Interspeech 2018
2-6 September 2018, Hyderabad, India

Speech Emotion Recognition Using Spectrogram & Phoneme Embedding

Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, Jithendra Vepa

[1]Samsung R&D Institute India - Bangalore

promod.y@samsung.com, abhay1.kumar@samsung.com, suraj.tri@samsung.com,
c.singh@samsumg.com, sibsambhukar@gmail.com, jithendra.v@samsung.com

Abstract

This paper proposes a speech emotion recognition method based on phoneme sequence and spectrogram. Both phoneme sequence and spectrogram retain emotion contents of speech which is missed if the speech is converted into text. We performed various experiments with different kinds of deep neural networks with phoneme and spectrogram as inputs. Three of those network architectures are presented here that helped to achieve better accuracy when compared to the state-of-the-art methods on benchmark dataset. A phoneme and spectrogram combined CNN model proved to be most accurate in recognizing emotions on IEMOCAP data. We achieved more than 4% increase in overall accuracy and average class accuracy as compared to the existing state-of-the-art methods.

Index Terms: Spectrogram, phoneme, phoneme embedding, speech emotion recognition, CNN

1. Introduction

A significant segment of natural language processing problem involves speech or audio input, such as Chatbot applications for automatic reply or personal assistance, emotion analysis and classification, voice activated systems etc. For some of the above tasks speech or audio is converted into texts at the first step (using Automatic Speech Recognition or ASR mechanism) followed by classification and other learning operations based on text data. When an ASR module generates texts from an audio, it (generated text) becomes speaker independent. ASR takes care of differences in audio from different users using probabilistic acoustic and language models. As a result, robust text processing technologies proved successful in various applications. Text processing through deep learning is already an established domain of research [1]-[2]. Though the state-of-the-art ASR techniques work with high accuracy, we lose significant information while converting the audio into text. Specifically, the emotion component present in the audio signal is missed in the converted text. To address this issue, speech-based emotion recognition (SER) became a research of interest in last few decades.

A Typical Speech Emotion Recognition (SER) system works on extracting features from the speech followed by a classification task to predict various classes of emotions [3]. Commonly used features are spectral features, pitch frequency features, formant features and energy related feature [4]. Traditional classification task involves various ML (Machine Learning) techniques such as Bayesian Network model [5, 6], HMM (Hidden Markov Model) [7], SVM (Support Vector Machines) [8], GMM (Gaussian Mixture Model) [9] and Multi Classifier Fusion [10]. Since last decade, Deep Learning techniques contributed to significant breakthrough in various research areas including natural language understanding (NLU). Deep Belief Networks (DBN) for SER proposed by Kim et al. [11] and Zheng et al. [12] showed a significant improvement over the baseline models [5]-[10] that do not employ deep learning. Later, Han et al. [13] proposed a DNN-ELM to extract high level features from raw data and a single hidden layer neural net to identify the segment level SER with a limited improvement in accuracies. Zheng et al. [14] used spectrogram with Deep Convolution Neural Network, whereas Fayek et al. [15] tried using data augmentation along with a DNN for SER. Lee et al. [3] used a bi-directional LSTM model to train the feature sequences and achieved an emotion recognition accuracy of 62.8% on IEMOCAP [17] dataset which is a significant improvement over DNN-ELM [13]. Jin et al. [4] tried fusion of acoustic and lexical feature representations and was able to achieve accuracy close to 69.2% on 4-class IEMOCAP dataset. Similarly, the usage of Mel-scale spectrograms by Satt et al. [18] on a deep CNN and a combination of CNN and LSTM helped in achieving better result on IEMOCAP dataset.

In this work, we propose a robust technique of emotion classification using phoneme sequence and spectrogram. The objective is to retain the emotion information of the speech or audio in phoneme sequence and spectrogram and use a deep learning based emotion classifier. We present different deep network architectures to classify emotion using phoneme and spectrogram. Main contributions of the current work are:

- Generating Phoneme embedding and applying phoneme sequence based CNN Model for emotion classification
- Spectrogram based CNN model for emotion classification
- Combined spectrogram and Phoneme based CNN model for emotion classification.

2. Proposed Method

As discussed before, to avoid loss of emotion information in speech to text conversion, we consider phoneme and spectrogram as input for emotion classification using deep neural networks in this work. Experiments have been performed on both phoneme and spectrogram independently as well as together to achieve better accuracy. For different inputs a number of different deep NN architectures are used. The detail of these methods and architectures are discussed in the following section.

