Speech Emotion Recognition System using CNN

Anushka Sharma, Arjun Tanwar, Aman Deol, Rajesh Kumar Singh

Department of computer science & engineering MIET Meerut

anushka.sharma.cs.2018@miet.ac.in, arjun.tanwar.cs.2018@miet.ac.in,

aman.deol.cs.2018@miet.ac.in, rajesh.singh@miet.ac.in

Abstract. We as a human being sound voice to express our emotion to other like crying, anger, laugh or shouting these some basic expression we speech. So,this paper of uspropose to automatically recognize texture of voice or expression of voice.

Here in this research, we proposed network architectures based on CNN and MFCC text method to get most efficient reorganization of emotion. After comparing with existing state of the art technique efficiency find 5.6 to 7 percent

Keywords:CNN,MFCC,expression,recognize

Introduction

Many natural language solutions such as IVR, Voice activated system, and Chatbot require voice as an input and the system for example, voice-enacted frameworks, chatbots, and so forth require discourse as info. Standard technique is to initially change over this discourse contribution to message utilizing Automatic Speech Recognition system which helped in learning process.Kim proposed CNNs on top on pre discovered sentence-level layout which also improved quality in some extent. Zhang et al. proposed character level CNN for text categorization by using RNN.

By hi level development in the field of computer science and hardware speed and accuracy the estimation of emotion recognition of human voice become accurate and purposefull. ASR uses probabilistic acoustic and linguistic models to resolve variances in a speech from multiple users, resulting in speaker-independent speech transcriptions. Modern ASR systems provide high-accuracy outputs but lose a considerable amount of information that signals emotion from speech.

As a result of this gap, Speech-based

Emotion Recognition (SER) systems have been a hot topic in recent years. A natural way for people to express their emotions is via language. In addition, audio is easy to collect and evaluate in real-time situations, so most applications that rely on emotional recognition use speech. In order to forecast diverse emotional states, a SER system gathers information from speech such as pitch recurrence qualities, formant highlights, and energy-related data before making a classification assignment. Traditional methods used to solve taxonomy problems include the Bayesian Network Model, the Hidden Markov Model (HMM), the Support Vector Machines (SVM), the Gaussian Mixture Model (GMM), and the integration of multiple taxonomies.

Deep Learning approaches have made major advances in natural language interpretation during the previous decade (NLU). To better recognise emotions, Kim et al. and Zheng et al. found that deep belief networks (DBNs) outperformed baseline models that did not apply deep learning. This recommends that high-request nonlinear connections are more suited for emotion recognition. Despite the fact that the ELM's exactness improvement was humble, Han et al. proposed utilizing a DNN Extreme Learning Machine (ELM) that utilizes syllable and portion level likelihood disseminations and a solitary secret layer brain net to perceive expression level feelings. When it came to the SER, Fayek et al. used deep hierarchical structures, data augmentation, and regularisation using DNNs, whereas Zheng et al. used Spectograms and Deep CNNs.

It has been shown that training DNNs with small speech interval acoustic features and a probabilistic-natured CTC loss function improves recognition accuracy and allows for the assessment of protracted utterances that comprise both emotional and unemotional components. Lee et al. used a bidirectional LSTM model to train feature sequences and generated an emotional response.

IEMOCAP [18] dataset recognition accuracy of 62.8 percent, which is a critical improvement above DNN-ELM. Satt et al. improved their performance on the IEMOCAP dataset by combining deep CNNs with LSTMs.

Analysts have been exploring the utilization of multimodal properties for feeling acknowledgment as of late. Tzirakis et al. suggested a SER system that captures emotional information from multiple speech styles using auditory and visual modalities. For the ephemeral nature of language on the internet, Zadeh et al. proposed a Tensor Fusion Network that gains intra-methodology and between methodology elements from the start as far as possible. Utilizing Convolutional Deep Belief Networks (CDBN), which learn major multimodal properties of articulations, Ranganathan et al. were able to achieve high accuracy.

STATEMENT OF PROBLEM

Spectrograms and MFCCs, coupled with their associated speech characteristics, furnish a profound brain network with both semantic connections and low-level data fundamental to consistently distinguish between various moods in this investigation. Various investigations have been embraced on voice records and discourse highlights in an effort to improve the current state-of-the-art methodologies. Different input

combinations have been employed in various DNN designs, the specifics of which are explained in the following section.

