Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions

Suraj Tripathi, Abhay Kumar, Abhiram Ramesh,
Chirag Singh, Promod Yenigalla

Paper# 307

Presented by

Abhay Kumar

Introduction

- This paper proposes a speech emotion recognition method based on speech features and speech transcriptions (text).

- Speech features such as Spectrogram and Mel-frequency Cepstral Coefficients (MFCC) help retain emotion related low-level characteristics in speech whereas text helps capture semantic meaning, both of which help in different aspects of emotion detection.

- The combined MFCC-Text Convolutional Neural Network (CNN) model proved to be the most accurate in recognizing emotions in IEMOCAP data.

Proposed Approach

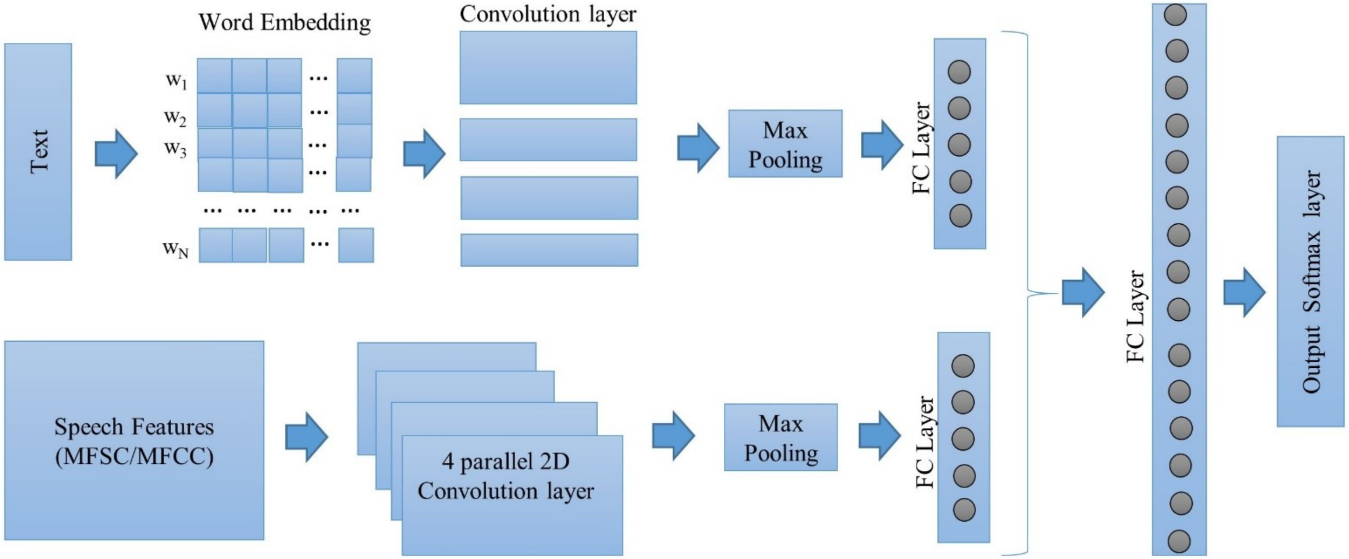


Fig. 1. Representative CNN architecture for Speech Emotion Recognition using Speech Features and Transcriptions

