Comparative Analysis of Deep Learning Architectures for Emotion Classification in Audio

Fredrik Nguyen
HKUST
HKUST
fnguyen@connect.ust.hk

Abstract

Speech emotional recognition (SER) involves analyzing audio signals (speech) to identify the speaker's emotional state. This is a task that has become more important for us as we continue to develop machines that we interact with on a daily basis. It has become especially crucial with the emergence of AI assistants that directly is designed to act human-like, that we communicate human-like with, and therefore it need to understand emotion well. In this study, we introduce and use different deep learning architectures to classify emotions. In particular, we use a CNN, RNN, a hybrid CNN-RNN and the state of art transformer wav2vec2 model. The results show that all of the models achieved better performance than random choice. The best model was the wav2vec model that achieved a f1-score of 0.84 and an accuracy of 0.82%.

1. Introduction

In the digital age, where smartphones have become ubiquitous, much of our communication has shifted from face-to-face interactions to phone calls and virtual exchanges. This transition often makes it challenging to discern the emotions of others without visual cues like facial expressions or body language. Research highlights that body language is the primary contributor to effective communication, followed by vocal tone, with words playing a lesser role. Consequently, understanding vocal tone becomes critical in decoding the emotional intent behind spoken messages.

With the emergence of AI as assistants in our daily lives the exploration of how machines can interpret human speech and emotions has become all more crucial. In particular, it can be integrated with vision-language-action (VLA) AI and robotics where audio input seem to be the next step in the development. Other applications, include criminal investigation, call center answering and mental health analysis, and much more where it is of importance to recognize the mental state conveyed through emotions.

In this study we investigate how speech tone influences message interpretation. Specifically, we aim to evaluate whether deep learning models can accurately predict a speaker's emotions based solely on audio input. By comparing various deep learning architectures, we seek to identify effective approaches for emotion classification, contributing to advancements in human-computer interaction and affective computing. In particular, we will study the effectiveness of transformer models such as wav2vec 2.0 and a hybrid LSTM-CNN model. We also compare performance of models where the data has been augmented. Finally, we do an ablation study of the LSTM-CNN by splitting it into the CNN and the LSTM stand-alone and observe their performance. We find that the wav2vec model outperformed the other models by good margin, followed by the LSTM-CNN hybrid model, LSTM stand-alone and lastly the CNN stand-alone model. We interestingly found that data augmentation resulted in poorer results.

In section 2 we will first introduce some related work. We will introduce, some common models used in the past up to current state of the art models. We will then in section 3 introduce the architecture of the models in a high level and the dataset used used in this study. We will also go through any pre-processing and augmentation of the data we have done. In section 4, we will introduce the metrics used for evaluating the classification ability of the models and comparing them. This section will also include any specific experimentation setup and details. Lastly, a high level comparison between models will be presented. Finally, in section 5, we will give a brief summary of the study and results, along with some future directions.

2. Related work

Speech emotion classification using machine learning has been a topic of extensive research for decades, but recent advancements in deep learning have significantly accelerated progress. Earlier approaches relied heavily on man- ual feature engineering, requiring extensive feature selection and extraction. Commonly used features included Mel-frequency cepstral coefficients (MFCCs), pitch, energy, and formant frequencies, which were combined with traditional classifiers such as Support Vector Machines (SVM) [4] and Hidden Markov models (HMM) [9]. These methods, while effective in the past, were limited by their dependence on features extraction and their inability to fully capture the complex temporal and spectral features of emotional speech.

The advent of deep learning has revolutionized speech emotion classification, with the majority of modern models leveraging neural network architectures. These models usually only rely on spectrograms to extract features. Some of the major break through came from using CNNs that excel at extracting local spectral features [1] and RNNs such as Long-Short-Time-Memory (LSTM) networks that has temporal modeling and can capture contextual dependencies in sequential speech data. In [11], the Hybrid model of CNN+LSTMs was proposed by Shiqing *et al*. This model combines the strengths of CNNs for spatial feature extraction with LSTMs for temporal sequence modeling, resulting in superior performance compared to standalone models.

