Speech Emotion Recognition Using Speech Feature and Word Embedding

Bagus Tris Atmaja
Kiyoaki Shirai
Masato Akagi
bagus@jaist.ac.jp

APSIPA 2019

Code, paper, and slide available at `https://github.com/bagustris/Apsipa2019_SpeechText`

Outline

Outline

Outline

Outline

Outline

Motivation

- Emotion can be automatically recognized by many modalities: face, speech, and motion of the body's parts.
- To obtain better recognition results, multimodal features are usually used together. (e.g. audio+visual, audio+text, or audio+visual+text).
- In the case of speech, both speech and text features (from transcription) can be used.
- We proposed to use speech feature (within the speech region by removing silence part) and text feature, from word embedding, to improve speech emotion recognition performance.

Motivation

- Emotion can be automatically recognized by manymodalities: face, speech, and motion of the body's parts.
- To obtain better recognition results, multimodal features are usually used together. (e.g. audio+visual, audio+text, or audio+visual+text).
- In the case of speech, both speech and text features (from transcription) can be used.
- We proposed to use speech feature (within the speech region by removing silence part) and text feature, from word embedding, to improve speech emotion recognition performance.

Motivation

- Emotion can be automatically recognized by manymodalities: face, speech, and motion of the body's parts.
- To obtain better recognition results, multimodal features are usually used together. (e.g. audio+visual, audio+text, or audio+visual+text).
- In the case of speech, both speech and text features (from transcription) can be used.
- We proposed to use speech feature (within the speech region by removing silence part) and text feature, from word embedding, to improve speech emotion recognition performance.

Motivation

- Emotion can be automatically recognized by manymodalities: face, speech, and motion of the body's parts.
- To obtain better recognition results, multimodal features are usually used together. (e.g. audio+visual, audio+text, or audio+visual+text).
- In the case of speech, both speech and text features (from transcription) can be used.
- We proposed to use speech feature (within the speech region by removing silence part) and text feature, from word embedding, to improve speech emotion recognition performance.

Motivation

- Emotion can be automatically recognized by manymodalities: face, speech, and motion of the body's parts.
- To obtain better recognition results, multimodal features are usually used together. (e.g. audio+visual, audio+text, or audio+visual+text).
- In the case of speech, both speech and text features (from transcription) can be used.
- We proposed to use speech feature (within the speech region by removing silence part) and text feature, from word embedding, to improve speech emotion recognition performance.

Motivation

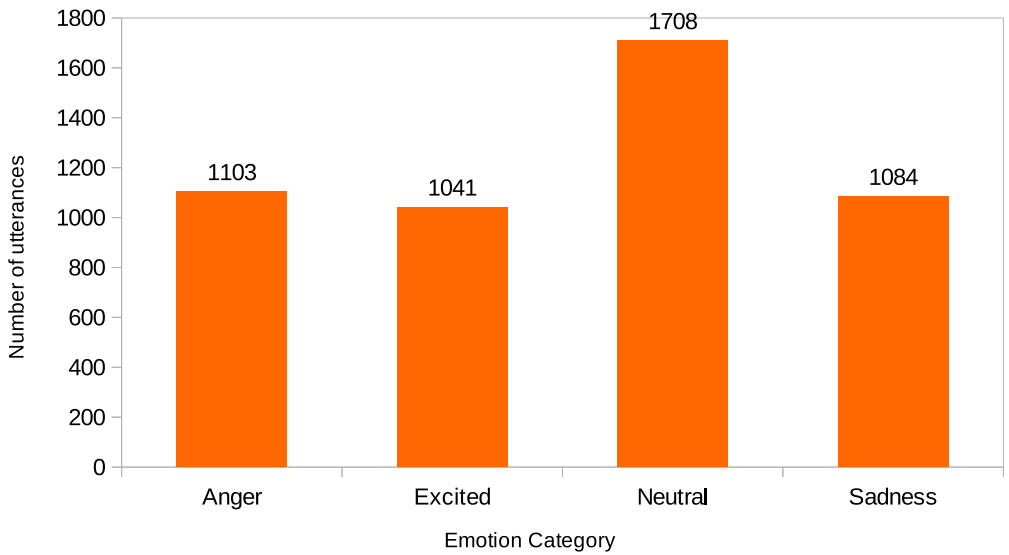
- Emotion can be automatically recognized by manymodalities: face, speech, and motion of the body's parts.
- To obtain better recognition results, multimodal features are usually used together. (e.g. audio+visual, audio+text, or audio+visual+text).
- In the case of speech, both speech and text features (from transcription) can be used.
- We proposed to use speech feature (within the speech region by removing silence part) and text feature, from word embedding, to improve speech emotion recognition performance.

