Speech Emotion Recognition Using Wav2Vec2 Pre-Trained Model

Dhruvil Dholariya, Devam Shah, Yaksh Desai

Abstract

The data sets for speech emotion recognition are extremely small, and it makes the use of different deep learning approaches very difficult. With the help of this paper, the new approach of transfer learning method has proposed for speech emotion recognition, where we have extracted features from pre-trained wav2vec 2.0 models and they are modeled using convolutional neural network. We have combined the output of different transformer blocks from the pre-trained wav2vec 2.0 model using weighted average for each block. Further, we performance of 2 different approaches, one is MFCC features and the other is wav2vec 2.0 features are compared for speech recognition. We evaluate our proposed approaches on four standard emotion databases RAVDESS, TESS, SAVEE and CREMA-D.

Index Terms: speech emotion recognition, transfer learning, wav2vec 2.0

Introduction

In recent years, more and more people are communicating with voice assistants such as Siri, Alexa, Cortana, and Google Assistant. Communication that ignores the emotional state of the user or fails to express appropriate emotions can seriously impair performance and jeopardize the perceived coldness, social ignorance, dishonesty, and poor performance [1]. As a result, speech recognition will become a more appropriate activity. Many databases have been developed over the years to train and evaluate sensory recognition models, including SAVEE [2], RAVDESS [3], EMODB [4], IEMOCAP [5], and MSP-Podcast [6]. With the exception of the MSP-Podcast, these data sets are small in size, usually including only a few speakers. In terms of modelling techniques, many traditional methods have been proposed, using Markov (HMMs) [7, 8], support vector systems (SVMs) [9, 10], decision trees [11, 12], and, more recently, deep neural networks (DNNs) [13, 14, 15]. However, while DNNs have shown significant gains in traditional methods in tasks such as automatic speech recognition (ASR) [16] and speaker recognition [17], the perceived benefits of emotional recognition are limited, perhaps due to the small size of the databases. A common strategy when working with small training databases is to apply transfer learning strategies. The transfer learning technique is to use a model that is useful for a specific task where large data sets are available to train to strengthen the intensity of the interest task the data is scarce in it. The model learned in the auxiliary work can be used as a feature identifier or a well-organized, after adding some of its final layers, to the work of interest. Recently, transfer learning methods have been explored in the field of speech recognition. In [18], a deep transformer [19] based neural network was first developed for LibriSpeech [20] using a number of objectives monitored on various time scales. After that, the model is used as a feature extractor or fine-tuned to detect speech emotions, among other downstream tasks. Similarly, in [21] and [22], they pre-train a deep encoder in a contrastive predictive coding task, and the embeddings of the results are used in datasets of speech emotion. Recently, several automated speech recognition (ASR) models have been developed using self-regulating self-control, including wav2vec [23] and VQ-wav2vec [24]. A few recent studies [25, 26, 27] have successfully used presentations from these genres as aspects of emotional recognition.

In line with those works, in this paper, we explore the use of the wav2vec 2.0 model [28], an improved version of the original wav2vec model, as a feature extractor for speech emotion recognition. The main contributions of our paper are use MFCC and Log-Mel spectrogram as features and feature extraction from weighted average of all layers of Wav2Vec 2.0 model and then feed features into CNN network and classify emotion.

Methods

In our study, we extracted features from two released wav2vec 2.0 models and used them for speech emotion recognition. In this section, we describe the wav2vec 2.0 model, the datasets used for training and evaluation, and the downstream models.

Wav2vec 2.0 model architecture

Wav2vec 2.0 [28], is a framework for self-supervised learning of representations from raw audio. The model consists of three stages. The first stage is a **local encoder**, which contains several convolutional blocks and encodes the raw audio into a sequence of embeddings with a stride of 20 ms and a receptive field of 25 ms. Two models have been released for public use, a large one and a base one where the embeddings are 1024- and 768-dimensional, respectively. The second stage is a **contextualized encoder**, which takes the local encoder representations as input. Its architecture consists of several transformer encoder blocks [19], The base model uses 12 transformer blocks with 8 attention heads each, while the large model uses 24 transformer blocks with 16 attention heads each. Finally, a **quantization module**, takes the local encoder representations as input and consists of 2 codebooks with 320 entries each. A linear map is used to turn the local encoder representations into logits. Given the logits, Gumbel-Softmax [29], is applied to sample from each codebook. The selected codes are concatenated and a linear transformation is applied to the resulting vector leading to a quantized representation of the local encoder output, which is used in the objective function, as explained below.

