) for each emotion. This dataset configuration is the same as the settings described in [6].

As shown in Table 1, these datasets collectively provide a range of labeled emotional expressions at the utterance level. However, truly frame-labeled data, as ZED [6], remains scarce, so we use these sets for utterance-level training.

Table 1: Durations in minutes for each dataset.

Dataset	Happy	Sad	Angry	Neutral	Total
IEMOCAP	96.14	50.56	66.16	76.82	289.68
RAVDESS	5.37	4.96	6.12	2.16	18.61
EmoV-DB	80.63	0.00	60.49	68.55	209.67
ESD	114.58	126.22	118.76	108.49	468.05
JL-Corpus	12.58	5.83	6.57	5.59	30.57
Total	309.3	187.57	258.1	261.61	1016.58

3.3. Training details

The model was trained for 15 epochs with a batch size of 4 on a single NVIDIA A100 GPU. Two separate Adam optimizers were employed: one for the fully connected (FC) layer and another for fine-tuning the WavLM model. The learning rate for the FC layer was set to 10^{-4} , while the learning rate for WavLM fine-tuning was set to 10^{-5} . A learning rate scheduler was applied to dynamically adjust the learning rates during training. We used Cross-entropy loss for both SER and SED. No early stopping was performed. The training was done using the SpeechBrain toolkit⁴, our code is available on github⁵.

4. Results

This section presents the evaluation of fusion strategies, sliding window configurations, and aggregation techniques for emotion diarization.

⁴<https://speechbrain.github.io/>

⁵https://github.com/zivta3/multimodal_emo_dia

Table 2: *EDER of the various fusion methods. A statistically significant difference with the audio-only baseline is marked with '*'. The best result in a column is boldfaced.*

Fusion Technique	Sliding Window	EDER (%) ↓	p-value
Only audio [6]	x	29.2	x
Late fusion	x	29.1	0.9309
Cross attention	x	31.1	0.1139
Concatenation	x	29.0	0.8591
Late fusion (proposed)	✓	27.5	0.2340
Cross attention (proposed)	✓	26.6	0.0604
Concatenation (proposed)	✓	25.0	0.0028(*)

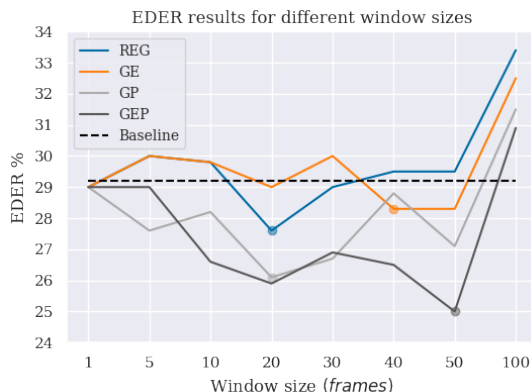


Figure 2: *The effect of window size in our concatenation-based technique on EDER.*

4.1. Fusion of text and audio embeddings

Table 2 reports the EDER of various fusion methods compared to the baseline, which uses only audio embeddings (EDER = 29.2%). Without a sliding window, the late fusion (29.1%) and the concatenation (29.0%) exhibit minimal improvements, with high p-values ($p > 0.85$) indicating no statistical significance with respect to the baseline (Only audio). In contrast, cross-attention (31.1%) leads to performance degradation.

When incorporating a sliding window mechanism, the EDER of all methods fusion improves. Concatenation with a sliding window achieves the lowest EDER (25.0%), which is statistically significantly lower than that of the baseline ($p = 0.0028$). Although cross-attention (26.6%) and late fusion (27.5%) also show reductions in EDER, their improvements remain statistically insignificant. These findings indicate that integrating text and speech embeddings at the frame level via concatenation, rather than fusion at the classifier level, yields superior performance, particularly when coupled with a sliding window that provides local temporal context.

4.2. Effect of window size

Figure 2 illustrates the effect of different window sizes on EDER for the various fusion methods. The Gaussian embeddings and probabilities (GEP) method consistently outperforms other techniques, achieving the lowest EDER (25.0%) with a window size of 50 frames. However, other smoothing techniques, such as GP and GE, perform best within the 20–40 frame range, indicating that this interval balances local context while mitigating over-smoothing.

The EDER of the REG method fluctuates around the baseline, with performance degradation at larger window sizes (>50 frames), likely due to excessive smoothing. The performance

Table 3: *The EDER of various polling techniques applied in our concatenation-based technique. The best result is boldfaced.*

Pooling Technique	Embeddings Pooling Eq.	Probabilities Pooling Eq.	EDER (%) ↓
REG	Eq. 1	Eq. 4	27.6
GE	Eq. 2	Eq. 5	28.3
GP	Eq. 1	Eq. 4	26.1
GEP	Eq. 2	Eq. 5	25.0

deteriorates for very small (≤ 5 frames) and large (100 frames) window sizes, suggesting that insufficient context does not provide meaningful temporal dependencies, while excessive windowing smoothes out fine-grained variations necessary for accurate diarization. The results indicate that an optimal trade-off exists within the 20–50 frame range, where the local context is preserved while minimizing the risk of over-smoothing, and different smoothing techniques reach their optimal performance within this interval.

