Detecting Emotions in Human Voice

Nineli Lashkarashvili, Lela Mirtskhulava

Department of Computer Science, Iv. Javakhishvili Tbilisi State University / San Diego State University,
Tbilisi, Georgia.

Email: nlashkarashvil6202@sdsu.edu; lela.mirtskhulava@tsu.ge

Abstract - **Expressing emotions is one of the most important factors in human communications. Words are not the only clues for conveying emotional information. Vocal features like timbre, loudness, tone, pitch, and facial expressions play a huge role. If there would be a good tool for recognizing human emotions this would make it possible to acquire machine intelligence with emotional intelligence. In this paper, we conducted extensive analysis of the impact of the number of MFCCs on several models' performances. We created 3 different models for this task: 2D Convolutional Neural Network with 4 convolution blocks and the LSTM and GRU models consisting of 256 neurons. We used the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) for training the models. We compare our models with the previous results and find that our best model for speech dataset has a 19% improvement. Two of our models for the song dataset achieved 78.4% accuracy.**

Keywords - emotion detection, mfccs, convolutional neural network, long-short term memory, gated recurrent unit

I. INTRODUCTION

Humans are social beings and our life is full of expressing emotions in day-to-day life. In order to fully understand the emotional context, several factors should be considered, including facial expressions, tone and choice of words. As emotions are inevitable in human's interactions, it is not surprising that there have been numerous attempts to implement the tools that would be capable of recognizing human emotions; this would facilitate the process of human-computer interactions and can be helpful for people with emotional disorders. [10]

Nowadays, the usage of voice assistant machines is increasing more and more. Virtual

Private Assistants such as Google Assistant, Siri and Alexa became a part of humans' lives. We are living in an era when voice-controlled services proliferate throughout the world. Nevertheless, they lack the ability to understand humans' emotional state. Thus, recognizing emotional signals will make a huge advance. [1]

Detecting emotions is not feasible, as several factors constitute the emotional context. These factors can be divided into verbal and nonverbal. In verbal, we can include sentences or words, while in non-verbal - tone of the voice, facial expressions, timbre and pitch.

The attempts to make the automatic recognition of emotions possible have started long ago (in the early 1990s) [3]. Some of the previous works focused on detecting emotions using only verbal or nonverbal emotional features, others have tried to use both of them (also called bimodal or multimodal recognition systems) [15].

In this paper, we focus on detecting emotions in the human voice without considering textual information. We use the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) for model training. This data contains [9]:"24 professional actors (12 female, 12 male) vocalizing two lexically-matched statements." We use only audio files. After Pre-processing raw audio files, we use Mel-Frequency Cepstral Coefficients to extract features. Then we are feeding the preprocessed data to our models.

In this paper, we have two research questions: 1. How does the number of MFCCs affect the performance of different models? 2. How do the results of our models compare with previous attempts?

Here is the short guideline for what will be covered in the next sections: firstly, we provide short summary of the previous studies (Related Work), then we detail input representation and model architectures (Methods and Data), after that we perform exhaustive analysis of the results (Results and Discussions), and lastly, we conclude our findings and suggest how this work can be extended in the future (Conclusions and Future).

II. RELATED WORK

Detecting emotions is an arduous task and numerous previous researchers tried to improve/solve at least one aspect of this massive problem.

It was found that timbre, loudness, pitch, and tone are very important for detecting emotion. These features vary across different emotions. Even though these variables can depend on the personal vocal characteristics, it was observed that for the nervous/panicked state, mean values of pitch/tone, the time between words and timbre ascend increase [4]. Various models ranging from HMM (Hidden Markov Models) to Transformers have been used to detect human emotions. [13-14] [16]

As we mentioned above, the audio data is transformed into MFCCs and then is fed to our models. Mel Frequency Cepstral Coefficients are widely used for speech recognition and they can also be applied to solve the emotion recognition problem. As Mel-Frequency Cepstral Coefficients are time-series data they need to be transformed and framed, so that they can be fed to Convolutional Neural Networks. [6] [8] [12] [15-16]

The results differed for the previous models. Some researchers decided to combine both speech features and transcriptions. The results showed that the combined CNN models based on both text and speech features had the highest overall accuracy, namely, 75.1 % (Text & Spectrogram) and 76.1% (Text & MFCC). [15] Several different models have been compared: LSTMs, CNNs, HMMs and DNNs with different input data fed to them: Log Mel Spectrograms/MFCCs/Pure Audio/MFCC with Deltas. The best results were achieved using CNNs with Log Mel Spectrograms' features [16]: "On the 14-class (2 genders × 7 emotions) classification task, an accuracy of 68% was achieved with a 4-layer 2 dimensional CNN using the Log-Mel Spectrogram features." Simple ANNs and neural network pattern recognition were used for supervising training and testing processes. The data contained German sentences and 7 different emotions. Feeding as an input MFCC + Cepstrum+ Frequency scaled MFCC showed the best results (85.7%). But the model completely failed in recognizing the sad emotional state. [8]

