Improved speech emotion recognition based on music-related audio features

Linh Vu
Monash University
linh.vu@monash.edu

Raphaël C.-W. Phan
Monash University
raphael.phan@monash.edu

Lim Wern Han
Monash University
lim.wern.han@monash.edu

Dinh Phung
Monash University
dinh.phung@monash.edu

***Abstract*—Emotions are essential for human communication as they reflect our inner states and influence our actions. Today, emotions provide crucial information to many applications, from virtual assistants to security systems, mood-tracking wearable devices, and autism robots. The speech emotion recognition (SER) model must be lightweight to run on varying devices with limited computational power. This research investigates the performance of music-related features for SER based on the auditory and neuropsychology evidence about the connection of emotional speech and music in human perception. Unlike prior works on low-level descriptors that primarily focus on differentiating human speech production, our method employs features extracted directly from raw speech signals through Discrete Fourier Transform and Constant-Q Transform. These features represent the perceptual pitches and timbre characteristics of the human voice. The 10-fold cross-validation results show that our method improves the accuracy of the audio feature-based approach on RAVDESS, CREMA-D and IEMOCAP datasets. Findings from the ablation study imply the significance of perceptual pitch, the perceptual loudness and the combination of pitch and timbre features in building a robust SER model. Compared to pretrained deep learning embeddings, our method demonstrates its generalizability and high efficiency despite a much smaller model size.**

***Index Terms*—speech emotion recognition, audio features, music features, LLD, MFCC, CQT, Mel spectrogram**

I. Introduction

Speech emotion recognition (SER) has been an active research area in Human-Computer Interaction for more than two decades. Many recent studies in SER borrow large pretrained models from Computer Vision (CV), and Natural Language Processing (NLP), which can range from hundreds to thousands of megabytes in size and be computationally expensive [1]–[4]. In terms of audio-based methods, the most popular approach relies on an excessive amount of hand-crafted features designed for general speech recognition tasks, including speech-to-text and speaker recognition. They are called Low-level descriptors (LLD) that include speech production-related features (e.g. energy, fundamental frequency, voice probability, harmony, jitter, shimmer, ...), and their derived values (e.g. max, min, mean, standard deviation, kurtosis, skewness, regression coefficients, ...). Notable ones are ComParE [5], GeMAPS and eGeMAPS [6]. ComParE [5] is the largest LLD set, which is also the most popular audio feature set for many speech classification tasks. Eyben et al. proposed GeMAPS and its extended version eGeMAPS as minimalistic acoustic sets for affective computing [6]. They achieve 76% to 80% UAR (Unweighted Average Recall) for arousal (intensity) regression, but only 64% to 68% for valence (positivity) regression [6]. These results align with insights from Sezgin et al. [7] where features based on the production of emotional speech are highly correlated to arousal. However, the author note these features have a low correlation with valence; hence they cannot help distinguish basic emotions.

To improve the ability of machines to mimic the human sense of emotions, SER models must learn from features that represent our perception of emotional speech. According to psychology studies [8], there is a strong link between emotional speech and music. Music is a more expressive form of emotional speech. Neuropsychology research has shown that people with music-specific disorders such as amusia have speech-based emotion recognition impairments [9]. This inspires us to explore features used in music recognition that capture the intonation characteristics of voice for emotion recognition, most notably pitch and timbre related features.

According to auditory research from Oxenham [10], pitch is a perceptual quality of sound that is ordered on the scale used for melody in music (e.g. C1, C1#, D1, ...). The absolute pitch spectrum plays an essential role in auditory perception that not only provides semantic and non-semantic information of speech but also helps us listen in a noisy environment. Oxenham mentioned that one of the key differences between human hearing and other species is the tendency to focus on relative pitch relations (i.e. 12 pitch classes, or chroma). To extract these features from raw audio signals, we use Constant-Q Transform (CQT). The CQT is well motivated by the musical theory that the fundamental frequencies F0 of the tones in Western music are geometrically spaced. Thus, the CQT spectrogram represents the absolute pitch spectrum of sound. CQT spectrogram provides the benchmark results for classification of music [11] and detection of spoofing attacks [12]. From the CQT spectrogram, we compute the chroma-based spectrogram (i.e. chromagram) by summing up all pitch coefficients that belong to the same pitch class $\{C, C^{\#}, D, D^{\#}, ..., B\}$.

