Vocal Sentiment Analysis

Authors: DR. Urooj Ainuddin (Assistant Professor NED University), Sukaina Asad, Shehroz Waseem, Abaad Murtaza, Aiman Nisar

***Abstract*— In this study, we combine many methods for improving emotion-level categorization, including context-level analysis, deep learning techniques like CNN, and conventional vocal feature extraction (MFCC). We investigate untested models to determine how they work and how accurate they are. To enhance performance, we use data augmentation and hyperparameter sweeps. Finally, we examine whether a real-time strategy is workable and easily incorporated into current systems.**

Keywords—speech emotion recognition, feature extraction, convolution neural network, MFCC, data augmentation, sentiment analysis, speech processing, artificial intelligence.

I. INTRODUCTION

Speech signals are among the most essential component of human communication, and they have the advantage of being easily measured in real-time. They include the speaker's emotions as well as inferred communicative and linguistic content. It's crucial to integrate appropriate audio characteristics in speech emotion identification since feature selection impacts categorization performance. Sentiment Analysis has been performed over the years to assess the behavior of individuals. The fact that the human voice can portray a vast spectrum of emotions, from joy to sorrow, anguish to happiness, spontaneity to rigidity, delicacy to harshness, health to disease, laughter to crying, justifies speech-based emotion recognition algorithms.[4]

Despite the significant advancements in artificial intelligence[1], there is still time for being able to connect with robots intuitively, in part because they are unable to comprehend our emotional states.

Speech emotion identification, which seeks to identify emotion states from speech signals, has recently attracted more and more interest. Speech emotion recognition is an extremely difficult task, and how to extract useful emotional elements is still up for debate.

According to scientific evidence, all human's emotions cause psychological and physiological changes that affect the voice. Emotional speech processing technologies use computer analysis of speech features to determine the user's emotional state. Pattern recognition algorithms can be used to evaluate voice features and prosodic aspects such as pitch variations and speech rate. Features extracted from human vocal data hold immense significance for sentiment analysis which gives insights into their emotional health.

II. SENTIMENT ANALYSIS

Sentiment Analysis is the translation and characterization of emotions (positive, negative, or neutral) within a set of data. It is possible to do so using content, sound, and video assessment methods. Sentiments or emotions are an important part of people's lives because they influence how they perceive or understand things. For the past two decades, a range of methodologies have been discovered and even implemented to facilitate sentiment analysis, ranging from manual methods such as questionnaires created by psychologists to computer-based methods.

There are multiple classes of sentiments which are expressed based on reactions received from human brain. Depending on the mood, the mind reacts, which is expressed in the action manner, and it is a physiologic state related to nerves. If the emotions were intense, the person could react in a variety of ways. The sentiment analyzer attempts to analyze emotion based on text, voice, and facial expressions in this environment.

Sentiment analysis is widely used to monitor an individual's behavior on social media and other platforms because it helps us to get a sense of how the general population feels about various issues.

III. VOCAL SENTIMENT ANALYSIS

A Speech signal can deliver a wide range of information, including paralinguistic and linguistic information. Sentiment is a prime example of content retrieved from the non-lexical elements of communication and it is primarily expressed through speech. Creating such machines that could understand these non-lexical features, such as emotion, enhances human-machine communication by making it more natural and understandable.

Researchers have been investigating emotion recognition for many years. The initial studies on emotion recognition relied on biological information like heart rates and skin textures as well as face cues to identify emotions. Emotion recognition from speech signals has been a center of attention in the recent times. It was because of the traditional concept which states that acoustic features and emotions have their strings attached. It can also be viewed in a way that acoustic and prosodic speech signal correlates such as speaking pace (rate), intonation, energy, formant frequencies, fundamental frequency (pitch), intensity (loudness), duration (length), and spectral characteristic (timbre) are used to encode emotion.

Several machine learning techniques have been investigated for classifying emotions based on acoustic correlates in spoken utterances. Hidden Markov models (HMM), Gaussian mixture models, nearest neighborhood classifiers, linear discriminant classifiers, artificial neural networks, and support vector machines are some examples of frequently used methods to categorize emotions on the basis of vocal features.