Purpose

How to improve accuracy of categorical speech emotion recognition by **fusing both speech and text features** using **deep neural network**?

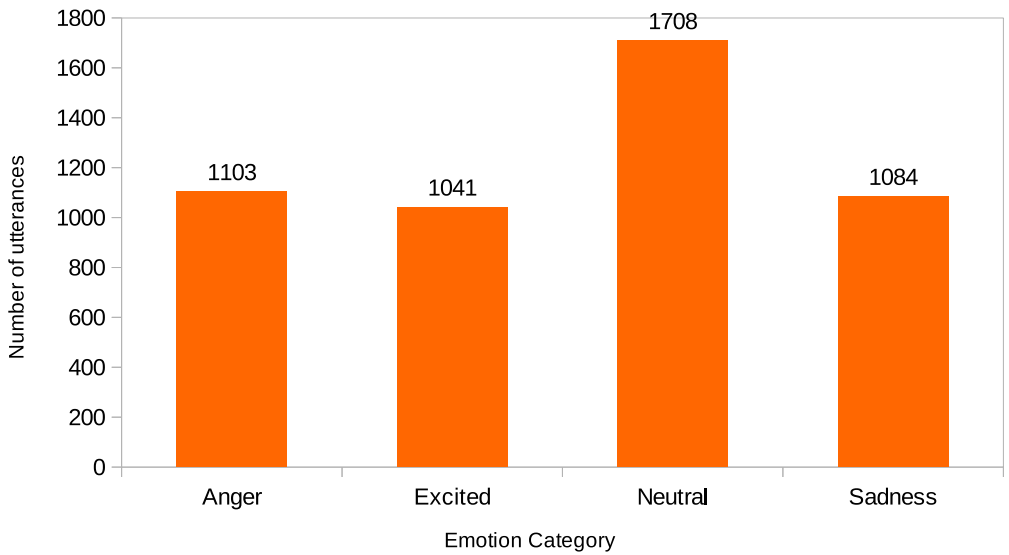
Dataset

- Speech and text transcription from Interactive Emotional Motion Capture (IEMOCAP).
- 4936 utterances from 10039 turns.
- Four emotion categories to balance samples.
- Training - test split = 80 : 20



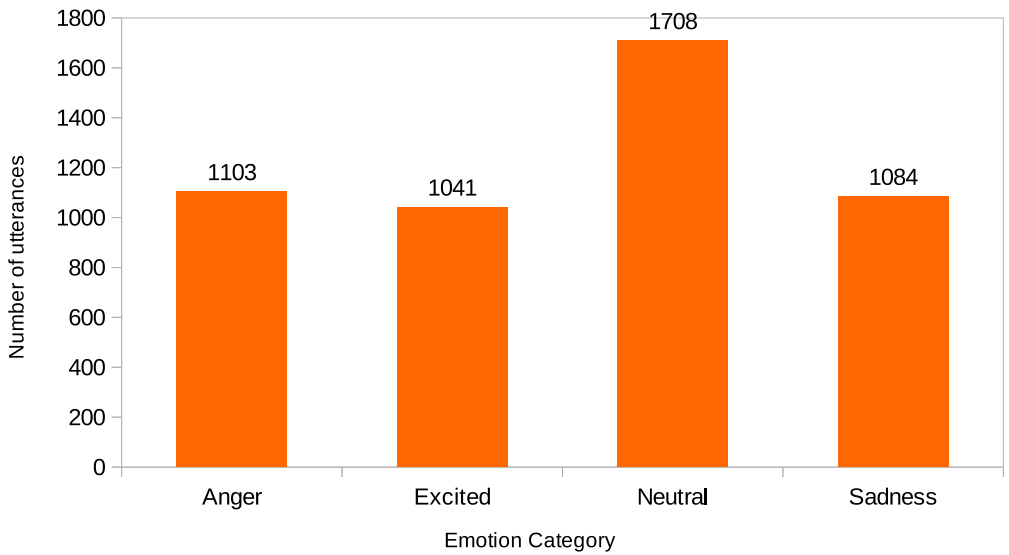
Dataset

- Speech and text transcription from Interactive Emotional Motion Capture (IEMOCAP).
- 4936 utterances from 10039 turns.
- Four emotion categories to balance samples.
- Training - test split = 80 : 20