Pitch and timbre have strong interaction and interference in the human perception of sound [10]. When listening to different instruments playing the same pitch note, our ears rely on timbre to distinguish them. This characteristic of sound is best described using Mel-frequency cepstral coefficients (MFCC) [13]. This feature captures the spectral envelope and the periodic structure of the Mel spectrogram. MFCC is generated using the Discrete Cosine Transform (DCT) on a Mel spectrogram; which is obtained from applying DFT and a Mel filterbank on raw speech signals. As DFT uses linearly spaced frequency bins (i.e. a constant bin size), the Mel spectrogram is faster to compute but lacks detailed information in the audible frequencies than the CQT spectrogram.

II. Methodology

In this research, we focus on analyzing audio features to improve SER. We propose new combinations of music-related features that best describe the pitch and timbre characteristics in distinguishing emotional speech. We extract and evaluate them against popular acoustic features and pretrained deep neural network embeddings.

A. Extracting music-related features for SER

We propose different combinations of the pitch-related features (CQT spectrogram, Chromagram) and the timbre-related features (MFCC, Mel spectrogram) for SER. They are called MuSER (Music-related features for Speech Emotion Recognition), the extended version eMuSER, and two variants MuSERn and eMuSERn. The details are included in the Table I.

We use *nnAudio* [14] for DFT and CQT transformations. Before transforming, each raw audio sample is resampled to 16000Hz and segmented into chunks of samples by a window function. With a window size of 2048 samples for a sampling rate of 16000, each chunk is 128-millisecond long, which is also the average length of an English word. With DFT, we apply the Mel filterbanks with 128 frequency bins on the DFT spectrogram to extract 128 Mel spectrogram, which is then converted to log-power spectrum and used DCT type II to get 128 MFCCs. With CQT, we apply the Chroma filterbanks to extract 12 Chromagram features. Since Mel spectrogram and CQT spectrogram features have log-normal distribution, we convert them to log-power spectrum and apply normalization. Log-power spectrum also represents the human perception of loudness. The total number of extracted features for each audio speech are MuSER $(380, 1)$ and eMuSER $(520, 1)$ after averaging through the time axis.

Recent research in SER employed a similar set of features for SER that includes Mel spectrogram, MFCC, and Chromagram (MMC) [15]. We also extract and evaluate its performance on the three datasets. We use the exact feature resolution described above to have a fair comparison.

In addition, Principal Component Analysis (PCA) was applied to all features to reduce the dimension to 12.5% of the original feature vector size. It often helps reduce the complexity and increase the model performance.

B. Extracting Low-level descriptors sets for SER

To compare with the proposed feature sets, we evaluate common audio features for SER namely GeMAPS, eGeMAPS and ComParE. We used the OpenSmile library [16] to extract them. Table II describes the details of each feature set.

Table I: Music-related audio feature sets

MMC: 268 features [15] Mel spectrogram, MFCC, Chromagram.
MuSER: 380 features Pitch-related features: CQT spectrogram, Timbre-related features: MFCC. **MuSERn: 380 features** Log-CQT spectrogram, MFCC.
eMuSER: 520 features Pitch-related features: CQT spectrogram, Chromagram, Timbre-related features: MFCC, Mel spectrogram. **eMuSERn: 520 features** Log-CQT spectrogram, Chromagram, MFCC, Log-Mel spectrogram.

Table II: Low-level descriptors sets

ComParE: 6373 features [5] Voicing probability, Jitter, F0, MFCC, Spectral roll off, flux, slope, sharpness, loudness, RMS energy, ... and their derived values (deltas and functionals).
GeMAPS: 66 features; eGeMAPS: 88 features [6] Voicing probability, Jitter, Shimmer, F0, Formants, MFCC, loudness, Harmonic-to-noise ratio, Alpha ratio, Energy peak, Spectral flux, slope, ... and their derived values (deltas and functionals).