In addition to hybrid models, purely convolutional architectures have also gained traction for temporal feature extraction. Temporal Convolutional Networks (TCNs), for instance, utilize 1D convolutions to effectively capture sequential dependencies in speech data TCNs offer advantages such as parallel processing and reduced computational complexity compared to RNNs, making them a good alternative for certain applications. [5].

Recent advancement in speech emotion classification has incorporated attention mechanism in the models along with CNNs or LSTMs such as the Bi-directional Long-Short Term Memory with Directional Self-Attention model (BLSTM-DSA) where the self-attention is used for calculating similarity between frames and therefore more easily find the autocorrelation of speech frames in speech. [7]

The introduction of transformer architectures has further transformed the field, owing to their ability to capture long-range dependencies and contextual information in sequential data. Transformer-based models have rapidly become the state-of-the-art in speech emotion recognition. For example, Vision Transformers (ViTs) have been adapted to process mel-spectrograms, capturing spatial dependencies and high-level features more effectively than traditional CNNs, which often exhibit image-specific inductive biases [2]. There are also models that are combination of both CNN and Transformers such as the Convolution Transformer (CvT) introduced in [10] or the Wav2Vec model that uses CNN to encode raw audio followed by a transformer [3]. These models are often pretrained on general data and requires finetuning.

Beyond architectural innovations, recent research has explored multimodal approaches to speech emotion classification, integrating audio with complementary modalities such as text or visual data. For instance, models that combine speech features with transcribed text or facial expressions have shown improved robustness and accuracy, particularly in noisy or ambiguous scenarios.

In our study we will implement a version of the CNN-LSTM hybrid network as mentioned above and a Wav2Vec model. Which are models that are same or inspired by models from earlier studies.

3. Methodology

In this section we will describe the data and any processing step and the models used for the experiment.

3.1. Data

3.1.1 Dataset

The data used in the study is the RAVDESS Emotional speech audio dataset [8] retrieved from kaggle [1]. This particular subset of the whole RAVDESS dataset consist of 1440 speech-only .wav audio files. It is broken down to 24 voice actors (12 male, 12 female) with 60 voice recording each of two statements in neutral North-American accent. The speech emotion of the recordings include calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is made in two levels of intensity normal or strong. The audio consist of either the statements ”Kids are talking by the door” or the phrase ”Dogs are sitting by the door”.

3.1.2 Data processing

The different architectures requires different data processing.

- Wav2Vec: We first resample all audio waveforms to 16000Hz. We then compress it to one channel audio by taking the mean of the channels. Then we pad the audio to equal length going by the longest one. Lastly, we use Wav2Vec2 internal processor that we will mention in section 3.2.2. to process the audio.

- Similarily to the Wav2Vec, for the LSTM model we first standardize the audiofiles such that they are equal length using padding and have the same sample rate to ensure consistent input size in the models. To reduce the dimension we performed feature extraction, by performing normalized mel-spectrography which captures the frequency content over time and further

[1] Kaggle dataset

process them to extract the MFCC (Mel-Frequency Cepstral Coefficients) which is a set of coefficients that capture the spectral characteristics of an audio signal.

The data will be split into 70:15:15 split for training, validation and testing data, by stratifying by the emotion label to ensure balanced data in all datasets.

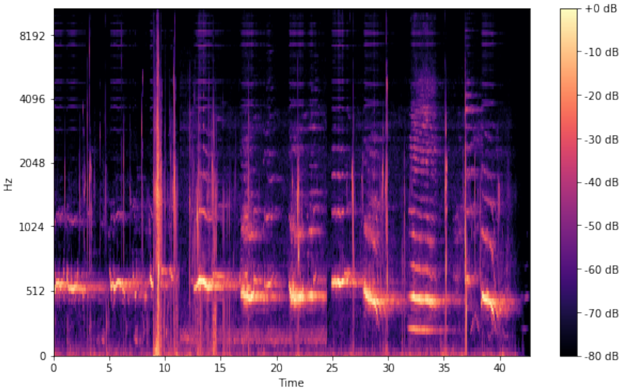

Figure 1. Mel spectrogram for reference

3.1.3 Data augmentation

For the LSTM-CNN model we also run it on a augmented data set. The augmentation consist of application of audio shift, pitch shifting, time stretching and noise which are conventional and typical methods for augmenting audio data. The values where chosen arbitrarily but with reference from other studies.