2.1. Model 1: CNN model with phoneme input

A phoneme is a unit of sound that distinguishes the pronunciation of same or different words. In an ASR system, a sequence phoneme is extracted from the spectrum of input speech or audio that helps to identify a word or set of words hidden in the speech. Both textual and non-textual (for example,

10.21437/Interspeech.2018-1811

laugh) information can be captured by different sets of phoneme. Hence, for a single word or sentence, there might be a number of different phoneme sequences depending on the way the word or the sentence was uttered. But, when the sequence of phoneme is converted into text through the decoding process of ASR, the variation in utterances is missed in the text. So, phoneme based emotion detection may work better than the text.

Like word or character based deep neural network models, phoneme based neural network models require embedding to represent each phoneme as numeric vector. A set of 47 phonemes (as presented in Table 1) is used in this work for representing an audio input as a sequence of phonemes. The embedding is generated using IEMOCAP speech data and word2vec model [19] as a part of model training. We use a dimension of 100 for phoneme embedding.

Table 1: Phoneme categories

Vowels	Consonants	Others
AA, AE, AH, AO, AW, AX, AXR, AY, EH, ER, EY, IH, IX, IY, OW, OY, UH, UW	B, CH, D, DD, DX, DH, F, G, HH, JH, K, KD, L, M, N, NG, P, R, S, SH, T, TD, TH, TS, V, W, Y, Z, ZH	SIL, LAUGHTER, LIPSMACK, GARBAGE, BREATHING

A 2D view of phoneme vectors is shown in Figure 1. We observed phonemes like AX, IH, IY, EY, EH, OW, IY, AE are close to each other and phonemes like AX, DX or EH, TS are well separated which indicates that phoneme embedding is able to capture pronunciation related information.

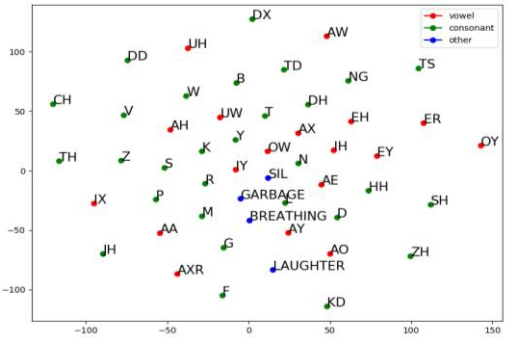
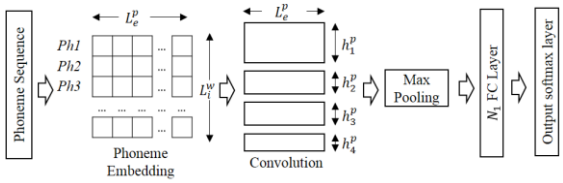


Figure 1: *t-SNE visualization of phoneme vectors*

As a part of experiment, we also generate phoneme embedding using text data. The text data is converted into phoneme sequence using G2P tool [20] and the same word2vec is used to generate the embedding. The advantage of text based phoneme embedding is availability of large volume of data. We used google's 1 billion text dataset [21] for phoneme embedding. Though the phoneme embedding from text is more stable, keeping the problem statement (speech based emotion classification) in mind we used speech based phoneme embedding in this work.

The architecture of phoneme based CNN is shown in Figure 2. The model takes phoneme sequence (in the form of embedded vector) as input, followed by convolution with multiple kernels. The maximum phoneme sequence length is fixed at 512 which covers almost 75% of the max length of the dataset. The max-pool layer takes one feature out of the features generated after convolution. The extracted features are flattened and fed to multiple fully connected (FC) layer one after the other. Finally, a softmax layer is used to perform classification. We used different training techniques like dropout (.25 to .75), batch-normalization that helps in reducing overfitting, sensitivity towards initial starting weights. We also observed improvement in convergence rate with the use of batch-normalization.



$L_e^p = Phoneme\ embedding\ length,$
$L_i^p = Input\ sequence\ length, h_i^p = filter\ height,\ i = 1,2,3,4$

Figure 2: *Proposed CNN model*

The model uses 4 convolutional kernel of dimension $h_i^p \times L_e^p$ (3,9,13,17 × 100). A kernel of size 9 × 100 indicates the filter span over a length of 9 phonemes. These filters are applied in parallel and the coefficients are tuned as a part of training process. Number of filter of each type is 200 and max-pooling is used to extract the feature for FC layer.