OBJECTIVES

- A CNN model for emotion categorization based on speech characteristics is proposed (MFCC, Spectrogram)
- A CNN model for emotion categorization based on speech characteristics (MFCC, Spectrogram) and transcriptions was proposed.

CNN Model with MFCC input (Model3)

. As shown in Fig. 5, the Mel Frequency Cepstrum (MFC) represents the Short-Term Power Spectra of sound. On a log power spectrum, a non-linear linear cosine transform is used. We chose to use MFCC in our emotion identification testing since it is a widely used speech feature in many speech processing applications.

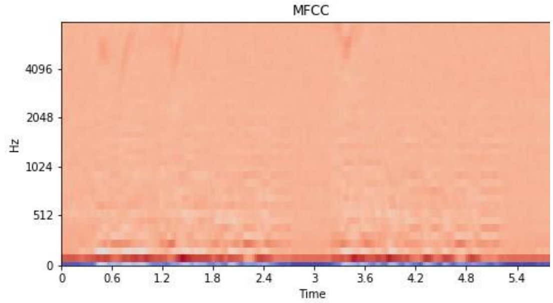


Fig.5.MFCCofthe audiofile being referredtoinFig.3

Similar to the Spectrogram generating procedure, the MFCC generation procedure uses the same hyperparameters and Python library (librosa). The only change is that instead of 128 Spectrogram coefficients per window, 40 MFCCs are constructed each window.

This model similarly has four equivalent convolutional layers, trailed by max-pooling layers and two further FC layers, as portrayed in the earlier area. Several different kernel sizes were tried since the input size varied from Models 2A and 2B. In the end, we settled on four different sizes of kernels, each with its own unique set of advantages and disadvantages, for this model.

CNN model with both text and speech features

Spectrograms, MFCCs, and speech transcriptions were combined in three different models to see if we could get the best of each.

We utilise distinct CNN channels for each input, as illustrated in Fig. 6, since the inputs are heterogeneous and hence have different dimensions. The following are the models:

MFCC and spectrogram (Model 4A) a

spectrogram and a text based model (Model 4B)

Text and MFCC (Model 4C)

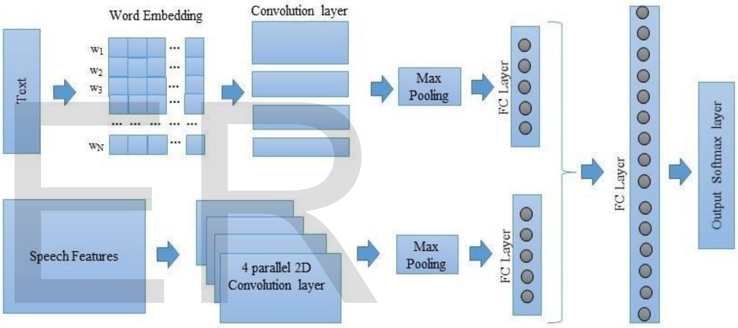


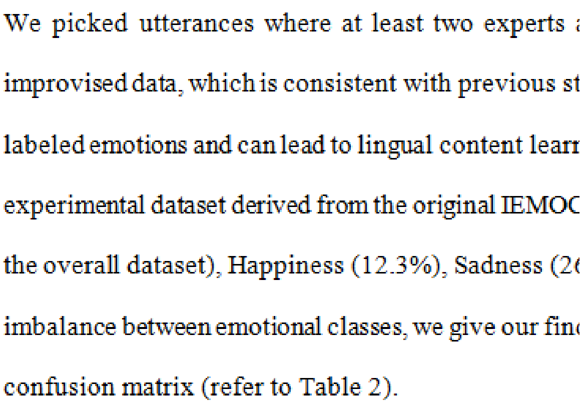
Figure 6 shows a representational CNN architecture.

The Spectrogram channel in Model 4A is made up of four parallel 2D-CNN layers with varying kernel sizes. The MFCC channel, like the Spectrogram channel, is made up of four equal 2D-CNN layers. The results of the two channels are steered to isolate FC layers. To feed the second FC layer, the normalised outputs from the first two FC levels are combined. Softmax layers are the next stage in processing the outputs of the previous FC layer. When comparing designs, keep in mind that Model 4A is a close cousin to Models 4B and C. There are text input channels for both models, which can receive word embeddings, but the voice feature alternates between

Spectrogram and MFCC.

