Leveraging Wav2Vec2.0 and DistilBERT with Autoencoder-Based Dimensionality Reduction for Continuous Multimodal Emotion Recognition

1st Awatef Messaoudi
University of Tunis El Manar
National Engineering School of, Tunis
Signal, Image and Information Technologies laboratory (LR-11-ES17)
Tunis, Tunisia
Email: awatef.messoudi@enit.utm.tn

2nd Hayet Boughrara
University of Tunis El Manar
National Engineering School of, Tunis
Signal, Image and Information Technologies laboratory (LR-11-ES17)
Tunis, Tunisia
Email: hayet.boughrara@insat.ucar.tn

3rd Zied Lachiri
University of Tunis El Manar
National Engineering School of, Tunis
Signal, Image and Information Technologies laboratory (LR-11-ES17)
Tunis, Tunisia
Email: Zied.lachiri@enit.utm.tn

***Abstract*—The dawn of the Transformer era has advanced emotion recognition systems. Transformers capture contextual dependencies in data, making them highly effective for complex applications such as sentiment analysis and audio emotion recognition. In this study, we used self-supervised learning models such as wav2vec for speech and DistilBERT, a transformer-based model for text, to enhance the continuous emotion recognition system. These models are evaluated in the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database, which contains speech data labeled with emotional dimensions such as valence, arousal, and dominance(VAD). Due to the complexity of the data and computational cost in fusing DistilBERT and wav2vec, we used an autoencoder for dimensionality reduction. To our knowledge, this is the first study to combine wav2vec and DistilBERT-like pre-trained features for Continuous Multimodal Emotion Recognition (CMER), addressing the challenge of limited labeled training data. Our experiments, evaluated using the Concordance Correlation Coefficient (CCC), show a significant performance boost, achieving a CCC of 0.808, 0.719 and 0.635 respectively for VAD dimensions compared to 0.603, 0.736 and 0.647 when using traditional feature extraction techniques (LSTM/CNN1D) on the IEMOCAP dataset. This demonstrates the effectiveness of using SSL models for emotion recognition tasks that typically suffer from small amounts of labeled data.**

***Index Terms*—Continuous Multimodal Emotion Recognition, Wav2vec2.0, DistilBERT, Autoencoder, Valence-Arousal-Dominance (VAD), early fusion.**

I. Introduction

Emotion Recognition (ER) is essential not only for human interactions, but also to enhance human-computer interaction, making systems more reactive and responsive to the emotional states of users [1]. Traditionally, most ER studies have focused on classifying emotions into discrete categories such as happiness, anger, sadness, and fear. However, recent research has shifted towards continuous emotion recognition, where emotional states are represented in a multidimensional space defined by VAD dimensions. By integrating these dimensions, a more comprehensive emotional spectrum, and a richer contextual basis for predicting specific emotional intensities is provided [23]. The emergence of deep learning (DL) technology has enabled the recognition of human emotions through speech [2], text [3], facial expressions [4], and physiological signals [5]. However, performance based on a single modality, remains limited. To overcome these challenges, researchers have integrated two or more modalities [6], [14]. Multimodal Emotion Recognition (MER) field has been significantly advanced with the introduction of Transformer-based models [7]. Specifically, the fusion of speech and text signals which has gained increasing attention in recent years. Self-supervised models such as Wav2Vec2.0 [8] and HuBERT [8] have improved feature extraction and representation learning in the Speec Emotion Recognition (SER) system. Similarly, pre-trained language models such as BERT [9] have demonstrated remarkable improvements by capturing deep contextual representations in textual emotion recognition. Although HuBERT and Wav2vec2.0 [8] perform well for activation and dominance, they struggle with valence prediction. To address this, extensive research has been conducted to combine text and speech modalities to predict emotions expressed in the VAD dimensions and to improve the SER system. In this context, Srinivasan et al. [10] proposed a teacher-student approach that improves SER by integrating lexical information. Their approach achieves state-of-the-art CCC scores on IEMOCAP (0.582 Valence, 0.667 Arousal, 0.545 Dominance), demonstrating the effectiveness of multimodal fusion. Triantafyllopoulos et al. [11] proposed a multistage fusion approach that integrates acoustic and linguistic information. Their architecture consists of a CNN14, a 14 layer convolutional neural network originally adapted from the PANNs (Pretrained Audio Neural Networks) framework [24], for speech feature extraction and a BERT-based text model, with fusion occurring at different stages within the CNN. Their results demonstrated that multistage fusion outperformed both baselines, achieving CCC scores of 0.714 for valence, 0.639 for arousal, and 0.575 for dominance, highlighting the advantage of incorporating linguistic context for improved emotion recognition specifically for the valence

score. Zhang et al. [12] propose a novel emotion recognition approach using a transformer-based model that integrates pre-trained wav2vec 2.0 for the extraction of speech features and BERT for the extraction of text features. Furthermore, LSTM layers are used to learn hidden representations from the merged speech and text data. Evaluated on the Iemocap dataset , their models showed competitive results.