- Gaussian Noise: We use min amplitude=0.001 and max amplitude=0.015 with probability 0.5
- TimeStretch: We use min rate=0.8 and max rate=1.2, with probability 0.5
- PitchShift: We use min semitones=-4 and max semitones=4 with probability 0.5
- TimeShift: We use min shift=-0.5s and max shift=0.5s with probability p=0.5

3.2. Models

3.2.1 CNN-LSTM

The implementation of the CNN-LSTM is inspired by the following paper [6]. The CNN-LSTM network is sequential neural network that has a convolution network which takes the the mel spectrogram as the input followed by a LSTM network and finally fully connected layers for prediction.

The Convolution network consist:

- Convolution Layers with (3, 3) kernels and stride (1, 1) and padding to get the same size as input.
- Batch Normalization Layer
- ELU activation Layer
- 2D MaxPooling Layer with (2, 2) kernels and stride (2, 2)
- Dropout Layer with 0.2 probability

The output channels of each convolution layer is 64, 64, 128, 128 in that order.

Then the output is resized to 1D and fed into a LSTM layer with 256 before being fed into a fully connected network with 128 channels in each layers and output channels 8 for each emotion. See fig 2 for reference of the architecture.

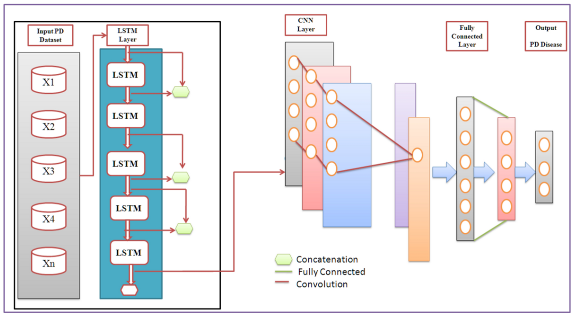

Figure 2. Illustration of the CNN-LSTM hybrid network

3.2.2 Wav2Vec 2.0

The Wav2Vec 2.0 model was introduced by META and is a self-supervised learning model for automatic speech recognition [3]. It leverages large amounts of unlabeled audio data to pre-train a model. It is trained by masking the speech input in the latent space and solves a contrastive task. It consists of several layers, see figure 3. First it has a feature encoder which transforms the raw audio into a sequence of low-level feature vectors. Then it has a contextualized representations network with transformers which processes the masked feature vectors to produce contextualized representations. Continuing, for the pre-training it has a quantization module which discretizes the feature encoder's output to create a set of discrete speech units, which are used for self-supervised learning. Finally, we will add a fully connected network in the finetuning for classification.

Wav2Vec2 is pre-trained using a self-supervised contrastive loss, combining two objectives:

- Contrastive Loss: The model learns to distinguish the true quantized representation of a masked audio frame from distractors (randomly sampled alternatives). This encourages the model to learn robust contextual representations.

- Diversity Loss: Ensures the quantization module uses the codebook entries uniformly, preventing collapse to a few discrete units.

See fig 3 for reference of the architecture.

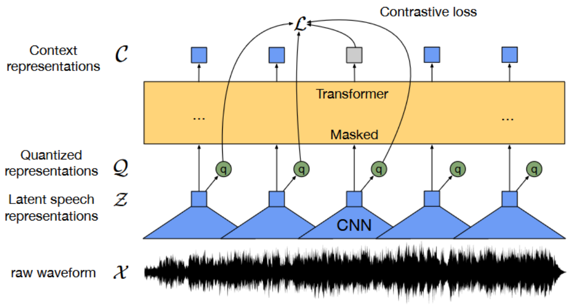

Figure 3. Illustration of the Wav2Vec 2.0 framework

4. Experiment

In below section we describe the experiment setup, evaluation metrics and finally present the result. We compare the results between models and analyze their resulting prediction distribution among the different emotion to come up with conclusions. Finally, we will do a simple ablation study of the LSTM-CNN hybrid network by splitting it in two networks and compare the result among them.