2.2. Model 2: CNN for spectrogram features

In this model we use spectrogram as input to the 2D CNN. Spectrogram is generated by STFT (Short Term Fourier Transform) of windowed audio or speech signal. We sampled the audio at 22050Hz sampling rate. Each frame of audio is windowed using *"hann"* window of length 2048. We applied 2048 length FFT windows on the windowed audio samples and used 512 as the hop length for the Short-Time Fourier transform (STFT). The computed magnitude spectrogram is mapped to the mel-scale to get mel-spectrogram. Mel-frequency scale emphasizes the low frequency over the high frequency, similar to the human ears perceptual capability. We used the "librosa" python package to compute the mel-spectrogram using above mentioned parameters. A sample spectrogram corresponding to audio "*Yeah...you wanna see my supervisor ah you wanna see my supervisor. fine I'll right back*", is shown in Figure 3. Spectrogram features are used instead of MFCC (Mel Frequency Cepstral Coefficients) features as Discrete Cosine Transformation (DCT) for generating MFCC destroys locality information.

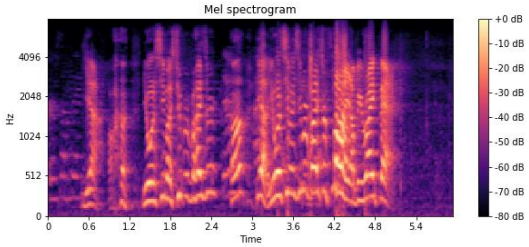


Figure 3: *Sample spectrogram extracted from speech*

We trimmed the long duration audio utterances to a duration which covers 75 percentile of all audio data samples of the dataset, under the assumption that the frequency variations responsible to capture the emotion content, will be present throughout the dialogue and hence will not be lost by trimming the end of long dialogues. Hence, the maximum duration considered is 6 seconds.

The 2D CNN architecture for working with spectrogram is presented in Figure 4. In this model, a set of 4 parallel 2D convolutions is performed to extract features from spectrogram. The input shape of the spectrogram Image is 128×256 (n-mels x no.of windows). For each parallel convolution, a set of 200 2D kernels were used. The kernel size is fixed for each parallel path. The size of different kernels at different parallel paths are $12 \times 16, 18 \times 24, 24 \times 32, 30 \times 40$. The features generated in the convolution layers are fed to max-pool layer to extract 4 features for each filter, i.e. the pool size is half along width and height of convolution output. The extracted features are flattened and fed to fully connected (FC) layer. Two FC layers are used in this work of sizes 400 and 200. In fully connected layers we performed batch normalization as well as different dropout rates like 50% and 75% for first FC and no dropout for the 2nd FC. "ReLU" activation function is used in Convolution and FC layers. Finally, an output softmax layer is used to perform the four class classification. "Adadelta" is used for training Optimization.

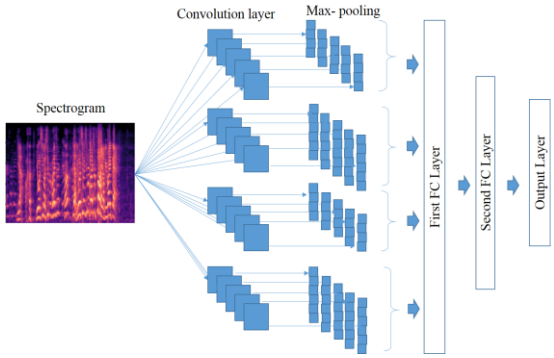


Figure 4: *Proposed CNN model for spectrogram*

Spectrogram is a time-frequency representation of speech. 2D convolution filters captures 2-dimensional feature from the spectrogram. Such 2D features are not available when converted into phoneme or text. So, spectrogram with such rich features may be better for emotion recognition.

We performed experiment with another variant of this model (referred a Model-2A) in this paper. This is similar to Model 2, but it has four additional parallel convolution layers which take down-sampled (by 2 in both time and frequency) spectrogram as input. The convolutional filters remain same as Model 2. The Objective of this experiment is to capture higher level features as compared to Model 2. The performance of this model is discussed in Table 2.