DATASET

In this study, we used the Interactive Emotional Motion Capture (USC-IEMOCAP) database from the University of Southern California. The IEMOCAP corpus is divided into five sessions, each of which includes a discussion between two persons on both planned and impromptu themes, aswell as their accompanying tagged spoken text (both phoneme and word level). To eliminate any gender prejudice, Each session has both male and female actors and speakers. A brief speech (between three and fifteen seconds) is created from the audio-visual data and afterwards labelled by evaluators. Each utterance is examined by three to four evaluators. The assessors had the choice of categorizing each speech into one of ten distinct emotion classifications (Neutral, joy, sadness, anger, surprise, fear, disgust, frustration, excitement, etc.).



We picked utterances where at least two experts a
improvised data, which is consistent with previous st
labeled emotions and can lead to lingual content learn
experimental dataset derived from the original IEMOC
the overall dataset), Happiness (12.3%), Sadness (2(
imbalance between emotional classes, we give our fin(
confusion matrix (refer to Table 2).

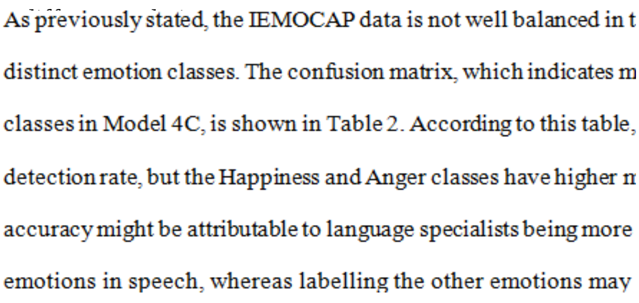
RESULT AND DISCUSSION

to continue previous research, we demonstrate the efficacy of the suggested approaches for emotion recognition using our benchmark findings on the IEMOCAP dataset. For separating training and test data, we employed a stratified K-fold. Table 1 shows some current results on emotion categorization, as well as cross verification based on five verifications. In general exactness depends on absolute counts paying little heed to class, while class precision is the mean of the precision accomplished in each class, in order to provide a more meaningful comparison between the two groups of results.

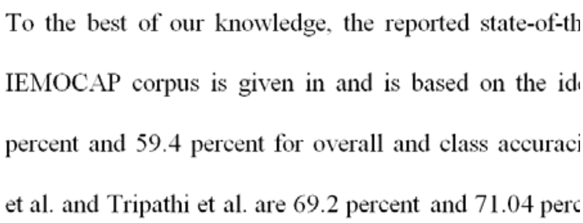
Table 1. Comparison of accuracies

Methods	**Input**	**Overall Accuracy**	**Class Accuracy**
Lee [4]	Spectrogram	62.8	63.9
Satt [19]	Spectrogram	68.8	59.4
Model 1	Text	64.4	47.9
Model 2A	Spectrogram	**71.2**	61.9
Model 2B	Spectrogram	**71.3**	61.6
Model 3	MFCC	**71.6**	59.9
Model 4A	Spectrogram & MFCC	**73.6**	62.9
Model 4B	Text & Spectrogram	**75.1**	**69.5**
Model 4C	Text & MFCC	**76.1**	**69.5**

Voice detection based on text and CNN technique have low level result hence results in low accuracy. whereas the combination of MFCC withspectrogram which is based on CNN iscomparatively high in accuracy. The MFCC based on text become the standard for measuring accuracy of emotion from

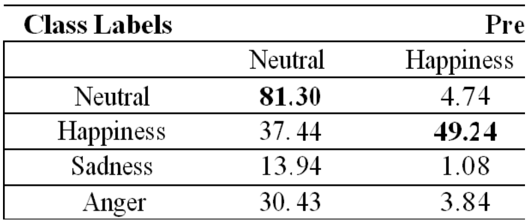


As previously stated, the IEMOCAP data is not well balanced in t
distinct emotion classes. The confusion matrix, which indicates m
classes in Model 4C, is shown in Table 2. According to this table,
detection rate, but the Happiness and Anger classes have higher n
accuracy might be attributable to language specialists being more
emotions in speech, whereas labelling the other emotions may