4.1. Evaluation Metrics

The models classification ability will be evaluated using accuracy and F1-score. We will evaluate the models using all recordings as well as on emotion specific, gender specific, statement specific and also intensity specific conditions such that we can extract any biases in the model.

4.2. Implementation Details

For the Wave2Vec 2.0 we will be using the model from Hugging Face transformer library with pretrained weights from the Librispeech dataset. [2]. It will be trained and tested using the Hugging face wrapper for PyTorch.

For the LSTM-CNN model, we will implement our own model according to its description above which we will build, train and test using Tensor Flow Keras Library framework.

The models will be trained using Adam optimizer with the default training rate 0.001, using categorical cross-entropy loss. We train both models for 10 epochs to ensure fairness when comparing models. This is with the long training time of the wav2vec model in mind. In actuality, we can see for instance, that the accuracy in fig. **??** and loss in fig. 5 of the LSTM-CNN model is still increasing and decreasing respectively (not plateauing) indicating that the performance of the model could still improve with more training. In each epoch we train on the training set and then save the best model depending on evaluation on the validation set. Finally we evaluate on the test set which is our final results.

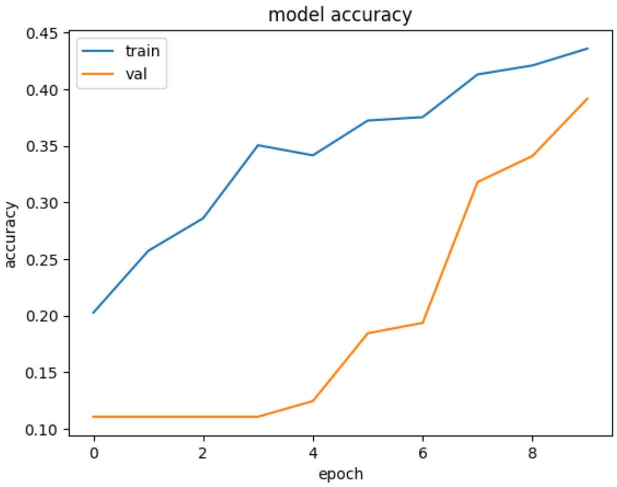

Figure 4. Validation and training accuracy

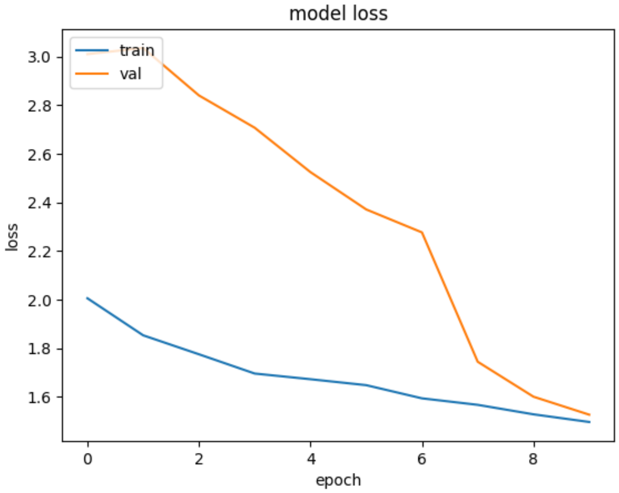

Figure 5. Validation and training loss

[2] Hugging Face Wav2Vec 2.0

4.3. Quantitative and Qualitative Analysis

The f1-score and accuracy evaluation on the test-set is presented in table 1. As we can see the Wav2Vec model outperform the other models significantly. As for the LSTM-CNN model it performs better than the baseline random choice which would have an accuracy of approximately 1/8. As we can the data augmented version of the before mentioned model is significantly worse. It could due to many reason, for instance the augmentation values where slightly arbitrary. However, our hypothesis is that it was due to only training for 10 epochs, which essentially made the model train on deformed/bad data for only a few generations making it not being able to generalize to the normal data that it usually will be able to do with data augmentation. In the confusion matrix presented in fig. 6, fig. 7, fig. 8 we see that both models perform differently depending on the emotion in the audio. For example, we see that the Wav2Vec excelled in predicting angry and calm, while the LSTM-CNN excelled in surprise. Interestingly, we see that while the LSTM-CNN with augmented data also got a lot of 'surprise' emotion correct it also got 'calm' correct a lot. In fact it seems to predict 'calm' more than normal which might suggest that the data augmentation removed features from the MFCC which made it less expressive like 'calm' MFCC.

Table 1. Model Performance Comparison

	Weighted F1 Score ↑	Accuracy (%) ↑
Wav2Vec2	0.8212	0.8194
LSTM-CNN	0.4463	0.4630
LSTM-CNN (Data augmented)	0.3462	0.3889
LSTM	0.4044	0.4352
CNN	0.3014	0.3519

4.4. Ablation study

By dividing the LSTM-CNN in either the LSTM part or the CNN part we can see how each part perform standalone. In table 1 we see that the LSTM-CNN perform better than both standalone parts as expected. The LSTM performed better than the CNN which suggest that the sequential features of the audio is more important than the tone/frequency of the audio which the CNN captures in the MFCC. We can also notice the difference in prediction in fig 9 and 10.