IV. RELATED WORKS

Sentiment Analysis has been performed over the years to assess the behavior of individuals. The fact that the human voice can transmit a wide range of emotions, from joy to sorrow, anguish to happiness, spontaneity to rigidity, delicacy to harshness, health to disease, laughter to crying, justifies speech-based emotion recognition algorithms.

Dasgupta[3] explained that there exist multiple algorithms that incorporate features like pitch, timbre, the

gap between the words, etc. in order to state the comparisons. Prosody features like Mel Frequency Cepstral Coefficients (MFCC) have also been used for sentiment analysis using vocal data.

In order to perform Sentiment Analysis extracting features from the data has been a matter of interest for a long time. Broadly there are 4 categories where the features fit.[2] First are the acoustic features which include characteristics like pitch and quality. Next are the lexical features which primarily deal with the textual or language features. Third, are context information features and last are the hybrid features. Sentiment Analysis has been performed on almost all these features.

The field of textual sentiment analysis has been another important viewpoint, with the goal of extracting evaluative meaning[2]. Using a person's tweets or text from a certain document to determine if the individual is feeling good, negative, neutral, or more. This sort of analysis is termed text analysis. It plays an important role during this sort of prediction of emotions supported sentiment analysis. In Natural Language Processing (NLP), there are numerous ways to do sentiment analysis covering both supervised and unsupervised methods, as well as the field's future goals and constraints. The goal of supervised sentiment analysis is to create predictive models for sentiment using annotated datasets and automated learning.

Textual Sentiment Analysis has a few limitations, the use of emotion keywords is a simple technique for discovering related emotions; nevertheless, the meanings of keywords may be numerous yet ambiguous, since most words may alter their meanings depending on context and usage.

With the advancement in the field of vision-based emotion recognition in recent times, computer vision sentiment analysis is a relatively new domain of research. Facial Emotion Recognition (FER) is a system that reads your facial expressions from your pictures and videos and provide you the information about the emotional state of that person. Recent advances in face detection, face tracking, and face recognition systems have sparked the interest in this field. Model-based methods, holistic methods, local methods, and motion extraction methods are some of the methodologies identified for facial expression analysis. They also took into account several obstacles in expression identification, such as stance, lighting, and occlusions. It corely depends on the algorithm being used that whether facial expressions can be classified into basic emotions (e.g. anger, disgust, fear, joy, sadness, and surprise). Many algorithms are proposed for FER one of that is the HAAR classifier which calculates the difference in average intensity in various sections of the image is encoded by features, which are black and white linked rectangles whose value equals the difference of the sum of pixel values in the monochromatic regions after that feature extraction is done from the collected data.

For vocal sentiment analysis or more commonly known as Speech Emotion Recognition, deep learning neural network techniques have been vastly used. A feed-forward neural network with multiple hidden layers separating its inputs and outputs is referred to as a deep neural network (DNN). It is capable of accurately identifying data and learning high-level representation from the bare features. When given enough training data and the right training methods, DNNs excel in many machine learning tasks, such as voice emotion identification. Compared to feature analysis in voice recognition, emotion recognition is significantly less explored. Most earlier studies used empirical feature selection to classify emotions. DNN is utilized to create segment-level emotion state probability distributions from the conventional acoustic data present in a speech segment. From these distributions, utterance-level features are created and used to identify the utterance-level emotion state.[5]

V. Methodology

This model is built on a dataset that constitutes two emotions. One negative emotion primarily the angry one and the other audios are termed as neutral. The goal of the model is to distinguish between negative and positive emotions of a human based on the vocal data provided to it. For a good model with high accuracy, data is cleaned, pre-processed and vocal features are extracted. We have extracted Mel Frequency Cepstral Coefficients (MFCC) features from our vocal data. The datasets are stripped of all superfluous data and redundant features. All of those attributes are preserved in the dataset and are included in the sensitivity list. Furthermore, we have used data augmentation to make the shape of our data compatible with the model.