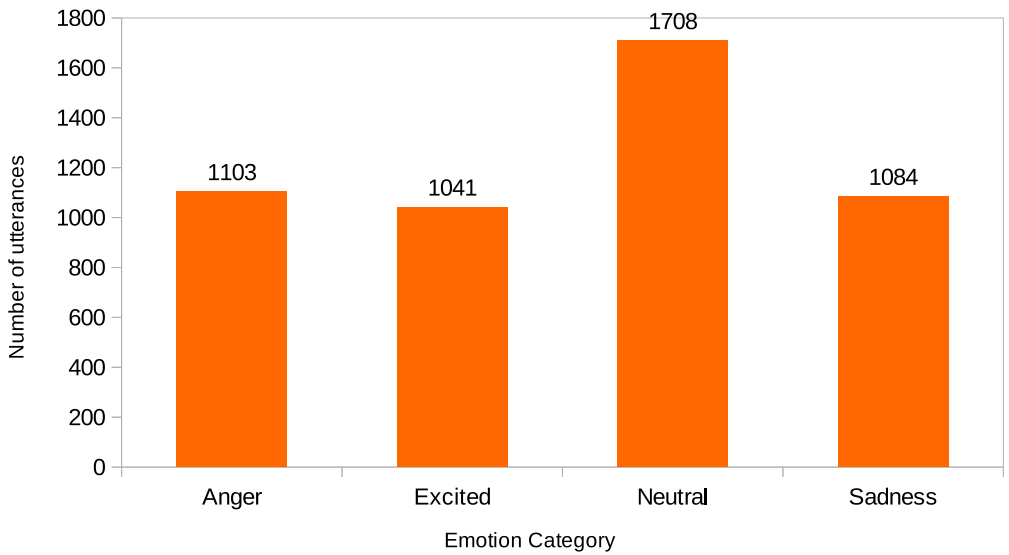
Dataset

- Speech and text transcription from Interactive Emotional Motion Capture (IEMOCAP).
- 4936 utterances from 10039 turns.
- Four emotion categories to balance samples.
- Training - test split = 80 : 20



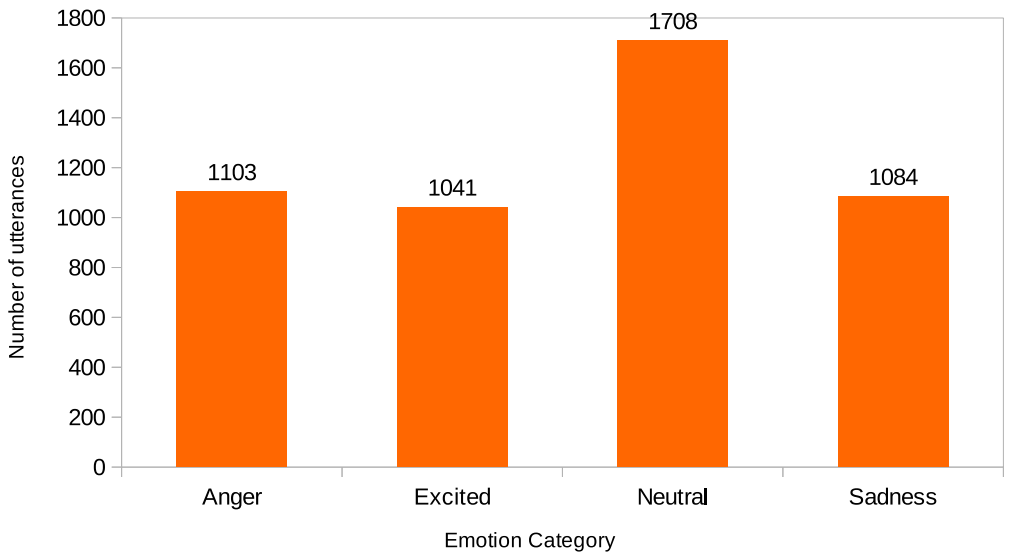
Dataset

- Speech and text transcription from Interactive Emotional Motion Capture (IEMOCAP).
- 4936 utterances from 10039 turns.
- Four emotion categories to balance samples.
- Training - test split = 80 : 20



Dataset

- Speech and text transcription from Interactive Emotional Motion Capture (IEMOCAP).
- 4936 utterances from 10039 turns.
- Four emotion categories to balance samples.
- Training - test split = 80 : 20