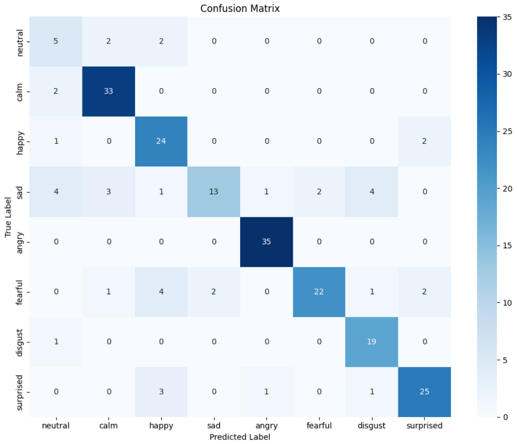

Figure 6. Confusion matrix for Wav2Vec

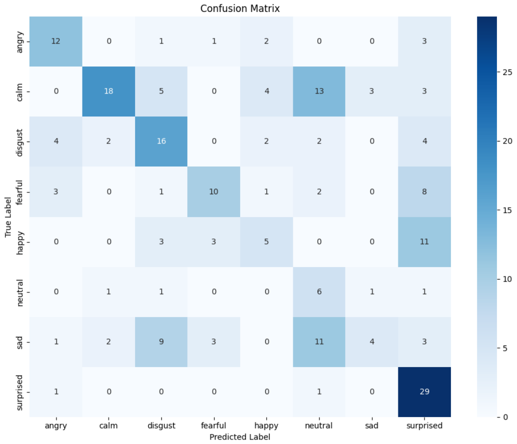

Figure 7. Confusion matrix for LSTM-CNN

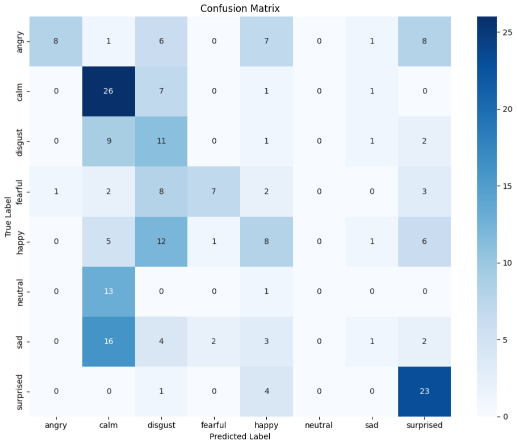

Figure 8. Confusion matrix for Data augmented LSTM-CNN

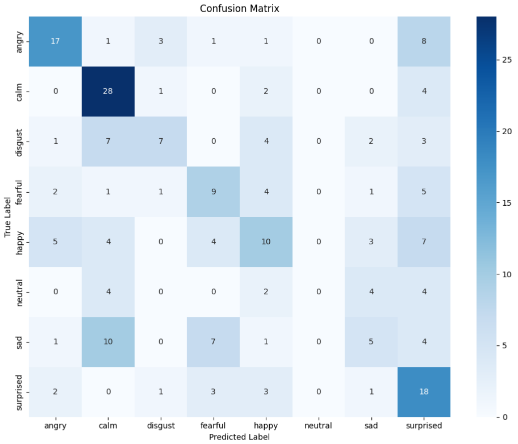

Figure 9. Confusion matrix for LSTM

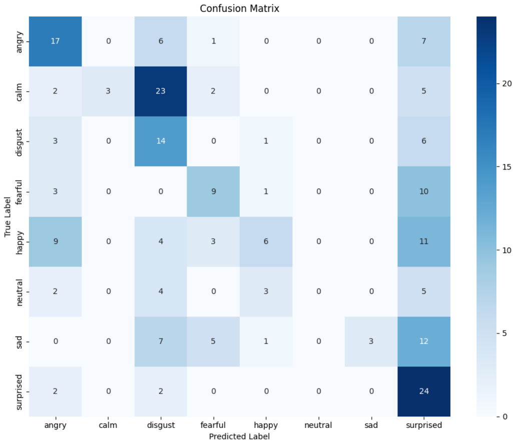

Figure 10. Confusion matrix for CNN

4.5. Other features

As mentioned above, other features included in the dataset besides emotion was the intensity of the emotion in the audio, gender and statement given in the audio. We studied these features further using the LSTM-CNN model. In fig. 11, fig. 12, fig. 13 we can see the accuracy given any of the features. Interestingly, the accuracy for strong emotion is worse than for normal intensity. However, the difference is so small it might be statistically insignificant. The same goes for the statement said. However, between men and women the difference is very large. This could be due to the frequency of female voice being higher which results in more contrast in the MFCC. We had hoped by augmenting the data, we could eliminate this bias. However, it resulted in approximately the same values.