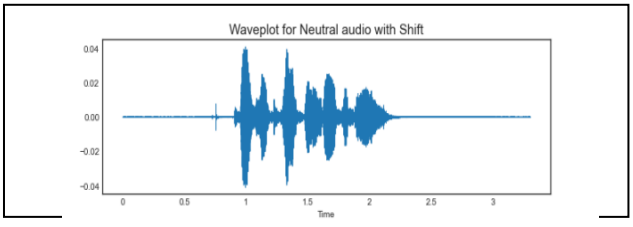
A convolutional neural network is used to extract sentiments from the audio data. A convolutional neural network is used concretely to process data

A. Data Augmentation

Because their full potential is increased when large data sets are utilized to train them, deep learning models are driven by data [9]. Indeed, it has been demonstrated that expanding the size of the training set lowers overfitting and enhances the generalizability of deep learning models. On the other hand, increasing the bulk of data is a costly and time-consuming procedure. Data augmentation, a regularization strategy, is used to fictitiously produce fresh training data and expand training sets in order to solve the problem with data gathering [7]. Data augmentation has gained a lot of attention in numerous machine learning tasks, especially classification. The number of training data sets can be successfully increased without impacting instance labels, for example, by rotating images for an object identification task or by encoding voice signals in background noise for a speech emotion detection task. There exist multiple techniques for data augmentation. The proposed technique utilizes noise injection, time shifts, pitch changes, and speed changes to create synthetic data. The aim was to improve the generalization ability of the model.

Adding perturbations must preserve the same label as the original training sample for this to operate [7].

In the proposed technique data augmentation is performed in several ways on a neutral-based emotion audio from our dataset. A sample wave plot is shown in the figure.



Shifting time is a relatively simple concept [8]. The audio is either shifted to left or right for a random second. if the audio is fast forwarded by x seconds, then the first x seconds will be marked as 0. (i.e., silence). Similarly, is the audio is moved to the right by x seconds then the last seconds will be marked as 0 (i.e., silence).

B. Feature Extraction

Mel Frequency Cepstral Coefficients (MFCCs) are retrieved from the vocal data in order to train the model. Mel Frequency Cepstral Coefficient is a popular and effective signal processing technique. When it comes to 1D signals, Mel frequency Cepstral coefficients (MFCC) are one of the most common and effective methods for extracting features from voice signals. By transforming input voice audio into a 1D signal, feature extraction was performed via the Mel Frequency Cepstral Coefficient (MFCC). It produces a cepstral coefficient, which is further utilized as a feature vector in the sentiment analysis procedure. [5]

This MFCC method lies on the frequency and bandwidth of the audible sound range of humans. It is incapable of xdetecting frequencies greater than 1 kHz. The results are obtained in multiple steps when a speech signal is given as input.

Step 01: Pre-Emphasis Filtering

This first step deals with the balancing of the vocal spectrum in order to have a steep roll-off in the high-frequency region. To achieve this balancing low-frequency signals are provided with some additional signal energy.

$$S'(n) = S(n) - \alpha * S(n-1)$$

Here α is referred to as the pre-emphasis coefficient whose value lies in the range of 0.9-1.

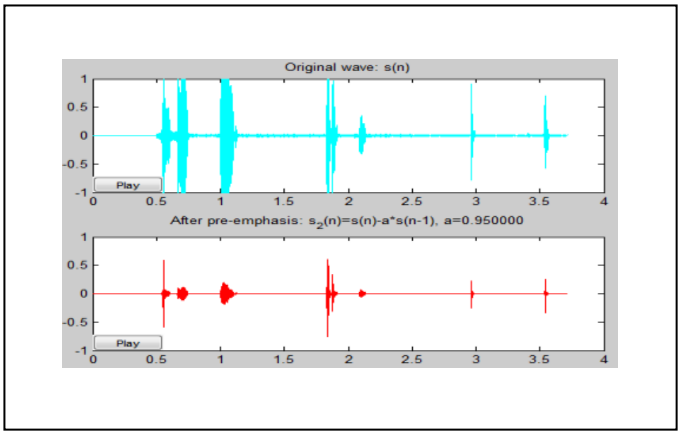


Figure 2. Pre Emphasis

Step 02: Frame Blocking

When the signal has passed the pre-emphasis step it is now ready to split into multiple frames. Speech must be studied over a sufficiently short period due to stable acoustic characteristics. As a result, speech analysis has to be performed always on segments of short length when the speech signal is presumed to be steady.