Wav2Vec 2.0 pretraining and finetuning

The wav2vec 2.0 model is pretrained in a self-supervised set- ting, similar to the masked language modelling used in BERT [30] for NLP. Contiguous time steps from the local encoder representations are randomly masked and the model is trained to reproduce the quantized local encoder representations for those masked frames at the output of the contextualized encoder.

The training objective is composed by terms of the form

$$L_m = -log \frac{\exp\left(\frac{sim(c_t, q_t)}{k}\right)}{\sum_{\tilde{q} \in \tilde{Q}} \exp\left(\frac{sim(c_t, q_t)}{k}\right)}$$

Where $sim(c_t, q_t)$ is the cosine distance between the contextualized encoder outputs and the quantized local encoder representations *qt*. *t* is the time step, *κ* is the temperature and Q˜ is

the union of a set of K *distractors* and *qt*. The distractors are outputs of the local encoder sampled from masked frames be- longing to the same utterance as *qt*. The contrastive loss is then given by *Lm* summed over all masked frames. Finally, terms to encourage diversity of the codebooks and L2 regularization are added to the contrastive loss.

The main goal of the wav2vec 2.0 paper was to use the learned representations to improve ASR performance, requiring less data for training and enabling its use for low resource languages. To this end, the model trained as described above is finetuned for ASR using a labelled speech corpus like LibriSpeech. A randomly initialized linear projection is added at the output of the contextual encoder and the connectionist temporal classification (CTC) loss [31] is minimized. The models finetuned in 960 hours of LibriSpeech reach state of the art results in automatic speech recognition when evaluated in different subsets of LibriSpeech. Even when finetuning using considerably less hours, wav2vec 2.0 models reach a performance comparable to the state of the art.

In this paper we compared the performance in speech emotion recognition when using both the wav2vec 2.0 base model pretrained in Librispeech without finetuning[1] (we will call this model Wav2vec2-PT), and a model finetuned for ASR using a 960-hour subset of Librispeech[2] (Wav2vec2-

FT). In both cases, we used the base model as we did not see significant performance improvements when using the large one, and it allowed us to reduce computational requirements.

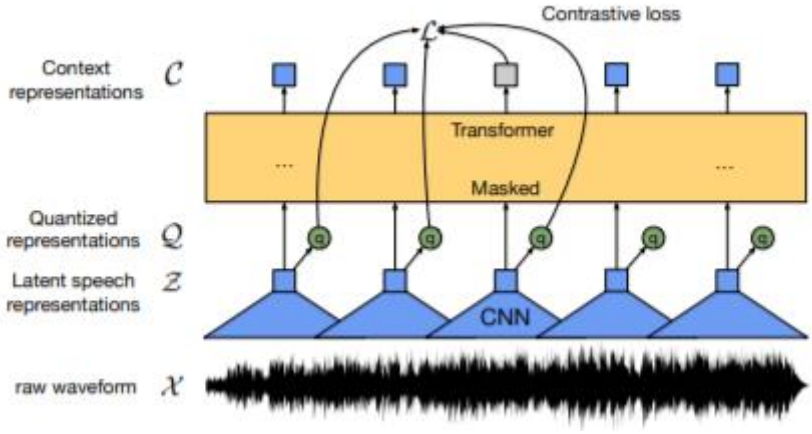

Fig 2:Wav2Vec 2.0 model representation

Features

For any classification or other task on audio, first we need features which represent audio in better way. And then we perform our desire task on that features. We used different types of four features then perform classification and compare results which are shown in Table 1. Once we extract features, then we provide this features as input in downstream model. For all four methods our downstream model remains same. Which we will explain in next section. In this section we will talk about features and feature extraction.

In first method we use MFCC as features to represent audio. You will find more information on MFCC in this paper [32]. MFCC is known to be a good feature to represent an audio. There are many different ways you can slice MFCC feature. It stands for Mel-frequency cepstral coefficient, and it is a good representation of the vocal tract that produces the sound. In most and common machine learning task we treats MFCC as an image and it is become feature. Main advantage of treating it as an image is, it provides more information, and gives one the ability to draw on transfer learning. This is certainly legit and yields good accuracy. However, research has also shown that statistics relating to MFCCs can carry good amount of information as well.