Additionally, Detecting emotions is very important to improve IoT voice-enabled services. From another perspective, voice signals convey important information about the user that can compromise user privacy. So it is important to maintain user privacy as well as to improve IoT voice-enabled services using emotion recognition. GANS (Generative Adversarial Networks) can be used to solve this issue using voice conversion [1]. Detecting emotions in human voice can improve various applications. The emotional state can change HRV (Heart Rate Variation). Human speech characteristics, including emotional state, can be used to measure HRV. Classical machine learning models have very poor performance when it comes to emotion detection.

Some studies focused on detecting emotions using both vocal and visual features (multi-cue approaches) [10] [12].

In order to recognize emotions in a human voice several datasets have been developed. The emotion labeling process can be quite difficult and mislabeled data can lead to poor performance. There are several open source datasets that can be used for emotion detection tasks: the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Suppes Brain Lab Psychotherapy EEG Dataset, and Indian EmoSpeech Command Dataset, [2] [3] [9]

In this paper we extended the work of [16] and unlike previous researches we analyzed the impact of the number of MFCCs on the models. Additionally, the previous studies mainly focused on detecting emotions on the speech audio files. Here we used both song and speech data.

III. METHODS AND DATA

We utilize three different models for emotion detection: the Convolutional Neural Network consisting of 4 convolution blocks, and LSTM and GRU with 256 neurons. In order to extract meaningful vocal features, we use MFCCs, as they are very successful in speech recognition tasks. Both data preparation part and model architectures were motivated by the findings of previous researches. These methodologies will be detailed in this section.

A. Dataset

In this paper, the RAVDESS dataset was used. To the best of our knowledge, in the previous models song data was not used for predicting emotions, and in this paper, we decided to use both speech and song datasets. The speech dataset contains 8 emotions: happy, fearful, calm, disgust, neutral, sad, surprised and angry. The song dataset contains only 6 emotions: happy, fearful, calm, neutral, sad, angry. In the dataset 12 actors and 12 actresses repeated two different sentences for these emotions.[9]

B. Data Preparation

In order to prepare data firstly each sample audio was trimmed and all sound below 30 decibels was considered silence. After this step we extracted different number of MFCCs: 10, 12, 13, 20, 30, 40, 50 and 60. In order to get MFCC coefficients, raw audio data goes through several steps [11]:

i. Pre-emphasis: Filter to Emphasize Higher Frequency:

$$H(z) = 1 - \beta z^{-1}$$

Here β stands for the slope control parameter,

ii. Discrete Fourier Transform (DFT):

$$X(k) = \sum_{n=0}^{N-1} x(n)^{\frac{-2j\pi k}{N}} \quad 0 \leq k \leq N-1$$

iii. Mel Spectrum: DFT Transformed is Passed

Through Mel-Filter Bank:

- Mel Scale:

$$f_{mel} = 2595 log_{10}(1 + \frac{f}{700})$$

- Triangular Weights Multiplied by Magnitum Spectrum (X(k)):

$$s(m) = \sum_{k=0}^{N-1} [|X(k)|^2 H_m(k)]$$

$$0 \leq m \leq M-1$$

Here M is for the number of Mel filters. $H_m(k)$ is the weight:

$$H_m(k) = \begin{cases} 0, & k \leq f(m-1) \\ \frac{2(k-f(m-1))}{f(m)-f(m-1)}, & f(m-1) \leq k \leq f(m) \\ \frac{2(f(m+1)-k)}{f(m+1)-f(m)}, & f(m) < k \leq f(m+1) \\ 0, & k > f(m+1) \end{cases}$$

iv. Discrete Cosine Transform (DCT):

$$c(n) = \sum_{m=0}^{M-1} log_{10}(s(m)) \cos(\frac{\pi n(m-0.5)}{M})$$

for $n = 0, \dots, C-1$

IV. THE MODEL

2D Convolutional Neural Network consists of 4 convolution blocks followed by the fully connected layer for predicting emotion. Each convolution block contains convolution and average pooling layers.

The LSTM model includes a single LSTM layer and dense layer. The LSTM layer has 256 neurons.

Similar to the LSTM model, GRU model consists of a single GRU layer with 256 neurons and a dense layer for emotion prediction.