Step 03: Windowing

The windowing technique has to be applied to every frame to reduce signal discontinuity generated by the frame blocking process at both ends of each frame. The Hamming window is utilized, and the frame from the preceding operation is multiplied by it. The formula can be used to compute hamming Windowing:

$$w(n) = 0.54 - 0.46\cos(2\pi n / N - 1)$$

Here n lies in the range of 0 till N-1.

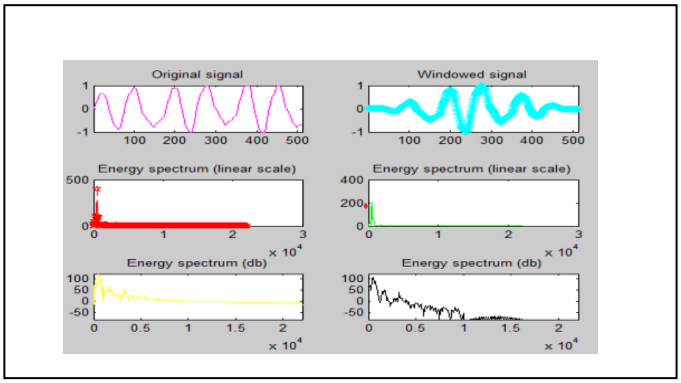


Figure 3. Sharp edges after windowing

Step 04: Fast Fourier Transform

It is important to translate every frame into frequency domain from their time domain version. For this Fast Fourier Transform is used. If the frames exist in frequency domain, they are more understandable and easier to understand.

$$Y(w) = FFT[h(t) * x(t) = H(w) * X(w)$$

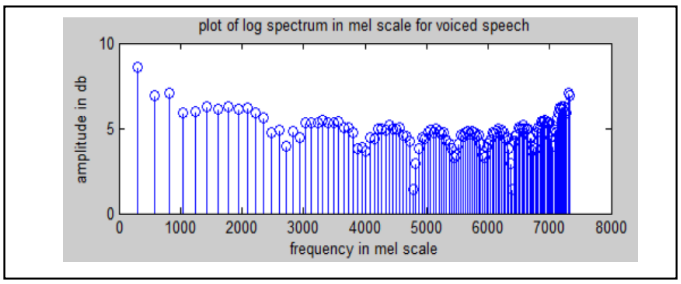


Figure 4. Voiced speech spectrum

Step 05: Mel Frequency Wrapping

Tones of various frequencies, calculated as f and in Hz, make up the speech signal. In the meanwhile, "Mel" units are used to express subjective pitch. The "Mel" frequency scale is linear for frequencies below 1 kHz and logarithmic for frequencies above 1 kHz.

$$fmel = 2595 \log10(1 + f/700)$$

Step 06: Discrete Cosine Transform (DCT)

Now log Mel spectrum is taken back into the time domain using DCT. The Mel Frequency Cepstral Coefficient is the result. We refer to the set of coefficients as the acoustic vector.

One of the best tools for removing characteristics from audio waveforms (and other digital signals in general) are Mel Frequency Cepstral Coefficients (MFCCs).

For extracting the MFCC feature from audio we have to define the sampling rate which is defined 22050 Hz here along with the number of MFCC's to be extracted from the audio are 58.

```
array([-6.97928345e+02,  7.73176041e+01, -1.61472988e+00,  2.17203465e+01,
        4.62321663e+00,  6.10153151e+00, -8.18462944e+00, -1.01108003e+00,
       -1.52747717e+01, -2.70904946e+00, -1.99016297e+00, -1.34955096e+00,
       -1.77466166e+00, -2.07293749e+00, -2.30235052e+00,  3.16991425e+00,
       -7.66876602e+00, -4.57050614e-02, -2.12107229e+00, -1.50315058e+00,
       -5.80852890e+00, -1.46739030e+00, -3.20313287e+00, -5.28320885e+00,
       -1.79240656e+00, -1.83985150e+00, -5.19961596e+00,  9.93070304e-01,
       -3.01291561e+00, -5.79474382e-02, -1.84313500e+00, -2.09675956e+00,
       -1.57865787e+00, -3.60660672e+00, -1.33066654e+00, -1.29110837e+00,
       -1.32152855e+00, -2.83525801e+00, -3.81865501e+00, -4.00056458e+00,
       -3.71870494e+00, -2.29128385e+00, -1.25041807e+00, -1.56987751e+00,
       -2.01045489e+00, -2.06818485e+00, -1.79611695e+00, -1.29479420e+00,
       -1.56357586e+00, -8.56688678e-01, -1.34954882e+00, -1.60825896e+00,
       -2.38189578e+00, -1.45623636e+00, -1.12093973e+00, -1.52607572e+00,
       -5.88103294e-01, -3.06453586e-01])
```