In these studies, wav2vec for speech and BERT for text have proven to be effective solutions for Multimodal Emotion Recognition (MER), enabling the fusion of both acoustic and linguistic information.
While prior work [12] explored the fusion of Wav2Vec2.0 and BERT features and demonstrated improvements using early fusion strategies, our work introduces two key novelties:

- We incorporate a lightweight Transformer, DistilBERT, for the text modality to reduce computational load while maintaining performance.
- We introduce a dedicated autoencoder module for dimensionality reduction of both modalities before fusion, addressing the challenge of high-dimensional multimodal data.

Furthermore, we conduct a comprehensive comparison of early and late fusion strategies, demonstrating that while early fusion yields strong results, late fusion combined with autoencoder-based compression achieves the best performance in valence and dominance recognition. The remainder of this paper is organized as follows: Section 2 gives an overview of the proposed system. Section 3 describes the experiments and the implementation details. Section 4 presents the results and a comparison with previous works. We conclude in the last section.

II. Methodology and Proposed System

The architecture of the proposed system for predicting VAD from speech and text signals is presented in this section. In line with [17], which explores reducing model size while maintaining performance using DistilHuBERT and linguistic informations, our work integrates DistilBERT for textual feature extraction and self-supervised learning Wav2Vec 2.0 representations for speech processing.
The proposed framework is built on the wav2vec2-large-robust model, specifically fine-tuned for emotion recognition, and the DistilBERT base available at: https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim and https://huggingface.co/distilbert/distilbert-base-uncased.
To evaluate the effectiveness of multimodal fusion, we experimented with late fusion and early fusion strategies. The baseline architecture is composed of four principal components:

1) Audio feature extraction Block: Using Wav2vec 2.0 pretrained model, high and complex acoustic features are extracted.

$$F_{\text{acoustics}} = \text{Wav2Vec2}(X_a), \quad F_{\text{acoustics}} \in \mathbb{R}^{B\times T\times d} \quad (1)$$

where X_a is the channel audio input, B is the batch size, T is the sequence length, and d is the dimension of the extracted representation.

2) Textual feature extraction Block: We use DistilBERT pre-trained model to extract deep contextualized representations from textual informations:

$$F_{\text{text}} = \text{DistilBERT}(X_t), \quad F_{\text{text}} \in \mathbb{R}^{B\times L\times h} \quad (2)$$

where X_t is tokenized text, L is the number of tokens in the input sequence, and h is the hidden size of DistilBERT representations.

3) Autoencoder: To reduce the dimensionality of both speech and text embeddings while preserving important features, we introduce an Autoencoder. An autoencoder has a very specific architecture, because the hidden layers are smaller than the input layers and this architecture is called a "bottleneck" architecture [18]. It contains two parts:

- The encoder which transforms the input into a representation in a lower-dimensional space called the latent space. The encoder therefore compresses the input into a less expensive representation. The encoder's formulation for speech and text signals is given by:

$$Z_{\text{acoustic}} = f_{\text{enc}}(F_{\text{acoustic}}), \quad Z_{\text{acoustic}} \in \mathbb{R}^{B\times d'} \quad (3)$$

$$Z_{\text{text}} = f_{\text{enc}}(F_{\text{text}}), \quad Z_{\text{text}} \in \mathbb{R}^{B\times h'} \quad (4)$$

  where d' and h' represent the reduced dimensions for speech and text, respectively.
- The second part is called the decoder, because it must reconstruct, using the latent representation of the input, an output that is as faithful as possible to the input.
  The Decoder (Reconstruction Loss - Training Phase Only) formula is given by:

$$D\text{acoustic} = f\text{dec}(z_{\text{acoustic}}), D\text{acoustic} \approx F\text{acoustic} \quad (5)$$

$$D\text{text} = f\text{dec}(Z_{\text{text}}), D\text{text} \approx F\text{text} \quad (6)$$