V. EXPERIMENTAL SETUP

Training set contained 10 actors/actresses data and both validation and test set 1 actor/actress audio files. All of the models were trained for 100 epochs. For CNN we chose the model with the minimum validation loss for testing. We used elu, relu and softmax activation functions. In the case of song/speech data we had different numbers of classes (8/6) and the number of output neurons differed respectively.

We decided to use 256 neurons for both LSTM and GRU in order to make them comparable with the previous study. As the evaluation metric we utilized accuracy, optimizer was Adam and the loss function was sparse categorical cross entropy. We used the Tensorflow framework to implement the models and librosa to prepare the data.

VI. RESULTS AND DISCUSSION

In this section, we analyze the results of the experiments that we conducted to test our models on a varying number of MFCCs. We evaluate our models based on how accurately they predict emotional state. We compare our results with the previous research [16] and find that even though we don't treat female/male emotions as separate classes, models are still able to give a reasonable accuracy, furthermore, compared to the previous CNN model, we have a 21% gain on the test set.

A. Research Question 1

For every different number of MFCCs 3 distinctive models were tried: 2D CNN, LSTM and GRU. The test accuracies for the speech and song data are displayed in Fig. 1 and Fig 2.

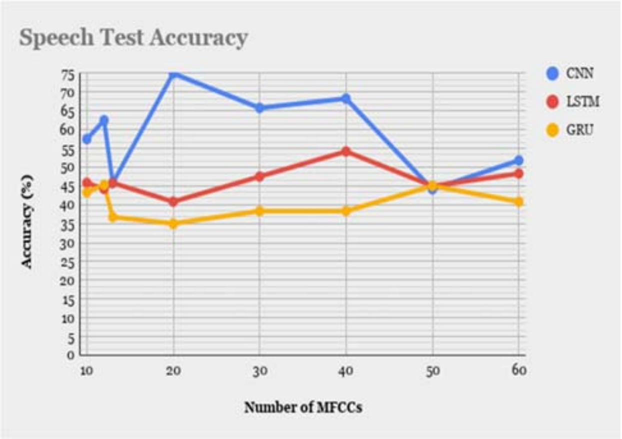

Fig. 1. Results: Speech Data

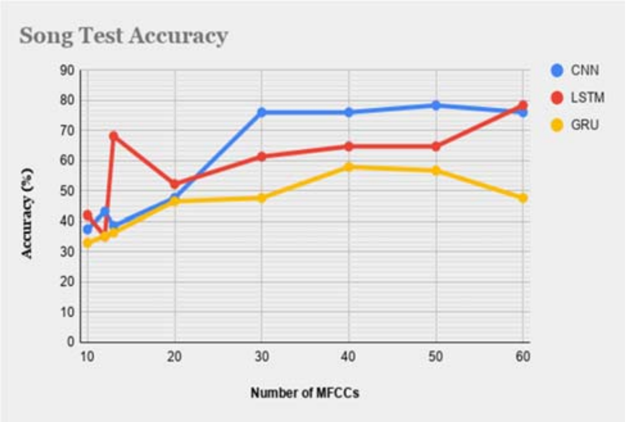

Fig. 2. Results: Song Data

We find that one of the most common choices of the number of MFCCs in the speech recognition and previous works, 12-13 MFCCs, mostly show poor results compared to the other choices. This is not surprising as the formulation of MFCC doesn't provide a deterministic way to compute the optimal choice for it. Our results show that the number of MFCCs can have a significant influence on the model performance and provide guidelines for the value ranges that are mightier to improve the results for the different architectures.

A1. Convolutional Neural Networks

Convolutional Neural Networks have shown the best results for both speech and song datasets. In the case of speech data, the best accuracy score was 75% when the number of MFCCs was set to 20. As it is clear from the Fig. 1, for the speech data model had the highest performance when the number of MFCCs was 20, 30, 40. For the song data, the model showed best performance(78.4%) when the number of MFCCs was 50. According to the chart, the best results were achieved when $nmfcc \geq 20$. The results supports the idea that on the speech datasets, Convolutional Neural Networks will give the best results when $20 \leq nmfcc \leq 40$ and for song: $nmfcc \geq 20$.

A2. LSTM

LSTMs have shown the best result for song dataset(78.4%) when the $nmfcc = 60$, which is remarkable, as very simple LSTM was used. Based on the figures, for the speech data we have highest accuracies when$30 \leq nmfcc \leq 60$ and the highest accuracy was 54.2%($nmfcc = 40$). In the case of song data if we would not consider $nmfcc = 13$, which seems like an outlier, we have the highest accuracies when $30 \leq nmfcc \leq 60$.