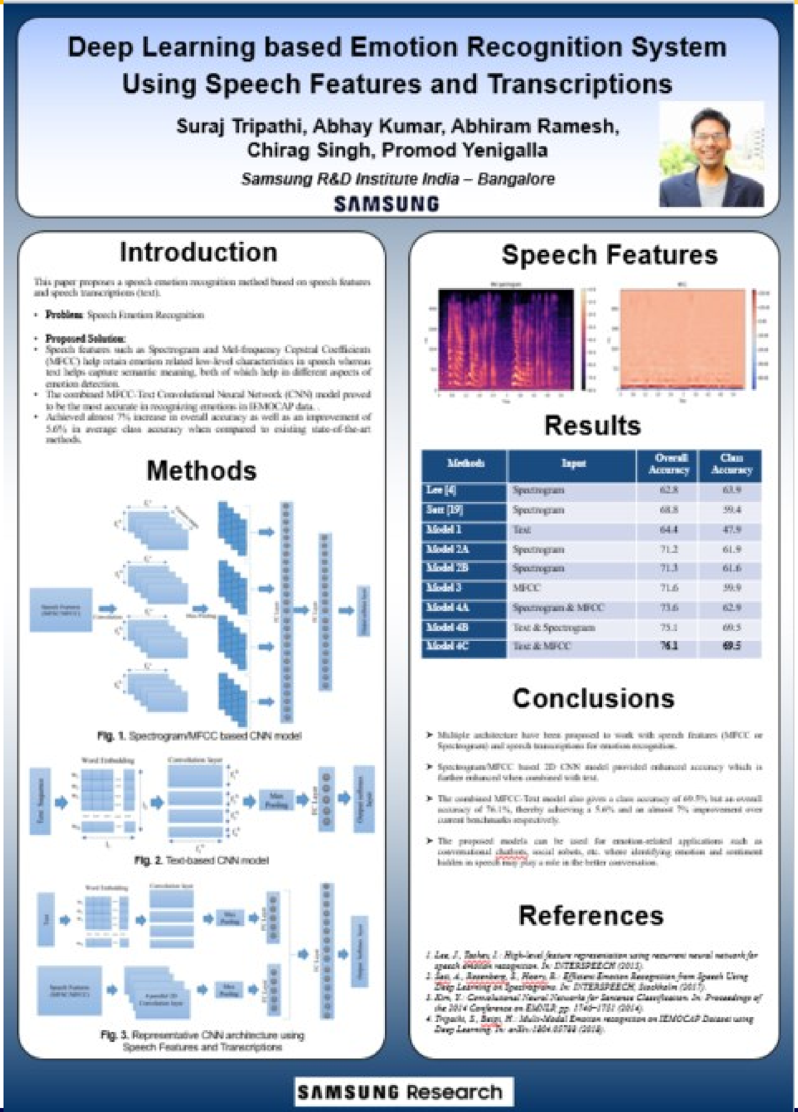
- ➢ Achieved almost 7% increase in overall accuracy as well as an improvement of 5.6% in average class accuracy when compared to existing state-of-the-art methods.

Results

Methods	Input	Overall Accuracy	Class Accuracy
Lee [1]	Spectrogram	62.8	63.9
Satt [2]	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**

1. *Lee, J., Tashev, I.: High-level feature representation using recurrent neural network for speech emotion recognition. In: INTERSPEECH (2015).*
2. *Satt, A., Rozenberg, S., Hoory, R.: Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In: INTERSPEECH, Stockholm (2017).*

Poster Screenshot



Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions

Suraj Tripathi, Abhay Kumar, Abhiram Ramesh,
Chirag Singh, Promod Yenigalla

Samsung R&D Institute India – Bangalore

SAMSUNG

Introduction

This paper proposes a speech emotion recognition method based on speech features and speech transcriptions (text).

- **Problem:** Speech Emotion Recognition

- **Proposed Solution:**
- Speech features such as Spectrogram and Mel-frequency Cepstral Coefficients (MFCC) help retain emotion related low-level characteristics in speech whereas text helps capture semantic meaning, both of which help in different aspects of emotion detection.
- The combined MFCC-Text Convolutional Neural Network (CNN) model proved to be the most accurate in recognizing emotions in IEMOCAP data. .
- Achieved almost 7% increase in overall accuracy as well as an improvement of 5.6% in average class accuracy when compared to existing state-of-the-art methods.

Methods

Fig. 1. Spectrogram/MFCC based CNN model

Fig. 2. Text-based CNN model

Fig. 3. Representative CNN architecture using Speech Features and Transcriptions

Speech Features

Results

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**

Conclusions

- Multiple architecture have been proposed to work with speech features (MFCC or Spectrogram) and speech transcriptions for emotion recognition.
- Spectrogram/MFCC based 2D CNN model provided enhanced accuracy which is further enhanced when combined with text.
- The combined MFCC-Text model also gives a class accuracy of 69.5% but an overall accuracy of 76.1%, thereby achieving a 5.6% and an almost 7% improvement over current benchmarks respectively.
- The proposed models can be used for emotion-related applications such as conversational chatbots, social robots, etc. where identifying emotion and sentiment hidden in speech may play a role in the better conversation.

References

1. Lee, J., Tashev, I.: *High-level feature representation using recurrent neural network for speech emotion recognition. In: INTERSPEECH (2015).*
2. Satt, A., Rozenberg, S., Hoory, R.: *Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In: INTERSPEECH, Stockholm (2017).*
3. Kim, Y.: *Convolutional Neural Networks for Sentence Classification. In: Proceedings of the 2014 Conference on EMNLP, pp. 1746–1751 (2014).*
4. Tripathi, S., Beigi, H.: *Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning. In: arXiv:1804.05788 (2018).*

SAMSUNG Research