Pre-processing: Silence Removal

Result: Speech signal without silence
TH=threshold;
min_dur=min_duration;
$n_i = 0$;
while $y_i(x) \leq TH * y_{max}(x)$ **do**
    $n_i = n_i + 1$;
    if $n_i \geq min\_dur$ **then**
        remove n_i samples;
    else
        keep samples;
    end
end

- Threshold = 0.001%
- minimum duration = 100 ms

Acoustic Features

- 3 time domain features: Energy, Zero Crossing Rate, Energy Entropy.
- 5 frequency domain features: Spectral Centroid, Spectral Spread, Spectral Roll-Off, Spectral Flux, Spectral Entropy.
- 13 MFCCs
- 13 Chromas
- Total: 34 acoustic features

Acoustic Features

- 3 time domain features: Energy, Zero Crossing Rate, Energy Entropy.
- 5 frequency domain features: Spectral Centroid, Spectral Spread, Spectral Roll-Off, Spectral Flux, Spectral Entropy.
- 13 MFCCs
- 13 Chromas
- Total: 34 acoustic features

Acoustic Features

- 3 time domain features: Energy, Zero Crossing Rate, Energy Entropy.
- 5 frequency domain features: Spectral Centroid, Spectral Spread, Spectral Roll-Off, Spectral Flux, Spectral Entropy.
- 13 MFCCs
- 13 Chromas
- Total: 34 acoustic features

Acoustic Features

- 3 time domain features: Energy, Zero Crossing Rate, Energy Entropy.
- 5 frequency domain features: Spectral Centroid, Spectral Spread, Spectral Roll-Off, Spectral Flux, Spectral Entropy.
- 13 MFCCs
- 13 Chromas
- Total: 34 acoustic features

Acoustic Features

- 3 time domain features: Energy, Zero Crossing Rate, Energy Entropy.
- 5 frequency domain features: Spectral Centroid, Spectral Spread, Spectral Roll-Off, Spectral Flux, Spectral Entropy.
- 13 MFCCs
- 13 Chromas
- Total: 34 acoustic features

Acoustic Features

- 3 time domain features: Energy, Zero Crossing Rate, Energy Entropy.
- 5 frequency domain features: Spectral Centroid, Spectral Spread, Spectral Roll-Off, Spectral Flux, Spectral Entropy.
- 13 MFCCs
- 13 Chromas
- Total: 34 acoustic features

Acoustic Feature Extraction

- Remove silence part by 0.001% threshold and 100 ms of minimum duration.
- Divide each speech utterances into frames (windows) of 200 ms.
- For each frame, do feature extraction (34 features).
- Move to next window with 50% overlap (100 ms).
- Collect feature until 100 frames for each utterance (size=(100,34)).
- Collect feature for all utterances (size=(4936, 100, 34)).
- Classifier: BLSTM, BLSTM + Attention

Acoustic Feature Extraction

- Remove silence part by 0.001% threshold and 100 ms of minimum duration.
- Divide each speech utterances into frames (windows) of 200 ms.
- For each frame, do feature extraction (34 features).
- Move to next window with 50% overlap (100 ms).
- Collect feature until 100 frames for each utterance (size=(100,34)).
- Collect feature for all utterances (size=(4936, 100, 34)).
- Classifier: BLSTM, BLSTM + Attention

Acoustic Feature Extraction

- Remove silence part by 0.001% threshold and 100 ms of minimum duration.
- Divide each speech utterances into frames (windows) of 200 ms.
- For each frame, do feature extraction (34 features).
- Move to next window with 50% overlap (100 ms).
- Collect feature until 100 frames for each utterance (size=(100,34)).
- Collect feature for all utterances (size=(4936, 100, 34)).
- Classifier: BLSTM, BLSTM + Attention

Acoustic Feature Extraction

- Remove silence part by 0.001% threshold and 100 ms of minimum duration.
- Divide each speech utterances into frames (windows) of 200 ms.
- For each frame, do feature extraction (34 features).
- Move to next window with 50% overlap (100 ms).
- Collect feature until 100 frames for each utterance (size=(100,34)).
- Collect feature for all utterances (size=(4936, 100, 34)).
- Classifier: BLSTM, BLSTM + Attention

Acoustic Feature Extraction

- Remove silence part by 0.001% threshold and 100 ms of minimum duration.
- Divide each speech utterances into frames (windows) of 200 ms.
- For each frame, do feature extraction (34 features).
- Move to next window with 50% overlap (100 ms).
- Collect feature until 100 frames for each utterance (size=(100,34)).
- Collect feature for all utterances (size=(4936, 100, 34)).
- Classifier: BLSTM, BLSTM + Attention