C. Extracting pretrained speech embeddings for SER

With the advancements in deep learning, many researchers have been trying to create deep neural networks (DNN) to learn the nature of the raw data and project it into a latent space that is most convenient for the model to handle downstream tasks such as classifying or generating new data without relying on feature preprocessing and engineering. The latent space feature vectors extracted from the hidden layers of the DNN, or embeddings, are often pretrained on extensive datasets to achieve high generalizability to deal with unseen data. For emotion classification, there are two popular embeddings, namely TRILL[1] and Wav2vec2-IEMOCAP[2].

TRILL [1] is a pretrained audio speech embedding for non-semantic classification tasks such as speaker identification, language identification, and voice-based medical diagnosis. It uses ResNet-50-based model as a backbone model to learn representation from Mel spectrogram. It was trained using the Triplet loss approach to discriminate same or different audio segment pairs from AudioSet[3] (4971 hours), which is a massive Youtube video dataset. The TRILL feature vector size is 512. **Wav2vec2-PT** [17] is a 768-dimensional embedding of size 768 fine-tuned on IEMOCAP based on Wav2vec2-base, which is the state-of-the-art embedding for Automatic Speech Recognition (ASR) [18]. Unlike TRILL, Wav2vec2-PT takes raw waveform signals as input. The model uses Transformer-based architecture which was first trained on ASR dataset using CTC loss, then fine-tuned on IEMOCAP for emotion classification task using Additive Margin Softmax Loss. Thus, for this feature, we only evaluate the pretrained model with

[1] https://tfhub.dev/google/nonsemantic-speech-benchmark/trill/3

[2] https://huggingface.co/speechbrain/emotion-recognition-wav2vec2-IEMOCAP

[3] https://research.google.com/audioset/

classifier on the whole RAVDESS and CREMA-D datasets. The pretrained ASR dataset is Librispeech (960 hours)[4], which is a huge ASR dataset for self-supervised training.

III. Experiment setup

A. Datasets

We evaluate the proposed method on three datasets that are standard for SER. First, **RAVDESS** [19] consists of very clean and high-quality audio speech performed by 24 North-American actors with eight emotional expressions. The top benchmark of machine learning models for this dataset is from Issa et al. with 71.61% for these eight emotions [20], which employed a 6-layer convolutional neural network to analyze an acoustic feature set includings the MMC set, tonnetz and spectral contrast features. The second dataset, **CREMA-D** [21] is a large speech dataset of six emotions with 7442 recordings from 91 speakers with a variety of races and ethnicities (e.g. African America, Asian, Caucasian, Hispanic, ...). However, its sound quality is lower compared to RAVDESS but is a good representation of real-world audio recordings. The state-of-the-art accuracy is 67.8% for these six emotions from the TRILL model [1]. Third, **IEMOCAP** [22] is a large conversational speech dataset of 12 hours of speech from 10 speakers. In the same paper [20], Issa et al. modified the CNN model and achieved 64.30% accuracy on IEMOCAP dataset for seven emotions. As IEMOCAP is very imbalanced (e.g. 2 samples in the *disgust* class, 40 samples in the *fear* class, 1849 samples in the *frustration* class, ...), we only use four classes of data out of total eight classes namely *happy, sad, angry, neutral*. We also use the same classes for all three datasets to have a fair comparison. We filter out speech samples more than 10-second long due to the limitation of the TRILL model. Table III describes the number of samples in each emotion class.

Table III: Number of samples used for each emotion class.

Dataset	RAVDESS	CREMA-D	IEMOCAP
Happy	192	1271	1041
Sad	192	1271	1084
Angry	192	1271	1103
Neutral	96	1087	1708

B. Classifier and evaluation

For the classifier, we use a dense neural network with the ReLU activation function, Adam optimizer and Cross Entropy loss. The network consists of two layers with 500 nodes in each layer.