Figure 5. MFCC's of an audio file

C. Model

We have used Convolution Neural Network (CNN) for training and testing our model because these neural networks are a sort of deep learning model that has seen a lot of success in Speech Data Processing. The architecture of our CNN contains four convolution layers (Conv1D) each followed by a pooling layer out of which the starting three are Average Pooling layers and the last one is the Max pooling Layer. The foregoing processes produce a standardized 1D-vector (array), which is then sent to the convolution layer and subsequently to the pooling layers. Fully linked layer is the final building block of Convolution Neural Network, which is effectively a standard MLP. It results a 2D-vector output (array).

Every layer had ReLU (Rectified Linear Activation) used. The ReLU activation function takes in input and yields an output of y = x when given positive values of x and y = 0 in case of negative values of x. Because of its simplicity, it is widely utilized. The ReLU function is simple to compute, which is important for hidden layers because it speeds up the whole process. The last layer's function is 'SoftMax,' which returns y = 1 if x > 0.5 and 0 otherwise. This will help in the output of the one-hot vector. The next part is the optimizer used. The optimizer we used is Adam. Because the Adam optimizer improved the CNN abilities in classification and segmentation with the highest accuracy.

```
def build_model(in_shape):

    model=Sequential()
    model.add(Conv1D(256, kernel_size=6, strides=1, padding='same', activation='relu', input_shape=(in_shape, 1)))
    model.add(AveragePooling1D(pool_size=4, strides = 2, padding = 'same'))

    model.add(Conv1D(128, kernel_size=6, strides=1, padding='same', activation='relu'))
    model.add(AveragePooling1D(pool_size=4, strides = 2, padding = 'same'))

    model.add(Conv1D(128, kernel_size=6, strides=1, padding='same', activation='relu'))
    model.add(AveragePooling1D(pool_size=4, strides = 2, padding = 'same'))
    model.add(Dropout(0.2))

    model.add(Conv1D(64, kernel_size=6, strides=1, padding='same', activation='relu'))
    model.add(MaxPooling1D(pool_size=4, strides = 2, padding = 'same'))

    model.add(Flatten())
    model.add(Dense(units=32, activation='relu'))
    model.add(Dropout(0.3))

    model.add(Dense(units=8, activation='softmax'))
    model.compile(optimizer = 'adam' , loss = 'categorical_crossentropy' , metrics = ['accuracy'])
    return model
```

Figure 6. Fully Linked Layer

D. Training and Testing

The complete dataset was further split into two subsets for training and testing of our model, while keeping the test size equals to 0.20 which means the model is trained for 80% of the data and the rest 20% is used in the testing phase of the model.

VI. Conclusion

The presented approach identified the best CNN Model after rigorously implementing other models in order to perform emotion classification. This model was able to obtain a training accuracy of 97 percent. If additional data is provided to the model, it would perform better. The graphs in the given figures show how the model anticipated the actual numbers. The detection of sensitive cues is based on the extraction of deep frequency characteristics from voice spectrograms, which are more robust and discriminative in speech emotion recognition. This is done using a simple and small convolutional neural network (CNN) architecture comprising of multiple layers using modified kernels and a pooling strategy.

A. . Results and Observations

During training, a Loss curves are frequently used curves to debug a neural network. It gives a broad perspective of both the network's learning curve and the training process.

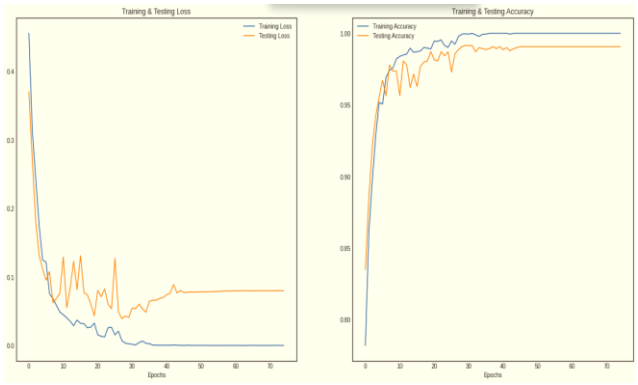


Figure 7. Training and Testing curves

Confusion Matrix was used to visualize the prediction of our True Positive results. As can be observed from the figure below, the network mostly correctly categorized emotions most of the time. For example, 678 audio samples which were actually angry were predicted angry by the model.