The autoencoder is trained to minimize the reconstruction loss:

$$\mathcal{L}\text{rec} = ||F\text{acoustic} - D\text{wav}||^2 + ||F\text{text} - D_{\text{text}}||^2 \quad (7)$$

4) Fusion and Regression Layer

- early Fusion: After obtaining the reduced speech and text embeddings, Features from linguistic and vocal informations are concatenated and fed into a regression layer as shown in Fig.1:

$$P_{\text{fusion}} = \text{Concat}(Z_{\text{acoustic}}, Z_{\text{text}}), \quad P_{\text{fusion}} \in \mathbb{R}^{B\times(d'+h')} \quad (8)$$

$$\hat{Y} = FC(P_{\text{fusion}}), \quad \hat{Y} \in \mathbb{R}^{B\times 3} \quad (9)$$

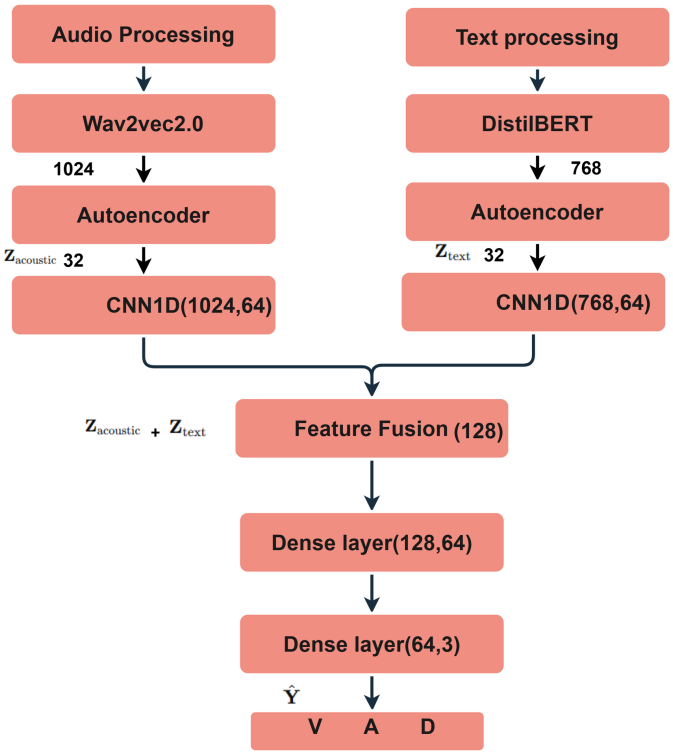

Fig. 1. MER System: Early Fusion

where: $\hat{Y}$ represents the predicted valence, arousal, and dominance scores and FC is the fully connected layers.

- Late Fusion: the features are processed separately before fusion as shown in Fig.2. After applying the autoencoder the speech features Z_{acoustic} go through a CNN1D Layer followed by linear layers. Similarly, the text features Z_{text} undergo the same processing. The predicted values from both modalities (speech and text) are then concatenated and passed through dense layers for final prediction. Formally, the late fusion can be written as:
For speech:

$$\hat{Y}_{\text{acoustic}} = \text{Dense}(\text{ReLU}(\text{Linear}(Z_{\text{acoustic}})))$$

For text:

$$\hat{Y}_{\text{text}} = \text{Dense}(\text{ReLU}(\text{Linear}(Z_{\text{text}})))$$

Predictions obtained from Speech and text are then concatenated as follow:

$$P_{\text{fusion}} = \text{Concat}(\hat{Y}_{\text{acoustic}}, \hat{Y}_{\text{text}}), \quad P_{\text{fusion}} \in \mathbb{R}^{B\times 6}$$

Where P_{fusion} represents the concatenated output from speech and text.
Finally, pass through two dense layers for final prediction:

$$\hat{Y}_{\text{final}} = \text{Dense}_2(\text{ReLU}(\text{Dense}_1(P_{\text{fusion}})))$$