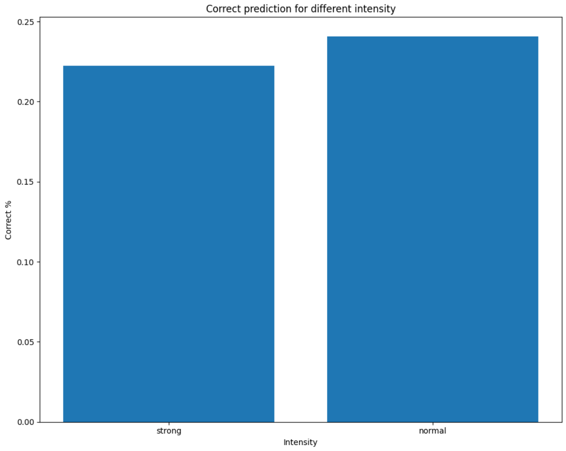

Figure 11. Accuracy for different emotion intensity

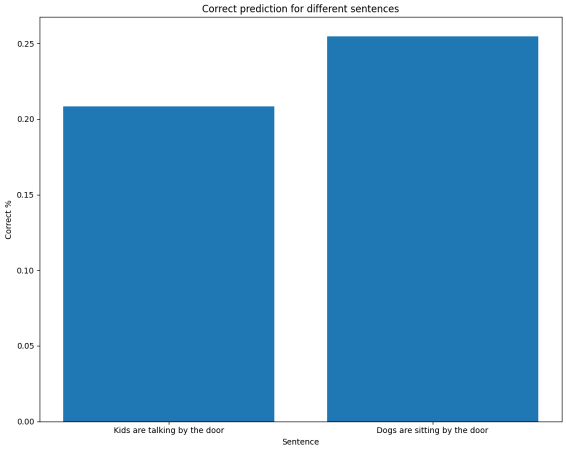

Figure 12. Accuracy for different statements

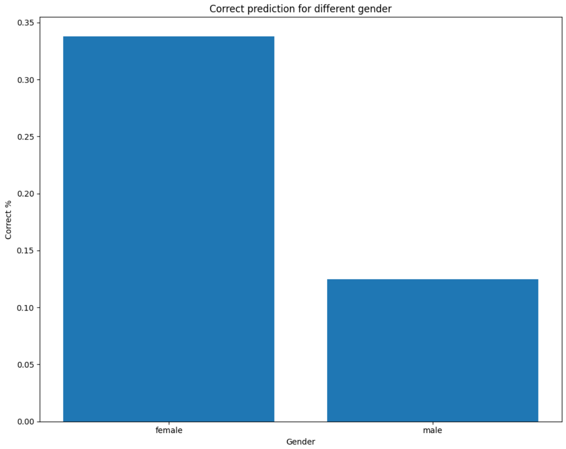

Figure 13. Accuracy for different gender of the actor

5. Conclusion

To summarize, both the LSTM-CNN and wav2vec model could perform the classification task well, even the RNN and CNN standalone derived from the LSTM-CNN model. Out the models the Wav2Vec had best results. Some interesting results suggest that men emotion is more difficult for the models to predict, and that the data augmentation used in this study gave worse result even though it normally enhances the performance which suggest the augmentation was employed wrongly. Some further extension of this study, could involve more extensive ablation study of the different layers, or other types of augmentation. Some interesting challenges remain for the speech emotion recognition in general is the generalization of models across diverse datasets, languages, and cultural contexts which is a significant hurdle, as emotional expression varies widely. In this study, we only used one dataset consisting of only American voice actors, which could be extended.

Furthermore, real-time processing requirements and computational efficiency are critical for deploying these models in practical applications, such as human-computer interaction or mental health monitoring. Future research is likely to focus on lightweight models, domain adaptation techniques, and the integration of cross-modal information to address some of these limitations. -

References