A3. GRU

GRUs have shown the lowest performance. For speech/song data the highest accuracies were 45.3%/56.8% for $nmfcc = 12/ nmfcc = 13$. In the case of song data when $30 \leq nmfcc \leq 60$ we have the highest values and for the speech data when $nmfcc = 10/nmfcc = 12$.

B. Research Question 2

Based on our results, our models have outperformed previous CNN models (using MFCC features) by a 21% gain in the case of CNN and 0.2% in the case of LSTM, see Table I. While for the LSTM we have only slight improvement, in the case of CNN proper choice of $nmfcc$ improves performance drastically. Here we included only those models from the previous study where MFCCs were used as input features.

In the previous study, for these models only 6 classes were used: angry, sad, neutral, disgust, happy and fearful [16]. We had classes for all 8 emotions that were present in the speech dataset. In Table I we included the results only for the speech data.

TABLE I. MODEL COMPARISON

Input	N MFCC	N Classes	Model	ACC
MFCC+Delta+Delta Delta Coefficient	29	6	CNN	47%
MFCC+Delta Coefficients	29	6	CNN	53%
MFCC+Delta Coefficients	40	6	DNN	55%
MFCC	**20**	**8**	**CNN**	**75%**
MFCC	25	12	LSTM	54%
MFCC	**40**	**8**	**LSTM**	**54.2%**

VII. CONCLUSION AND FUTURE WORK

In this paper we performed extensive analysis of how the number of MFCCs impact model performances. We used RAVDESS speech and song audio files in order to test our models. We trained 3 models: CNN with 4 blocks and LSTM and GRU with 256 neurons. We found that proper choice of MFCCs can improve the results drastically. We have a 19% gain compared to the previous results on the speech data and for the song dataset we achieved 78.4% accuracy.

We believe it would be interesting to utilize larger files since this data contained audio files ranging between 4-5 seconds. We consider that in this case it would be useful to use bimodal or multimodal architectures in which transcriptions would also be included.

REFERENCES

[1] Aloufi, R., Haddadi, H.,Boyle, D. (2019). Emotionless: privacy-preserving speech analysis for voice assistants. arXiv preprintarXiv:1908.03632.

[2] Banga, S., Upadhyay, U., Agarwal, P., Sharma, A.,Mukherjee, P.(2019). Indian EmoSpeech Command Dataset: A dataset for emotion based speech recognition in the wild. arXiv preprint arXiv:1910.13801.

[3] Crangle, C. E., Wang, R., Perreau-Guimaraes, M., Nguyen, M. U.,Nguyen, D. T., Suppes, P. (2019). Machine learning for the recognition of emotion in the speech of couples in psychotherapy using the Stanford Suppes Brain Lab Psychotherapy Dataset. arXiv preprint arXiv:1901.04110.

[4] Chung, J., Gulcehre, C., Cho, K.,Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

[5] Dasgupta, P. B. (2017). Detection and analysis of human emotions through voice and speech pattern processing. arXiv preprintarXiv:1710.10198.

[6] Davletcharova, A., Sugathan, S., Abraham, B., James, A. P. (2015). Detection and analysis of emotion from speech signals. Procedia Computer Science, 58, 91-96.

[7] Hochreiter, S.,Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[8] Lalitha, S., Geyasruti, D., Narayanan, R., Shravani, M. (2015). Emotion detection using MFCC and cepstrum features. Procedia Computer Science, 70, 29-35.

[9] Livingstone, S. R., Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North AmericanEnglish. PloS one, 13(5), e0196391.

[10] Ley, M., Egger, M., Hanke, S. (2019). Evaluating Methods for Emotion Recognition based on Facial and Vocal Features. In AmI(Workshops/Posters) (pp. 84-93).

[11] Rao, K. S.,Manjunath, K. E. (2017). Speech recognition using articulatory and excitation source features. Springer.(pp. 85-87).

[12] Ristea, N. C., Dutu, L. C.,Radoi, A. (2019, October). Emotio recognition system from speech and visual information based on convolutional neural networks. In 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD) (pp. 1-6). IEEE.

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

[14] Siriwardhana, S., Kaluarachchi, T., Billinghurst, M., Nanayakkara, S.(2020). Multimodal Emotion Recognition With Transformer-Based Self Supervised Feature Fusion. IEEE Access, 8, 176274-176285.

[15] Tripathi, S., Kumar, A., Ramesh, A., Singh, C., Yenigalla, P. (2019).Deep learning based emotion recognition system using speech features and transcriptions. arXiv preprint arXiv:1906.05681.

[16] Venkataramanan, K., Rajamohan, H. R. (2019). Emotion recognition from speech. arXiv preprint arXiv:1912.10458.