The mel spectrogram [33] can be thought of as a visual representation of an audio signal. Specifically, it represents how the spectrum of frequencies vary over. The Fourier transform is a mathematical formula that allows us to convert an audio signal into the frequency domain. It gives the amplitude at each frequency, and we call this the spectrum. Since frequency content typically varies over time, we perform the Fourier transform on overlapping windowed segments of the signal to get a visual of the spectrum of frequencies over time. This is called the spectrogram. Finally, since humans do not perceive frequency on a linear scale, we map the frequencies to the mel scale (a measure of pitch), which makes it so that equal distances in pitch sound equally distant to the human ear. What we get is the mel spectrogram.

Wav2vec 2.0 is a best model for automatic speech recognition. So we decided to use pre trained model as feature extractor. As we discussed Wav2vec 2.0 model consists of three stages local encoder, contextualized encoder, and quantization module. At first we take output of 12^{th} transformer block from contextualized encoder stage. And use this as our features for classification. It gives very low accuracy.

So from paper [34] we got an idea of use a weighted average of the outputs of the 12 transformer blocks, along with those of the local encoder and use this as a features for classification. All features were normalized by subtracting the mean and dividing by the standard deviation of that feature over all the data for the corresponding speaker. We didn't take global normalization per feature because it didn't work well as explained in this paper [34].

Downstream models

One audio is converted into size of (30,216) MFCC feature. Mel spectrogram feature size is also (30,216).
For one audio each Wav2Vec 2.0 layer gives output of size (124,768). In Wav2Vec 2.0 last layer method we directly feed last layer out to the downstream model. In case of all layer we get feature size (13,124,768) then we take weighted average on 1st dimension and we get feature of size (124,768) and now we feed this in downstream model. We take weights from this paper [34] for weighted average of all layers.
We applied same downstream model to all four features. Only difference is in input shape. For first two features input shape is (30,216,1). And for remaining two it is (124,768,1). Then we have applied 4 CNN blocks of the same architecture. In this block the first Conv2D of 32 filters having kernel size = (4,10) is applied while the padding value is set as the same. So there will be no change in shape. Then Batch Normalization and Relu activation is applied on the same. Then MaxPooling of (2,2) is applied so the new shape will be half of the old one and then drop out of 0.2. After these four blocks the final shape will be (7,48,32) and then it has got flattened and it will be of (10752). Then there is a dense layer of 64 neurons followed by drop out, batch normalization and relu activation and then the final output layer will be of 14 neurons with softmax activation.

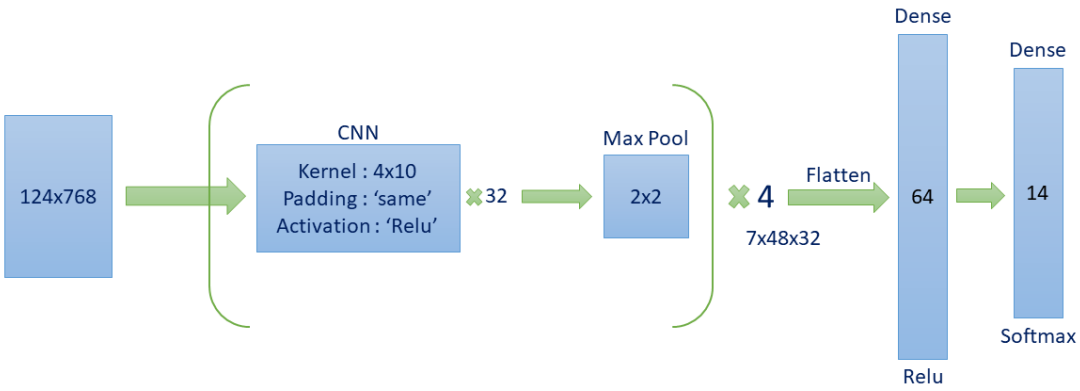

Fig 2: Figure of downstream model

Datasets

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [3]is a multi-modal database of emotional speech and song. It features 24 different actors (12 males and 12 females) enacting 2 statements: "Kids are talking by the door" and "Dogs are sitting by the door." with 8 different emotions: happy, sad, angry, fearful, surprise, disgust, calm, and neutral. These emotions are expressed in two different intensities: nor- mal and strong, except for neutral (normal only). Each of

the combinations was spoken and sung, and repeated 2 times, leading to 104 unique vocalizations per actor. Following [35], we merged the neutral and calm emotions, resulting in 7 emotions, and used the first 20 actors for training, actors 20-22 for validation to do early stopping, and actors 22-24 for test.