Acoustic Feature Extraction

- Remove silence part by 0.001% threshold and 100 ms of minimum duration.
- Divide each speech utterances into frames (windows) of 200 ms.
- For each frame, do feature extraction (34 features).
- Move to next window with 50% overlap (100 ms).
- Collect feature until 100 frames for each utterance (size=(100,34)).
- Collect feature for all utterances (size=(4936, 100, 34)).
- Classifier: BLSTM, BLSTM + Attention

Acoustic Feature Extraction

- Remove silence part by 0.001% threshold and 100 ms of minimum duration.
- Divide each speech utterances into frames (windows) of 200 ms.
- For each frame, do feature extraction (34 features).
- Move to next window with 50% overlap (100 ms).
- Collect feature until 100 frames for each utterance (size=(100,34)).
- Collect feature for all utterances (size=(4936, 100, 34)).
- Classifier: BLSTM, BLSTM + Attention

Acoustic Feature Extraction

- Remove silence part by 0.001% threshold and 100 ms of minimum duration.
- Divide each speech utterances into frames (windows) of 200 ms.
- For each frame, do feature extraction (34 features).
- Move to next window with 50% overlap (100 ms).
- Collect feature until 100 frames for each utterance (size=(100,34)).
- Collect feature for all utterances (size=(4936, 100, 34)).
- Classifier: BLSTM, BLSTM + Attention

Text Feature Extraction

- Tokenize utterance/sentence into words (max length = 537).
- For each token, generate 300 dimension of vector.
- Collect vector for each sentence (size=(537, 300)).
- Collect text feature for all sentences (size=(4936, 537, 300)).
- Classifier: CNN, LSTM, LSTM + Attention.

Text Feature Extraction

- Tokenize utterance/sentence into words (max length = 537).
- For each token, generate 300 dimension of vector.
- Collect vector for each sentence (size=(537, 300)).
- Collect text feature for all sentences (size=(4936, 537, 300)).
- Classifier: CNN, LSTM, LSTM + Attention.

Text Feature Extraction

- Tokenize utterance/sentence into words (max length = 537).
- For each token, generate 300 dimension of vector.
- Collect vector for each sentence (size=(537, 300)).
- Collect text feature for all sentences (size=(4936, 537, 300)).
- Classifier: CNN, LSTM, LSTM + Attention.

Text Feature Extraction

- Tokenize utterance/sentence into words (max length = 537).
- For each token, generate 300 dimension of vector.
- Collect vector for each sentence (size=(537, 300)).
- Collect text feature for all sentences (size=(4936, 537, 300)).
- Classifier: CNN, LSTM, LSTM + Attention.

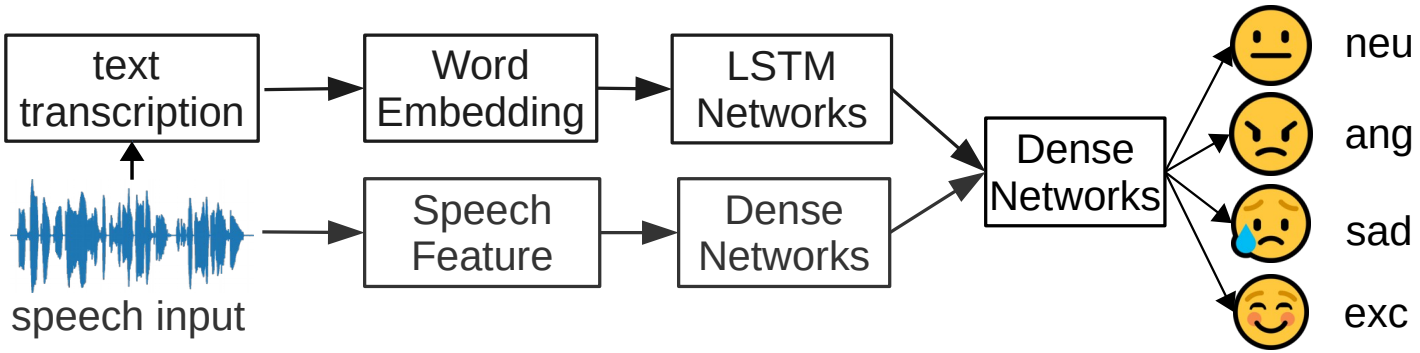
Text Feature Extraction

- Tokenize utterance/sentence into words (max length = 537).
- For each token, generate 300 dimension of vector.
- Collect vector for each sentence (size=(537, 300)).
- Collect text feature for all sentences (size=(4936, 537, 300)).
- Classifier: CNN, LSTM, LSTM + Attention.