2.3. Model 3: Multi-channel CNN model with phoneme and spectrogram features

In this model (see Figure 5), we combine phoneme and spectrogram both to achieve a better emotion classification accuracy. Since, both the inputs are of different dimensions, we use separate CNN channel for both. The phoneme channel consists of phoneme embedding, followed by four parallel convolution layers of different convolutional kernels which take phoneme sequence as input. The Spectrogram channel also has four parallel 2d-convolution layers with different filter shapes which takes Spectrogram image as input as in Model 2. Both the phoneme and spectrogram channel convolution outputs are fed to two separate FC layers and the output of each FC layer is concatenated and fed to the second level FC layer after normalization. The hyper-parameters and network

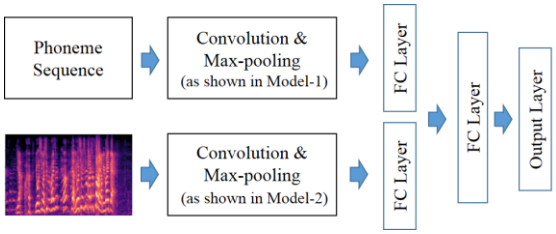


Figure 5: *Proposed Multi-channel CNN model for spectrogram and phoneme*

structure for individual channel are similar to their respective models as presented in Figure 2 and Figure 4.

2.4. Datasets

We used the University of Southern California's Interactive Emotional Motion Capture (USC-IEMOCAP) database in this work. The IEMOCAP corpus comprises of five sessions where each session includes the conversation between two people and its corresponding labeled speech text (both phoneme and word level). Each session consists of both male and female voices to remove any gender bias. The IEMOCAP corpus comprises of scripted and improvised dialogs. We are only using improvised data because scripted text shows strong correlation with labeled emotions and can lead to lingual content learning, which can be an undesired side effect. The final experimental dataset extracted from the original IEMOCAP data comprised of 4 classes named Neutral (48.8% of the total dataset), Happiness (12.3%), Sadness (26.9%) and Anger (12%). These emotions are selected on the basis of previous work and the state of art results which we try to enhance. As there is data imbalance between classes we presented the overall accuracy as well as average of the entire four class accuracy and the confusion matrix.

3. Evaluation and Discussions

3.1. Emotion classification on standard dataset

To show the effectiveness of the proposed method for emotion classification, we compared our methods with the state-of-the-art benchmark results on IEMOCAP dataset. Multiple attempts have been made before by multiple researchers on enhancing the classification accuracy. Some recent results are presented in Table 2 along with our 5-fold cross-validation experimental results. Both overall and class accuracies are presented for better comparison, where overall accuracy is measured based on total counts irrespective of classes and class accuracy is the mean of accuracies achieved in each class.

Table 2: Comparison of accuracies:

Methods	Input	Overall Accuracy	Class Accuracy
Lee [3]	Spectrogram	62.8	63.9
Satt [18]	Spectrogram	68.8	59.4
Model-1	Phoneme	59.1	46.7
Model-2	Spectrogram	**71.2**	61.9
Model-2A	Spectrogram	**71.3**	61.6
Model-3	Phoneme & Spectrogram	**73.9**	**68.5**

It is observed that the combined spectrogram and phoneme based CNN model is able to achieve a 4% higher average class accuracy compared to the existing state of art methods. Only Spectrogram models have performed equally well compared to existing models, but not as good as the combined model. As mentioned before, the data is not well balanced. So, in Table 3 we present the confusion matrix showing misclassification between each pair of classes for Model-3. From the table we can observe that Neutral and Sadness classes have shown high true positive whereas Happiness and Anger classes are more misclassification as neutral emotions.

Table 3: Confusion Matrix in Percentage on the Model-3:

Class Labels	**Prediction**			
	Neutral	Happiness	Sadness	Anger
Neutral	**75.5**	7	14.3	3.2
Happiness	26	**59.2**	11.1	3.7
Sadness	14	1.7	**83.5**	0.8
Anger	35.6	5.1	3.4	**55.9**

3.2. Discussion

Only-phoneme based model (Model-1) failed to exceed the accuracy of existing state-of-the-art methods. Spectrogram based approaches perform better as they contain both time and frequency information as input to the model. The overall accuracy of the only spectrogram model exceeds the existing best accuracy by more than 2.5%. The Model 2A, which is another variant of Spectrogram Model has performed similar to the Model 2 but it converged faster. The combined phoneme and spectrogram model enhances the accuracy further which is about 4% higher than the existing state of art result. As mentioned before, spectrogram contains time-frequency information and when convolved with 2d kernels, captures 2-dimensional features which are much richer than phoneme based features. In some scenarios phoneme may work better than spectrogram. For example, a set of discrete spectral bands along time axis forms a phoneme. When the phoneme is generated, the discrete nature of spectrogram is abolished in phoneme. Moreover, phoneme can differentiate between voiced and unvoiced part of the speech as well as the intervals between sub-words and words or phrases. And also phoneme embedding being trained on Word2Vec captures the semantic context too, which helps in identifying the emotion to some extent. For spectrogram based approach, both voiced part, unvoiced part and intervals behave in similar way. Thus, when we combine both phoneme and spectrogram, we are able to learn different kinds of features from them. This might be a probable reason of enhanced accuracy in combined phoneme and spectrogram model.