Surrey Audio-Visual Expressed Emotion (SAVEE) database [2] has been recorded as a pre-requisite for the development of an automatic emotion recognition system. The database consists of recordings from 4 male actors in 7 different emotions, 480 British English utterances in total. The sentences were chosen from the standard TIMIT corpus and phonetically-balanced for each emotion. The data were recorded in a visual media lab with high quality audio-visual equipment, processed and labeled. To check the quality of performance, the recordings were evaluated by 10 subjects under audio, visual and audio-visual conditions. Classification systems were built using standard features and classifiers for each of the audio, visual and audio-visual modalities, and speaker-independent recognition rates of 61%, 65% and 84% achieved respectively.

TESS stimuli were modeled on the Northwestern University Auditory Test No. 6 (NU-6; Tillman & Carhart, 1966) [36]. A set of 200 target words were spoken in the carrier phrase "Say the word _____" by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 stimuli in total. The two actresses were recruited from the Toronto area. Both actresses speak English as their first language, are university educated, and have musical training. Audiometric testing indicated that both actresses have thresholds within the normal range.

CREMA-D is an audio-visual data set for emotion recognition [37]. The data set consists of facial and vocal emotional expressions in sentences spoken in a range of basic emotional states (happy, sad, anger, fear, disgust, and neutral). 7,442 clips of 91 actors with diverse ethnic backgrounds were collected.

Results and discussion

We create and extract features and then provide it to the downstream model. All results are shown in Table 1 for dataset which contains all four datasets and in Table 2 for only Ravdess dataset.
Results for MFCC and Log-mel spectrogram features are shown in first two row of Table 1. This features are very good representation of audio and gives accuracy almost same. This features are work well for speech emotion recognition. We can treat this two features as an image so here we apply custom 2D CNN model but one can also try transfer learning. This is main advantage of this features that we can use CNN network.

Last two row showing results of feature extraction from pre trained Wav2Vec 2.0 model. First we extract feature from only last layers output which is 12^{th} transformer block of contextualized encoder stage of Wav2Vec 2.0 model. Which gives accuracy 29% which is very low. Because Wav2Vec 2.0 model is for automatic speech recognition so output layer gives features which is relevant to ASR and not related to speech emotion recognition.

And also fine tuning of model not gives better accuracy as experiment in this paper [34]. They hypothesize that reason behind this is, when Wav2vec 2.0 is finetuned for a speech emotion task, information that is not relevant for that task but might be relevant for speech emotion recognition is lost from the embeddings. For example, information about the pitch might not be important for speech recognition, while it is essential for speech emotion recognition.

After when we take weighted average of all layers we got an accuracy 63% which is far better than only extracting features from only last layers.

Features	Accuracy on RAVDESS, TESS, SAVEE, CREMA-D
MFCC	69.44%
Log-mel spectrogram	68%
From Wav2vec 2.0 last layer	29%
From Wav2vec 2.0 all layer	63%

Table 1: Accuracy on RAVDESS, TESS, SAVEE, CREMA-D datasets using different features and feature extractor methods

After that instead of taking all four datasets we trained model only on RAVDESS dataset. In which we try on MFCC feature and feature extraction from Wav2Vec 2.0 all layer and results are shown in Table 2. For MFCC accuracy is 69% and feature extraction from weighted average of all layers of pre trained model gives accuracy around 72%.

Features	Accuracy on only RAVDESS
MFCC	69%
From Wav2vec 2.0 all layer	72%

Table 2: Accuracy on RAVDESS dataset using MFCC features and feature extracting from pre trained Wav2vec 2.0 model

Conclusions

In this work, we explored different ways of extracting and modeling features from pretrained wav2vec 2.0 models for speech emotion recognition. We proposed to combine the different layers in the wav2vec 2.0 model using trainable weights and model the resulting features with a simple DNN with a time-wise pooling layer. We evaluated our models on two standard emotion datasets, IEMOCAP and RAVDESS, and showed superior results on both cases, compared to those in recent literature. We found that the combination of information from different layers in the wav2vec 2.0 model led to improved results over using only the encoder outputs, as in previous works. Further, we found that the combination of the wav2vec 2.0 features with a set of prosodic features gave additional gains, suggesting that the wav2vec 2.0 model does not contain all the prosodic in- formation needed for emotion recognition, so we need all the output features of every block of wav2vec 2.0.

References

[1] S. Brave and C. Nass, “Emotion in human–computer interaction,” in *Human-computer interaction fundamentals*. CRC Press Boca Raton, FL, USA, 2009, vol. 20094635.