To evaluate the performance of the proposed models, we split each dataset into two sets: 80% of the data for the training set and 20% for the test set using stratified splitting to ensure the proportion of each class in both sets. For training, we apply 10-fold cross-validation to get the best model to evaluate on the test set. To measure the performance on the test set, we report three different metrics: unweighted accuracy (UA, or micro-F1), unweighted-F1 (UF1, or macro-F1) and weighted-F1 (WF1).

IV. Results and discussions

A. Feature sets comparison

In the first experiment, we extracted eight feature sets and two embeddings; training the classification model separately on each dataset RAVDESS, CREMA-D and IEMOCAP. Table IV shows the result of this experiment. As seen in Table IV, the models for TRILL, MMC, and all the proposed feature sets namely MuSER, MuSERn, eMuSER and eMuSERn learn better with the clean audio from the RAVDESS dataset compared to the noisier CREMA-D and IEMOCAP datasets; while the LLD sets seem to no benefit from the increase in data quality.

On the RAVDESS dataset, all the proposed feature sets surpassed all the baseline acoustic feature sets. MuSER-PCA approach yielded the highest accuracy with 86.67% UA, which is equal to the deep learning embedding TRILL despite having a smaller input feature size (380 compared to 512). Among all the baseline approaches, MMC-PCA achieved the best result with 80% UA, which is still 6.67% lower than MuSER-PCA. While on RAVDESS, MuSER-based methods are better than eMuSER-based ones, the latter gave more accurate results on CREMA-D dataset.

On both CREMA-D and IEMOCAP, the TRILL embedding is the most discriminative feature, which gave 75.1% and 71.56% UA, respectively. This is followed by eMuSERn-PCA for CREMA-D with 71.73% UA, and eMuSERn for IEMOCAP with 67.91% UA. On average, TRILL also achieved the best UA with 78% accuracy. eMUSERn-PCA yielded the second-best UA with 75% accuracy, followed by MuSERn, MuSERn-PCA, MuSER-PCA, and eMuSERn with 1 to 2% lower results.

According to SpeechBrain[5], Wav2vec2-PT was trained on IEMOCAP with 5-fold cross-validation setting and reached 75.28% UA. When evaluating this pretrained model on CREMA-D, it gave 50.63% UA and 43.5% WF1. Surprisingly, it gave the lowest WF1 20.28% on the RAVDESS dataset, which is a very clean and high-quality audio dataset. This indicates overfitting and low generalizability. However, to have a fair comparison with these results, we also need to conduct cross-dataset evaluation on other features. We leave this for future work.

Among all the LLD sets, GeMAPS is the best feature set for SER achieving 55.39% and 61.76% WF1 on RAVDESS and CREMA-D respectively, followed closely by eGeMAPS with 1-2% difference. On IEMOCAP, eGeMAPS performed better than GeMAPS with 51.59% versus 49.76% WF1 in particular. ComParE gave the worst performance among all feature sets despite being the most extensive feature. Its WF1 results on RAVDESS, CREMA-D and IEMOCAP are 21.03%, 23.25% and 17.8%, respectively. When projecting the ComParE feature

[4] https://www.openslr.org/12

[5] https://huggingface.co/speechbrain/emotion-recognition-wav2vec2-IEMOCAP

Table IV: Results of feature sets and embeddings on three datasets in percentage (%)