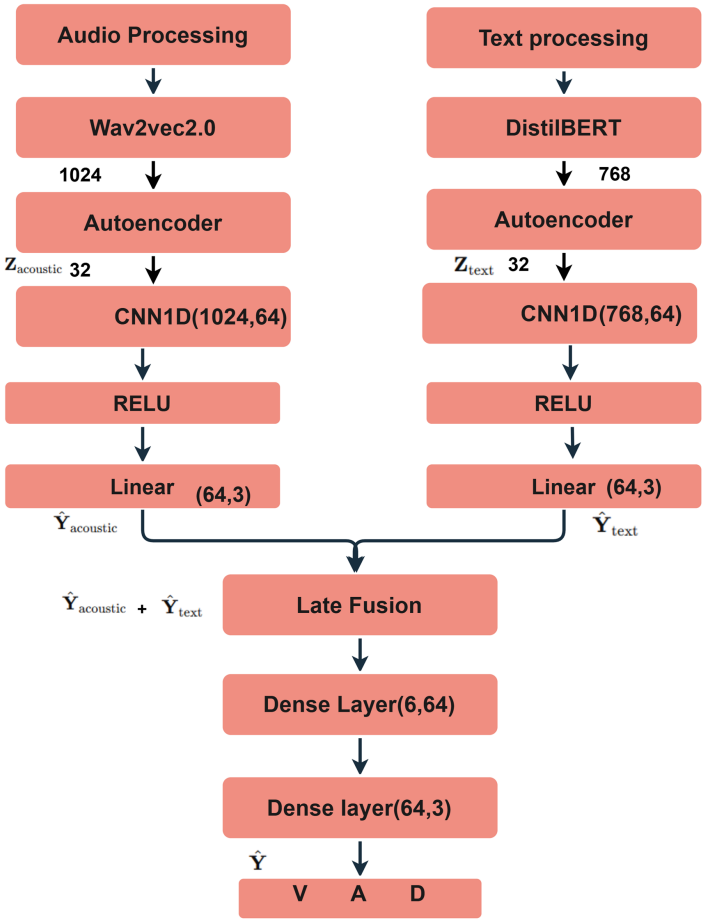

Fig. 2. MER system: Late Fusion

Where:
- Dense_1 and Dense_2 are the two fully connected layers in the late fusion architecture.
- $\hat{Y}_{\text{final}}$ is the predicted value for VAD emotional state.

III. Experimental Setup

We evaluated this framework on the IEMOCAP dataset, a benchmark corpus for emotion recognition.

A. IEMOCAP Database and Data Splitting

IEMOCAP database is utilized [13]. This database comprises approximately 12 hours of data, encompassing 10,039 utterances, all of which are included in our analysis. Although the database includes measurements of speech, facial expressions, head, and affective dyadic session movements, our study focuses solely on speech data. Emotional labels within continuous space are attributed dimensions of valence, arousal, and dominance. We adopt a speaker-independent split, reserving Session 5 for testing. The remaining sessions are divided into 90% for training and 10% for validation.

B. preprocessing data

The preprocessing pipeline is specifically designed for the structure of the IEMOCAP data set, handling audio and textual modalities. Audio processing begins by loading WAV files and resampling them to 16kHz for consistency. Textual content undergoes tokenization using DistilBERT's tokenizer with a fixed 128-token length. The VAD attributes are assigned within the range of [1,5]. Following the approach of previous work [2], these labels fall into the range of [-1,1] which is influenced by the correspondence between the VAD model and the discrete model of emotions proposed by Russell and Mehrabian [15].

C. Hyperparameters

The model components (Wav2Vec2.0, DistilBERT, autoencoder, and fusion layers) are trained end-to-end using the Adam optimizer. Table I details the hyperparameters. Wav2Vec2.0 uses a higher learning rate due to its larger architecture, while DistilBERT uses a smaller rate to avoid overfitting.

TABLE I
HYPERPARAMETERS OF PIPELINE TRAINING.

	Wav2vec2.0	**DistilBERT**
Number of layer/encoder	24 encoders	6layers
Number of units	1024	768
Output Activation	GELU	GELU
learning Rate	5e-4	5e-5
Batch Size	8	8
Maximum Epochs	5	5
Optimizer	Adam	Adam

Wav2Vec2 uses a higher learning rate to adapt its larger pre-trained architecture [21], while DistilBERT's smaller size benefits from a lower rate to avoid overfitting [22].