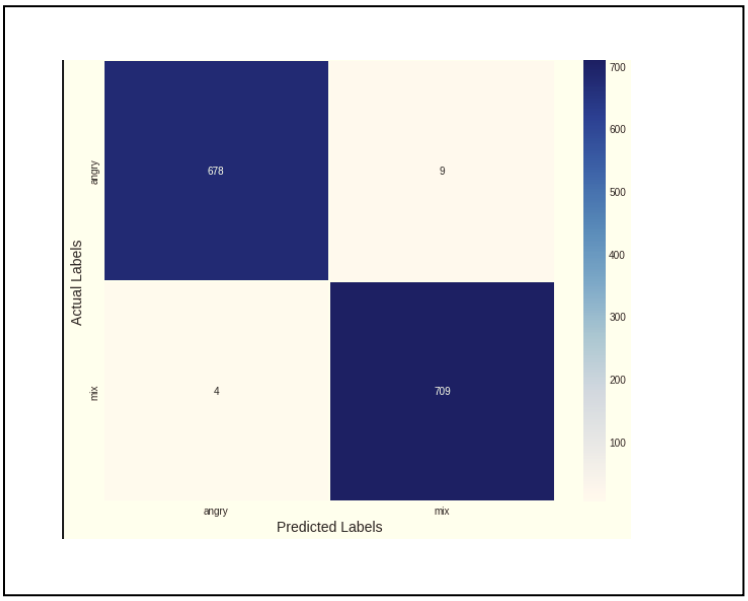


Figure 8. Confusion Matrix

B. Future Recommendations

This model has delivered accurate results on both training and testing datasets up to this point, but it still fails to make proper predictions when pre-recorded audio from a human is fed to it. Now, to increase this accuracy on real-time audio or pre-recorded audio from a user, we may either alter our model if and only if the outcomes stats aren't up to par, or we can modify or acquire new voice-based data for our model.

REFERENCES

[1] Huang, Andrew & Bao, Puwei. (2019). Human Vocal Sentiment Analysis. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-Level Multimodal Sentiment Analysis. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 973–982, Sofia, Bulgaria. Association for Computational Linguistics.

[3] Banerjee Dasgupta, Poorna. (2017). Detection and Analysis of Human Emotions through Voice and Speech Pattern Processing. International Journal of Computer Trends and Technology. 52. 10.14445/22312803/IJCTT-V52P101.

[4] Kwon, Oh-Wook & Chan, Kwokleung & Hao, Jiucang & Lee, Te-Won. (2003). Emotion recognition by speech signals. Proc Eurospeech. 10.21437/Eurospeech.2003-80.

[5] Han, Kun & Yu, Dong & Tashev, Ivan. (2014). Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.

[6] Zhang, Shiqing & Zhang, Shilliang & Huang, Tiejun & Gao, Wen. (2017). Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. IEEE Transactions on Multimedia. PP. 1-1. 10.1109/TMM.2017.2766843.

[7] X. Cui, V. Goel and B. Kingsbury, "Data Augmentation for Deep Neural Network Acoustic Modeling," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 9, pp. 1469-1477, Sept. 2015, doi: 10.1109/TASLP.2015.2438544.

[8] Yan Zhou, Jing Sun, "Speech Recognition Using Double Data Augmentation Strategy", 2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), vol.10, pp.715-719, 2022.

[9] Dino Oglic, Zoran Cvetkovic, Peter Sollich, Steve Renals, Bin Yu, "Towards Robust Waveform-Based Acoustic Models", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.30, pp.1977-1992, 2022.

[10] Jing, S.; Mao, X.; Chen, L. Prominence features Effective emotional features for speech emotion recognition. Digit. Signal Process. 2018, 72, 216–231.