Dataset	**RAVDESS**			**CREMA-D**			**IEMOCAP**			**Average**		
Feature	**UA**	**UF1**	**WF1**	**UA**	**UF1**	**WF1**	**UA**	**UF1**	**WF1**	**UA**	**UF1**	**WF1**
ComParE	31.11	18.37	21.03	30.41	22.94	23.25	34.62	12.86	17.80	32 ±16	18 ±10	21 ±11
ComParE-PCA	30.37	19.32	22.06	31.84	21.64	22.37	34.72	27.26	29.88	32 ±16	23 ±12	25 ±13
eGeMAPS	57.78	47.11	54.02	60.00	58.45	58.57	52.63	50.68	51.59	57 ±29	52 ±26	55 ±28
eGeMAPS-PCA	42.96	39.54	42.71	39.29	38.13	38.56	38.56	35.95	37.30	40 ±20	38 ±19	40 ±20
GeMAPS	57.78	50.42	55.39	62.14	61.56	61.76	51.01	49.11	49.76	57 ±29	54 ±27	56 ±28
GeMAPS-PCA	41.48	38.16	40.27	42.35	41.83	41.90	40.18	38.09	39.25	41 ±21	39 ±20	40 ±20
MMC	79.26	78.15	79.34	65.00	65.01	65.24	67.51	67.30	67.31	71 ±36	70 ±36	71 ±36
MMC-PCA	80.00	80.01	79.64	60.82	59.71	59.99	57.59	57.29	57.38	66 ±35	66 ±34	66 ±34
MuSER	82.22	80.30	81.98	66.53	65.59	65.99	65.69	65.40	65.59	71 ±37	70 ±36	71 ±36
MuSER-PCA	**86.67**	**86.13**	**86.77**	66.63	66.11	66.31	67.00	67.07	67.08	73 ±38	73 ±38	73 ±38
MuSERn	84.44	83.12	84.40	69.29	69.17	69.20	67.71	67.51	67.58	74 ±38	73 ±37	74 ±38
MuSERn-PCA	**86.67**	86.01	86.75	67.04	66.59	66.81	65.69	65.61	65.74	73 ±38	73 ±38	73 ±38
eMuSER	84.44	83.93	84.50	66.63	65.98	66.19	65.18	65.02	65.20	72 ±37	72 ±37	72 ±37
eMuSER-PCA	83.70	82.64	83.77	64.69	64.34	64.53	62.75	62.47	62.93	70 ±36	70 ±36	70 ±36
eMuSERn	82.96	81.26	82.83	67.45	67.01	67.41	67.91	68.06	67.88	73 ±37	72 ±37	73 ±37
eMuSERn-PCA	84.44	84.17	84.52	71.73	71.37	71.57	67.61	67.33	67.57	75 ±38	74 ±38	75 ±38
TRILL	**86.67**	85.88	**86.77**	**75.10**	**75.17**	**75.31**	**71.56**	**71.52**	**71.66**	**78 ±39**	**78 ±39**	**78 ±39**
Wav2vec2-PT	*32.59*	*19.62*	*20.28*	*50.63*	*43.78*	*43.50*	75.28†	–	–	*53 ±32*	–	–

Notation: **bold-underline**: best results, underline: second-best results, *italic*: evaluation of SpeechBrain's classifier pretrained on IEMOCAP.
†: 5-fold cross-validation result from https://huggingface.co/speechbrain/emotion-recognition-wav2vec2-IEMOCAP

into a lower dimension space using PCA method, the WF1 results on RAVDESS and IEMOCAP increased by 1.03% and 12.08% respectively, while the WF1 on CREMA-D slightly decreased by 0.88%.

The PCA method increased the average UF1 classification results for ComParE, MuSER, eMuSERn by 5%, 3% and 2% respectively. On the clean speech dataset RAVDESS, this simple dimensionality reduction method also enhanced the UF1 of the ComParE, MMC, MuSERn and eMuSERn by 1 to 3%, especially 6% on the MuSER model. The MuSER-PCA model also surpassed the MuSER model on the CREMA-D and IEMOCAP. PCA's most significant accuracy boost is on the ComParE model with the IEMOCAP dataset, which is 14.4% UF1. However, the PCA method does not always guarantee its effectiveness, especially with the GeMAPS and eGeMAPS features.

B. Model size comparison

Table V describes the model sizes. For all acoustic audio features (music-related features and LLD except ComParE), the total training and testing time of the model without PCA on a regular CPU (AMD Ryzen 7 3700X 8-Core Processor) for RAVDESS, CREMA-D and IEMOCAP datasets took less than 12, 55 and 120 seconds respectively. It needs less than 10ms for inference with the model size below 6MB. To sum up, the proposed method with one of the proposed music-related feature sets, and the 2-layer neural network model has shown its effectiveness and efficiency for speech emotion recognition across different datasets.