D. Evaluation Metrics:

In the field of affective computing, CCC is the selected metric to evaluate the performance of dimensional emotion recognition [17]. This metric indicates the agreement between the predicted and the ground truth attribute scores for Iemocap dataset. Denoting the mean and variance of the ground truth by μ_g, σ_g^2 and the mean and variance of the predicted scores by μ_p, σ_p^2, ρ is the Pearson correlation coefficient. The CCC is defined as:

$$\text{CCC} = \frac{2\rho\sigma_g\sigma_p}{\sigma_g^2 + \sigma_p^2 + (\mu_g - \mu_p)^2} \tag{10}$$

CCC is considered superior to Pearson correlation because it penalizes deviations in scale [17], and offers a quantitative measure of the model's prediction performance in relation to actual emotional dimensions.

IV. Results and Discussion

Table II displays the CCC scores for VAD emotional states, from different methods.

TABLE II
SPEAKER-INDEPENDENT EVALUATION RESULTS ON IEMOCAP: IMPACT OF FEATURE FUSION AND DIMENSIONALITY REDUCTION ON SER PERFORMANCE.

Model	**Valence (CCC)**	**Arousal (CCC)**	**Dominance (CCC)**
DistilBERT	0.685	0.453	0.485
Wav2Vec2	0.475	0.708	0.456
Early Fusion	0.732	0.730	0.610
Late Fusion	**0.808**	0.719	**0.635**

In this study, we evaluated two fusion strategies for predicting VAD by integrating DistilBERT for text and Wav2Vec 2.0 for speech.

Table II presents the results of our experiments, highlighting the impact of early fusion and late fusion compared to unimodal approaches. Our baseline models include DistilBERT, which achieved CCC scores of 0.685 for valence, 0.453 for arousal, and 0.456 for dominance, and Wav2Vec2 with CCC scores of 0.475 for valence, 0.708 for arousal, and 0.456 for dominance. Our results demonstrate that Wav2Vec2 with its strong ability to capture audio features, outperforms in predicting arousal compared to valence and dominance. In contrast, DistilBERT excels at predicting valence, which aligns with its strong performance in capturing semantic and contextual information from text. These findings highlight the complementary nature of speech and text features for emotion recognition, suggesting that combining these two signals may further improve overall performance. While early Fusion approach improved CCC scores to 0.732 for valence, 0.730 for arousal, and 0.610 for dominance, showing a notable gain in valence prediction, further improvements were achieved through late fusion, where predictions from both modalities are combined at a decision level. This method led to the best performance, with CCC scores of 0.808 for valence, 0.719 for arousal, and 0.635 for dominance. These findings underscore the effectiveness of multimodal fusion strategies, where late fusion outperforms early fusion in capturing emotional cues, especially for valence and dominance prediction. The autoencoder's dimensionality reduction improved fusion efficiency by removing redundant features and reducing overfitting. To assess the performance of these models, a cross-comparison was performed with previous results published on the IEMOCAP dataset, as shown in Table III. Compared with [11], our model enhances the CCC scores of valence by 13.17%, also the CCC scores of Arousal by 12.52% and the value of CCC dominance by 10.43%.

V. Conclusion

In this study, we propose a novel approach to fuse deep speech embeddings from Wav2Vec 2.0 with textual representations from DistilBERT to recognize emotions in continuous space. To address the high dimensionality of the combined features, we incorporate an autoencoder for dimensionality reduction for both early and late fusion. Evaluated on the IEMOCAP dataset, our approach achieved a competitive CCC

TABLE III
COMPARISON OF PREVIOUS WORKS ON IEMOCAP WITH OUR PROPOSED APPROACH.

Work	Models	Fusion	V	A	D
[14]	LSTM + CNN1D	-	0.603	0.736	0.606
[10]	HuBERT + BERT	-	0.582	0.667	0.545
[11]	CNN14 + BERT	-	0.714	0.639	0.575
[12]	Wav2Vec2-b + BERT-b	Early Fusion	0.625	0.661	0.570
Our Work	Wav2Vec2-l + DistilBERT	Early Fusion	**0.732**	**0.730**	**0.610**
		Late Fusion	**0.808**	**0.719**	**0.635**

scores of 0.808 for valence, 0.719 for arousal, and 0.635 for dominance using late fusion. These results emphasize the effectiveness of Transformer-based models and dimensionality reduction via autoencoders in capturing nuanced emotional cues, especially for valence and dominance. More sophisticated fusion techniques, which can dynamically model interactions between modalities like hybrid fusion and cross-modal attention mechanisms will be explored in future work. Additionally, we aim to evaluate the generalizability of our proposed system on larger and more diverse datasets, including MSP-Podcast.