4.3. Comparison of aggregation methods

Table 3 presents a comparison of the best performing aggregation techniques at their respective optimal window sizes. REG (27.6%) provides moderate improvements over the baseline. GE (28.3%) increases EDER, suggesting that modifying embeddings alone without adjusting probability distributions may introduce inconsistencies. Conversely, GP (26.1%) yields a more substantial reduction in EDER, underscoring the benefits of weighting probability distributions rather than embeddings.

The best performance is achieved with the GEP method (25.0%), which applies Gaussian weighting to both embeddings and probabilities. This result highlights the importance of weighting both the representations and their corresponding probability distributions to emphasize more reliable frames during classification. By jointly refining feature embeddings and frame-wise probability estimates, this method provides a more robust diarization framework.

5. Conclusion and Future Work

We introduced a novel multimodal speech emotion diarization framework that integrates frame-level text and audio embeddings using temporal alignment and direct concatenation. The proposed fusion method, combined with a context-aware sliding window smoothing mechanism, significantly improves emotion diarization performance. Experimental results demonstrated that our concatenation-based fusion approach achieved an EDER of 25.0%, significantly outperforming baseline methods such as score fusion and cross-attention-based fusion strategies.

A key limitation of this approach is its reliance on pseudo-labeling derived from utterance-level datasets, which assumes uniform emotion labels within an utterance and may introduce inconsistencies. In addition, since CopyPaste needs to concatenate utterances, the augmentation might lack semantic context by randomly selecting utterances. Furthermore, the evaluation is constrained by the limited size of the Zion Emotion Dataset (ZED), restricting the diversity of emotional expressions available for testing. Addressing these challenges through larger frame-annotated datasets, using semantic similarity for more natural concatenation, and self-supervised learning techniques will be essential for advancing robust emotion diarization systems.

6. References

- [1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.
- [2] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The msp-conversation corpus," *Interspeech 2020*, 2020.
- [3] M. Dhuheir, A. Albaseer, E. Baccour, A. Erbad, M. Abdallah, and M. Hamdi, "Emotion recognition for healthcare surveillance systems using neural networks: A survey," in *2021 International Wireless Communications and Mobile Computing (IWCMC)*. IEEE, 2021, pp. 681–687.
- [4] S. Alisamir and F. Ringeval, "On the evolution of speech representations for affective computing: A brief history and critical overview," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 12–21, 2021.
- [5] H. Wu, H.-C. Chou, K.-W. Chang, L. Goncalves, J. Du, J.-S. R. Jang, C.-C. Lee, and H.-Y. Lee, "Emo-superb: An in-depth look at speech emotion recognition," *arXiv preprint arXiv:2402.13018*, 2024.
- [6] Y. Wang, M. Ravanelli, and A. Yacoubi, "Speech emotion diarization: Which emotion appears when?" in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–7.
- [7] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.
- [8] S. R. Livingstone and F. 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. 1–35, 05 2018. [Online]. Available: <https://doi.org/10.1371/journal.pone.0196391>
- [9] A. Adigwe, N. Tits, K. E. Haddad, S. Ostadabbas, and T. Dutoit, "The emotional voices database: Towards controlling the emotion dimension in voice generation systems," *arXiv preprint arXiv:1806.09514*, 2018.
- [10] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," 2021.
- [11] J. James, L. Tian, and C. Watson, "An open source emotional speech corpus for human robot interaction applications," 09 2018, pp. 2768–2772.
- [12] T. Kim and P. Vossen, "Emoberta: Speaker-aware emotion recognition in conversation with roberta," *arXiv preprint arXiv:2108.12009*, 2021.
- [13] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE spoken language technology workshop (SLT)*. IEEE, 2018, pp. 112–118.
- [14] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE access*, vol. 9, pp. 47 795–47 814, 2021.
- [15] T. Thebaud, A. Favaro, Y. Guan, Y. Yang, P. Singh, J. Villalba, L. Mono-Velazquez, and N. Dehak, "Multimodal emotion recognition harnessing the complementarity of speech, language, and vision," in *Proceedings of the 26th International Conference on Multimodal Interaction*, 2024, pp. 684–689.
- [16] L. Sun, B. Liu, J. Tao, and Z. Lian, "Multimodal cross-and self-attention network for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4275–4279.
- [17] M. Khan, W. Gueaieb, A. El Saddik, and S. Kwon, "Mser: Multimodal speech emotion recognition using cross-attention with deep fusion," *Expert Systems with Applications*, vol. 245, p. 122946, 2024.
- [18] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion approaches for emotion recognition from speech using acoustic and text-based features," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6484–6488.
- [19] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, p. 1505–1518, Oct. 2022. [Online]. Available: <http://dx.doi.org/10.1109/JSTSP.2022.3188113>
- [20] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [21] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *arXiv preprint arXiv:2303.00747*, 2023.
- [22] S. L. Tóth, D. Sztahó, and K. Vicsi, "Speech emotion perception by human and machine," in *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, A. Esposito, N. G. Bourbakis, N. Avouris, and I. Hatzilygeroudis, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 213–224.
- [23] R. Pappagari, J. Villalba, P. Želasko, L. Moro-Velazquez, and N. Dehak, "Coppaste: An augmentation method for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6324–6328.