C. Ablation study

To further investigate the contribution of each feature in the proposed feature sets, we extended the first experiment to individual features namely Mel spectrogram, MFCC and CQT spectrogram, Chromagram as well as the log-power spectrum variants of Mel spectrogram and CQT spectrogram. Table VI illustrates the performance of each feature set without PCA and with PCA 12.5% on three datasets.

Table V: Model size comparison.

Feature	**Acoustic features**	**TRILL**	**Wav2vec2-PT**
Model size	3 to 6 MB	92.56 MB	377.6 MB

It can be observed that without PCA, MFCC yielded the highest accuracy on RAVDESS and IEMOCAP with 79.26% and 65.79% UA respectively, while CQT spectrogram gave the best result on CREMA-D with 67.45%. The main difference of the CREMA-D dataset compared to RAVDESS and IEMOCAP is a large number of speakers with different ethnicities. This reveals the importance of combining pitch and timbre features for building a robust SER model.

On the performance of pitch-related features, the CQT spectrogram surpassed the Mel spectrogram, especially the CQT spectrogram average UA result is 6 to 7% higher than the Mel spectrogram result with or without PCA. Chromagram gave the lowest results among all individual features as the dimension is only 12. However, it achieved 50% UA without PCA, which is 18% higher on average when compared to ComParE with a much larger set of 6,373 LDD features. These results strongly support the effectiveness of pitch-related features in speech emotion recognition.

On average, log-CQT spectrogram achieved 68% UA with PCA, which is 1% higher than MFCC and log-Mel spectrogram. The transformation into a normalized logarithmic representation significantly improved the performance of CQT

Table VI: Results of each feature with and without PCA in percentage (%).

Dataset	**RAVDESS**			**CREMA-D**			**IEMOCAP**			**Average**		
Feature	**UA**	**UF1**	**WF1**	**UA**	**UF1**	**WF1**	**UA**	**UF1**	**WF1**	**UA**	**UF1**	**WF1**
No PCA												
MFCC	**79.26**	**77.87**	**79.24**	63.67	63.24	63.53	**65.79**	**66.02**	**65.89**	**70 ±08**	**69 ±08**	**70 ±08**
Mel	54.07	52.43	52.74	61.94	61.69	61.81	53.74	52.87	53.52	57 ±05	56 ±05	56 ±05
Log-Mel	60.74	48.85	56.02	63.37	63.63	63.82	57.39	57.27	57.32	60 ±03	57 ±07	59 ±04
CQT	66.67	62.17	65.25	**67.45**	**67.49**	**67.65**	57.09	56.70	57.00	64 ±06	62 ±05	63 ±06
Log-CQT	76.30	75.44	76.19	62.86	62.44	62.62	63.26	62.83	62.85	67 ±08	67 ±07	67 ±08
Chromagram	54.81	53.58	54.88	50.20	49.61	49.92	43.83	42.97	43.62	50 ±06	49 ±05	49 ±06
PCA 12.5%												
MFCC	**85.19**	**84.36**	**85.49**	60.20	59.86	60.05	56.78	56.56	56.79	67 ±16	66 ±15	67 ±16
Mel	54.81	44.20	50.73	55.61	54.44	54.67	48.18	47.40	47.88	53 ±04	49 ±05	51 ±03
Log-Mel	80.00	78.19	80.11	62.86	62.63	62.83	59.01	58.71	59.08	67 ±11	66 ±10	67 ±11
CQT	57.04	56.03	56.27	62.76	62.37	62.46	56.38	55.92	56.26	59 ±04	58 ±04	58 ±04
Log-CQT	80.00	78.79	79.71	**64.18**	**63.89**	**64.11**	**59.41**	**59.48**	**59.36**	**68 ±11**	**67 ±10**	**68 ±11**
Chromagram	46.67	42.19	42.19	54.59	54.99	55.21	44.53	38.18	40.37	49 ±05	45 ±09	46 ±08

Notation: **bold-underline**: best results, underline: second-best results.

and Mel spectrogram by 3% without PCA, and by 9-14% with PCA. This emphasizes the importance of perceptual loudness.