[1] Alif Bin Abdul Qayyum, Asiful Arefeen, and Celia Shahnaz. Convolutional neural network (cnn) based speech-emotion recognition. In *2019 IEEE International Conference on Signal Processing, Information, Communication Systems (SPICSCON)*, pages 122–125, 2019. 2

[2] Samson Akinpelu, Serestina Viriri, and Adekanmi Adegun. An enhanced speech emotion recognition using vision transformer. *Scientific Reports*, 14, 06 2024. 2

[3] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. 2, 3

[4] S. Casale, A. Russo, G. Scebba, and S. Serrano. Speech emotion classification using machine learning algorithms. In *2008 IEEE International Conference on Semantic Computing*, pages 158–165, 2008. 2

[5] Muhammad Ishaq, Mustaqeem Khan, and Soonil Kwon. Tc-net: A modest lightweight emotion recognition system using temporal convolution network. *Computer Systems Science and Engineering*, 46:3355–3369, 01 2023. 2

[6] Baij Kaushik. A hybrid technique using cnn+lstm for speech emotion recognition. *International Journal of Engineering and Advanced Technology*, 9:1126–1130, 08 2020. 3

[7] Dongdong Li, Jinlin Liu, Zhuo Yang, Linyu Sun, and Zhe Wang. Speech emotion recognition using recurrent neural networks with directional self-attention. *Expert Systems with Applications*, 173:114683, 2021. 2

[8] Steven R. Livingstone and Frank 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*, 13(5):1–35, 05 2018. 2

[9] Tin Lay Nwe, Foo Say Wei, and L.C. De Silva. Speech based emotion classification. In *Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology. TENCON 2001 (Cat. No.01CH37239)*, volume 1, pages 297–301 vol.1, 2001. 2

[10] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers, 2021. 2

[11] Shiqing Zhang, Xiaoming Zhao, and Qi Tian. Spontaneous speech emotion recognition using multiscale deep convolutional lstm. *IEEE Transactions on Affective Computing*, 13(2):680–688, 2022. 2