REFERENCES

[1] ZHAO, Z., WANG, Y., et WANG, Y. Multi-level fusion of wav2vec 2.0 and BERT for multimodal emotion recognition. arXiv 2022. arXiv preprint arXiv:2207.04697.

[2] Messaoudi, Awatef, Hayet Boughrara, and Zied Lachiri. "Speech emotion recognition in continuous space using IEMOCAP database." 2024 IEEE 12th International Symposium on Signal, Image, Video and Communications (ISIVC). IEEE, 2024.

[3] Messaoudi, Awatef, Hayet Boughrara, and Zied Lachiri. "Modeling Continuous Emotions in Text Data using IEMOCAP Database." 2024 IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP). Vol. 1. IEEE, 2024.

[4] Aayushi Chaudhari, Chintan Bhatt, Achyut Krishna, Sai Adiraju, Pier Luigi Mazzeo, "ViTFER: Facial Emotion Recognition with Vision Transformers," Applied System Innovation 5(4):80,August 2022.

[5] Oana Mitruț, Gabriela Moise, Livia Petrescu, Marius Leordeanu, Alin Moldoveanu, Florica Moldoveanu, "Emotion Classification Based on Biophysical Signals and Machine Learning Techniques," Symmetry 12(1):21, December 2019.

[6] Sun, D., He, Y., and Han, J. (2023, June). Using auxiliary tasks in multimodal fusion of wav2vec 2.0 and bert for multimodal emotion recognition. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[7] Macary, M., Tahon, M., Estève, Y., and Rousseau, A. (2021, January). On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition. In 2021 IEEE Spoken Language Technology Workshop (SLT) (pp. 373-380). IEEE.

[8] Wagner, Johannes, et al. "Dawn of the transformer era in speech emotion recognition: closing the valence gap." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.9 (2023): 10745-10759.

[9] Acheampong, Francisca Adoma, Henry Nunoo-Mensah, and Wenyu Chen. "Transformer models for text-based emotion detection: a review of BERT-based approaches." Artificial Intelligence Review 54.8 (2021): 5789-5829.

[10] Srinivasan, Sundararajan, Zhaocheng Huang, and Katrin Kirchhoff. "Representation learning through cross-modal conditional teacher-student training for speech emotion recognition." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[11] Triantafyllopoulos, Andreas, et al. "Multistage linguistic conditioning of convolutional layers for speech emotion recognition." Frontiers in Computer Science 5 (2023): 1072479.

[12] Zhang, Enshi, Rafael Trujillo, and Christian Poellabauer. "The MERSA Dataset and a Transformer-Based Approach for Speech Emotion Recognition." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024.

[13] Busso, Carlos, et al. "IEMOCAP: Interactive emotional dyadic motion capture database." Language resources and evaluation 42 (2008): 335-359.

[14] Awatef, Messaoudi, Boughrara Hayet, and Lachiri Zied. "Multimodal emotion recognition: integrating speech and text for improved valence, arousal, and dominance prediction." Annals of Telecommunications (2025): 1-15.

[15] Bakker, Iris, et al. "Pleasure, arousal, dominance: Mehrabian and Russell revisited." Current psychology 33 (2014): 405-421.

[16] Chen, Li-Wei, and Alexander Rudnicky. "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

[17] de Oliveira, Danilo, Navin Raj Prabhu, and Timo Gerkmann. "Leveraging semantic information for efficient self-supervised emotion recognition with audio-textual distilled models." arXiv preprint arXiv:2305.19184 (2023).

[18] Michelucci, Umberto. "An introduction to autoencoders." arXiv preprint arXiv:2201.03898 (2022).

[19] Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019). recognition," Journal of Physics Conference Series 1896(1):012004, April 2021.

[20] Pepino, Leonardo, Pablo Riera, and Luciana Ferrer. "Emotion recognition from speech using wav2vec 2.0 embeddings." arXiv preprint arXiv:2104.03502 (2021).

[21] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in neural information processing systems 33 (2020): 12449-12460.

[22] Kenton, Jacob Devlin Ming-Wei Chang, and Lee Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." Proceedings of naacL-HLT. Vol. 1. No. 2. 2019.

[23] Hallmen, Tobias, et al. "Unimodal multi-task fusion for emotional mimicry intensity prediction." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[24] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.