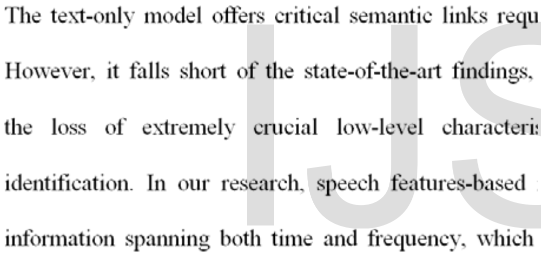
To the best of our knowledge, the reported state-of-th IEMOCAP corpus is given in and is based on the id percent and 59.4 percent for overall and class accuraci et al. and Tripathi et al. are 69.2 percent and 71.04 perc

Table 2. Confusion Matrix inPercentage onthe Model4C



Class Labels		Pre
	Neutral	Happiness
Neutral	**81.30**	4.74
Happiness	37. 44	**49.24**
Sadness	13.94	1.08
Anger	30. 43	3.84

Discussion



The text-only model offers critical semantic links requ However, it falls short of the state-of-the-art findings, the loss of extremely crucial low-level characteri identification. In our research, speech features-based information spanning both time and frequency, which

2D feature maps. It is possible to accurately identify emotions with 71.2 percent accuracy using Model 2A's Spectrogram model and 71.3 percent accuracy using Model 2B's more complex form of the model. With the MFCC-based paradigm, there is an additional benefit.

CONCLUSION

We suggested a number of CNN-based designsfor working with voice characteristics and transcriptions. When paired with text, the speech detection based on CNN and 2D perform more accurately. The technique of MFCC used with spectrogram is found to be 71 percent efficient which is almost 4 percent higher than previous results. When speech characteristics are employed in conjunction with speech transcriptions, the results are improved. The combination of Spectrogram with text technique has accuracy of 69.5 percent and total accuracy of 75.1 percent, but combining with MFCC with Text method has accuracy of 69.5 percent but which is total accuracy of 76.1 percent,

REFERENCES

1. Kim, Y.: Convolutional Neural Networks for Sentence Classification. In: Proceedings of the 2014 Conference on EMNLP, pp. 1746–1751 (2014).

2. Zhang, X., Zhao, J., LeCun, Y.: Character-level Convolutional Networks for Text Classifi- cation. In: Proceedings of the Annual Conference of the International Speech Communica- tion Association, INTERSPEECH, pp. 3057–3061 (2015).

3. Hinton, G.: Deep neural networks for acoustic modeling in speech recognition. In: IEEE Signal Processing Magazine 29, pp. 82–97 (2012).

4. Lee, J., Tashev, I.: High-level feature representation using recurrent neural network for speech emotion recognition. In: INTERSPEECH (2015).

5. Jin, Q., Li, C., Chen, S., Wu, H.: Speech emotion recognition with acoustic and lexical fea-

tures. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4749-4753 (2015).

6. Ververidis, D., Kotropoulos, C.: Fast and accurate sequential floating forward feature selectionwith the bayes classifier applied to speech emotion recognition. In: Signal Processing, vol. 88, no. 12, pp. 2956– 2970 (2008).

7. Mao, X., Chen, L., Fu, L.: Multi-level Speech Emotion Recognition Based on HMM and ANN. In: WRI World Congress on Computer Science and Information Engineering, pp. 225-229 (2009).

8. Ntalampiras, S., Fakotakis, N.: Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition. In: IEEE Transactions on Affective Computing 3.99, pp. 116-125 (2012).

9. Hao, H., Xu, M. X., Wu, W.: GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP, pp. 413-416 (2007).

10. Neiberg, D., Laskowski, K., Elenius, K.: Emotion Recognition in Spontaneous Speech Using GMMs. In: INTERSPEECH (2006).

11. Wu, C. H., Liang, W. B.: Emotion Recognition of Affective Speech Based on Multiple Clas- sifiers Using Acoustic-Prosodic Information and Semantic Labels. In: IEEE Transactions on Affective Computing 2.1, pp. 10-21 (2011).

12. Kim, Y., Lee, H., Provost, E. M.: Deep learning for robust feature generation in audiovisual emotion recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3687-3691 (2013).