4. Conclusions

Multiple architectures have been proposed to work with phoneme and spectrogram on emotion recognition. Spectrogram based 2D CNN model provided enhanced accuracy which is further enhanced when combined with phoneme. The combined method provided more than 4% increase in overall accuracy as well as average class accuracy over existing state-of-the-art method. The proposed method can be used for similar other applications such as conversational Chatbot where identifying the emotion and sentiment hidden in the speech may play a role in better conversation. We use phoneme embedding which is generated from IEMOCAP speech data. This phoneme embedding can be used as pre-trained embedding for similar other analysis.

5. Acknowledgements

The authors would like to acknowledge the support of Samsung R&D Institute-India, Bangalore in this work.

6. References

[1] X. Zhang, J. Zhao, & Y. LeCun, "Character-level Convolutional Networks for Text Classification," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 3057–3061, 2015.

[2] Y. Kim, "Convolutional Neural Networks for Sentence Classification," *Proceedings of the 2014 Conference on EMNLP*, pp. 1746–*1751, 2014.*

[3] J. Lee and I Tashev. High-level feature representation using recurrent neural network for speech emotion recognition. In *INTERSPEECH* , 2015.

[4] Q. Jin, C. Li, S. Chen, H. Wu . "Speech emotion recognition with acoustic and lexical features."*in IEEE International Conference on Acoustics, Speech and Signal Processing* 2015:4749-4753.

[5] V. Dimitrios, and C. Kotropoulos. "Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition." *Signal Processing* 88.12, pp. 2956-2970, 2008.

[6] X. Mao, L. Chen, and L. Fu. "Multi-level Speech Emotion Recognition Based on HMM and ANN." *WRI World Congress on Computer Science and Information Engineering*, 225-229, 2009.

[7] S. Ntalampiras and N. Fakotakis. "Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition." *IEEE Transactions on Affective Computing* 3.99, pp. 116-125, 2012.

[8] H. Hao, M. X. Xu, and W. Wu. "GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition." *IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP*, pp. 413-416, 2007.

[9] D. Neiberg, K. Laskowski, and K. Elenius. "Emotion Recognition in Spontaneous Speech Using GMMs." *INTERSPEECH, 2006.*

[10] C. H. Wu, and W. B. Liang. "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels." *IEEE Transactions on Affective Computing* 2.1, pp. 10-21, 2011.

[11] Y. Kim, H. Lee, and E. M. Provost. "Deep learning for robust feature generation in audiovisual emotion recognition." *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3687-3691, 2013.

[12] W. L. Zheng, J. Zhu, Y. Peng. "EEG-based emotion classification using deep belief networks." *IEEE International Conference on Multimedia & Expo*, pp. 1-6, 2014.

[13] K. Han, D. Yu, and I. Tashev. Speech emotion recognition using deep neural network and extreme learning machine. *In INTERSPEECH*, 2014.

[14] W. Q. Zheng, J. S. Yu, and Y. X. Zou. "An experimental study of speech emotion recognition based on deep convolutional neural networks." *International Conference on Affective Computing and Intelligent Interaction*, pp. 827-831, 2015.

[15] H. M. Fayek, M. Lech, and L. Cavedon. "Towards real-time Speech Emotion Recognition using deep neural networks." *International Conference on Signal Processing and Communication Systems*, pp.1-5, 2015.

[16] V. Chernykh, G. Sterling. "Emotion Recognition From Speech With Recurrent Neural Networks" *arXiv:1701.08071v1*, 2017.

[17] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335-359, 2008.

[18] A. Satt, S. Rozenberg, R. Hoory. "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms" in *INTERSPEECH*, Stockholm, 2017.

[19] T. Mikolov, K. Chen, G. Corrado, & J. Dean, "Efficient Estimation of Word Representations in Vector Space," *In Proceedings of Workshop at ICLR*, 2013.

[20] M. Bisani and H. Ney. "Joint-Sequence Models for Grapheme-to-Phoneme Conversion". *Speech Communication*, Volume 50, Issue 5, pp. 434-451, 2008.
[21] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, and P. Koehn, “One billion word benchmark for measuring progress in statistical language modeling,” CoRR, abs/1312.3005, 2013.