Text Feature Extraction

- Tokenize utterance/sentence into words (max length = 537).
- For each token, generate 300 dimension of vector.
- Collect vector for each sentence (size=(537, 300)).
- Collect text feature for all sentences (size=(4936, 537, 300)).
- Classifier: CNN, LSTM, LSTM + Attention.

Early Fusion: Acoustic + Text Network



- Feature-Level (FL) fusion (also known as early fusion): combines features from different modalities before performing recognition.
- Variate network on each modality (text & acoustic networks), observe the result (accuracy).
- Network to be evaluated: CNN, LSTM, Dense.

Result: Unimodal

Model	Accuracy
Acoustic Network	
Whole speech + BLSTM	52.83%
Speech segment + BLSTM	55.26%
Whole speech + BLSTM + Attention	53.64%
Speech segment + BLSTM + Attention	58.29%
Text Networks	
CNN	65.69%
LSTM	66.59%
LSTM with Attention	68.01%

Result: Multimodal

#	text network	acoustic networks	accuracy
1	CNN	Dense	68.83%
2	LSTM	LSTM	69.13%
3	LSTM	Dense	75.49%

- Some combinations are tried, but those three are the highest one.
- For each network, the three same layers are used (CNN, LSTM, or Dense).
- The same two dense layers with 256 and 4 units are added to all combinations.
- Comparison between models are made by selecting the similar number of trainable parameters (5 million parameters)

Comparison with previous works

References	Method	Accuracy
Tripathi [1]	LSTM (GloVe Embedding) + Dense	69.74%
Yenigalla [2]	CNN (Spectrogram) + CNN (Phoneme Embedding)	73.9%
Yoon [3]	RNN + RNN	71.8%
Ours	LSTM + Dense	75.49%

[1] Tripathi, Samarth, and Homayoon Beigi. "Multi-modal emotion recognition on iemocap dataset using deep learning." arXiv preprint arXiv:1804.05788 (2018).

[2] P. Yenigalla et al., Speech Emotion Recognition Using Spectrogram & Phoneme Embedding, In Interspeech, pp. 3688-3692, 2018.

[3] Yoon, Seunghyun, Seokhyun Byun, and Kyomin Jung. "Multimodal speech emotion recognition using audio and text." In 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 112-118, 2018.

Conclusion

- Multimodal emotion recognition from acoustic and text is presented, the fuse of two features set is done using early fusion of acoustic network and text network.
- For each network, the architecture of three networks are evaluated i.e., text network, acoustic network, and concatenation network. The result shows that the highest performance was obtained using LSTM, dense, and dense network.
- Using multimodal feature, an improvement of accuracy of categorical emotion recognition is obtained. The improvement is about 17% from acoustic feature, and 7% from text feature.

Conclusion

- Multimodal emotion recognition from acoustic and text is presented, the fuse of two features set is done using early fusion of acoustic network and text network.
- For each network, the architecture of three networks are evaluated i.e., text network, acoustic network, and concatenation network. The result shows that the highest performance was obtained using LSTM, dense, and dense network.
- Using multimodal feature, an improvement of accuracy of categorical emotion recognition is obtained. The improvement is about 17% from acoustic feature, and 7% from text feature.

Conclusion

- Multimodal emotion recognition from acoustic and text is presented, the fuse of two features set is done using early fusion of acoustic network and text network.
- For each network, the architecture of three networks are evaluated i.e., text network, acoustic network, and concatenation network. The result shows that the highest performance was obtained using LSTM, dense, and dense network.
- Using multimodal feature, an improvement of accuracy of categorical emotion recognition is obtained. The improvement is about 17% from acoustic feature, and 7% from text feature.

Conclusion

- Multimodal emotion recognition from acoustic and text is presented, the fuse of two features set is done using early fusion of acoustic network and text network.
- For each network, the architecture of three networks are evaluated i.e., text network, acoustic network, and concatenation network. The result shows that the highest performance was obtained using LSTM, dense, and dense network.
- Using multimodal feature, an improvement of accuracy of categorical emotion recognition is obtained. The improvement is about 17% from acoustic feature, and 7% from text feature.