V. Conclusions

This research gave a detailed analysis of audio-based features and their performance on the RAVDESS, CREMA-D and IEMOCAP datasets in emotion classification. Based on the psychology evidence on the similarity in the human perception of emotional speech and music, we proposed new sets of music-related audio features for SER. Our method achieved comparable results with the deep learning model embedding TRILL despite being up to 15 times smaller in model size. The ablation study showed that without PCA, while MFCC is the best feature for RAVDESS and IEMOCAP, CQT can handle a large number of speakers with different ethnicities from the CREMA-D dataset. Thus, the combination of pitch and timbre features is crucial to distinguish emotional speech. The limitation of this method is its sensitivity towards background noise, which can be deduced based on the results of the CREMA-D and IEMOCAP datasets. In conclusion, the proposed method is effective in terms of accuracy and model size for basic emotion detection from audio speech.

Nevertheless, there is still room for improvement over the model robustness against noise. Regarding feature extraction, we can fine-tune the parameters of the DFT and CQT kernels as well as filterbanks together with the neural network model via gradient descent to learn higher discriminative features and reduce the sensitivity to noise [14]. Another suggestion is to replace PCA with other non-linear dimensionality reduction methods such as autoencoders. This can enable the model to learn the underlying manifold structure of data to improve the classification result.

References

[1] J. Shor, A. Jansen, R. Maor, and et al., "Towards learning a universal non-semantic representation of speech," *Interspeech*, 2020.

[2] Z. Zhao, K. Wang, and et al., "Self-attention transfer networks for speech emotion recognition," *Virtual Reality & Intelligent Hardware*, 2021.

[3] V. Dissanayake, H. Zhang, and et al., "Speech emotion recognition 'in the wild' using an autoencoder," *Interspeech*, 2020.

[4] S. Siriwardhana, A. Reis, and et al., "Jointly fine-tuning 'bert-like' self supervised models to improve multimodal speech emotion recognition," *Interspeech*, 2020.

[5] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, and et al., "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech*, 2013.

[6] F. Eyben, K. R. Scherer, and et al., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, 2015.

[7] M. C. Sezgin, B. Gunsel, and G. K. Kurt, "Perceptual audio features for emotion detection," *EURASIP Journal on Audio, Speech, and Music Processing*, 2012.

[8] A. Bhatara and et al., "Expression of emotion in music and vocal communication: Introduction to the research topic," *Frontiers in psychology*, 2014.

[9] C. F. Lima and et al., "Impaired socio-emotional processing in a developmental music disorder," *Scientific Reports*, 2016.

[10] A. J. Oxenham, "Revisiting place and temporal theories of pitch," *Acoustical science and technology*, vol. 34, no. 6, pp. 388–396, 2013.

[11] Z. Meng and W. Chen, "Automatic music transcription based on convolutional neural network, constant q transform and mfcc," in *Journal of Physics: Conference Series*. IOP Publishing, 2020.

[12] K. Khoria, A. T. Patil, and H. A. Patil, "Significance of constant-q transform for voice liveness detection," in *Proc. EUSIPCO*, 2021.

[13] H. Terasawa, M. Slaney, and J. Berger, "A timbre space for speech," in *Interspeech*, 2005.

[14] K. W. Cheuk, H. Anderson, K. Agres, and D. Herremans, "nnaudio: An on-the-fly gpu audio to spectrogram conversion toolbox using 1d convolutional neural networks," *IEEE Access*, 2020.

[15] P. Kumar, V. Kaushik, and B. Raman, "Towards the explainability of multimodal speech emotion recognition," *Interspeech*, 2021.

[16] F. Eyben and et al., "Opensmile: the munich versatile and fast open-source audio feature extractor," in *ACM International Conference on Multimedia*, 2010.

[17] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, and et al., "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

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

[19] 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, 2018.

[20] D. Issa and et al., "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, 2020.

[21] H. Cao and et al., "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, 2014.

[22] C. Busso and et al., "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, 2008.