[2] S. Haq and P. Jackson, “Machine Audition: Principles, Algo- rithms and Systems,” W. Wang, Ed. IGI Global, 2010.

[3] S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dy- namic, multimodal set of facial and vocal expressions in North American English,” *PLOS ONE*, vol. 13, no. 5, 2018.

[4] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech,” in *Ninth Eu- ropean Conference on Speech Communication and Technology*, 2005.

[5] 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, no. 4, 2008.

[6] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, 2019.

[7] B. Schuller, G. Rigoll, and M. Lang, *Hidden Markov Model-based Speech Emotion Recognition*, 2003, vol. 2.

[8] N. Sato and Y. Obuchi, "Emotion recognition using mel- frequency cepstral coefficients," *Information and Media Tech- nologies*, vol. 2, no. 3, 2007.

[9] P. Shen, Z. Changjun, and X. Chen, "Automatic speech emotion recognition using support vector machine," in *Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology*, vol. 2. IEEE, 2011.

[10] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, "Emotion recognition from speech using global and local prosodic features," *In- ternational journal of speech technology*, vol. 16, no. 2, 2013.

[11] M. Borchert and A. Dusterhoft, "Emotions in speech-experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments," in *International Conference on Natural Language Processing and Knowledge Engineering*. IEEE, 2005.

[12] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, 2011.

[13] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[14] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms." in *Inter- speech*, 2017.

[15] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, and N. Dehak, "Emotion identification from raw speech signals using dnns." in *Interspeech*, 2018.

[16] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han,S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution- augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[17] P. Mateˇjka, O. Glembek, O. Novotny`, O. Plchot, F. Gre´zl, L. Bur- get, and J. H. Cernocky`, "Analysis of dnn approaches to speaker identification," in *ICASSP*. IEEE, 2016.

[18] Y. Zhao, D. Yin, C. Luo, Z. Zhao, C. Tang, W. Zeng, and Z.-J. Zha, "General-Purpose Speech Representation Learn- ing through a Self-Supervised Multi-Granularity Framework," *arXiv:2102.01930*, 2021.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

[21] M. Li, B. Yang, J. Levy, A. Stolcke, V. Rozgic, S. Matsoukas, C. Papayiannis, D. Bone, and C. Wang, "Contrastive Unsupervised Learning for Speech Emotion Recognition," arXiv:2102.06357, 2021.

[22] R. Zhang, H. Wu, W. Li, D. Jiang, W. Zou, and X. Li, "Trans- former based unsupervised pre-training for acoustic representation learning," arXiv:2007.14602, 2021.

[23] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-training for Speech Recognition," arXiv:1904.05862, 2019.

[24] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self- supervised learning of discrete speech representations," in Inter- national Conference on Learning Representations, 2020.

[25] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly Fine-Tuning "BERT-like" Self-Supervised Models to Improve Multimodal Speech Emotion Recognition," arXiv:2008.06682, 2020.

[26] M. Macary, M. Tahon, Y. Este`ve, and A. Rousseau, "On the use of Self-supervised Pre-trained Acoustic and Linguistic Features for Continuous Speech Emotion Recognition," arXiv:2011.09212, 2020.

[27] J. Boigne, B. Liyanage, and T. O¨ strem, "Recognizing More Emotions with Less Data Using Self-supervised Transfer Learning," arXiv:2011.05585, 2020.

[28] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," arXiv:2006.11477, 2020.

[29] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," arXiv preprint arXiv:1611.01144, 2016.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre- training of deep bidirectional transformers for language under- standing," arXiv preprint arXiv:1810.04805, 2018.

[31] A. Graves, S. Ferna´ndez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in ICML, 2006.

[32] Likitha, M. S., et al. "Speech based human emotion recognition using MFCC." *2017 international conference on wireless communications, signal processing and networking (WiSPNET)*. IEEE, 2017.

[33] H. Meng, T. Yan, F. Yuan and H. Wei, "Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network," in IEEE Access, vol. 7, pp. 125868-125881, 2019, doi: 10.1109/ACCESS.2019.2938007.F. Eyben, M. Wo¨llmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in Proceedings of the 18th ACM international conference on Multimedia, 2010.

[34] Pepino, Leonardo, Pablo Riera, and Luciana Ferrer. "Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings." *arXiv preprint arXiv:2104.03502* (2021). Kate Dupuis, M. Kathleen Pichora-Fuller, "Toronto emotional speech set (TESS) Collection" arXiv:2102.10326.

[35] 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.