EMOCEPTION: AN INCEPTION INSPIRED EFFICIENT SPEECH EMOTION RECOGNITION NETWORK

Chirag Singh, Abhay Kumar, Ajay Nagar*, Suraj Tripathi*, Promod Yenigalla*

[1]Samsung R&D Institute India - Bangalore

c.singh@samsung.com, abhay1.kumar@samsung.com, ajay.nagar@samsung.com
surajtripathi93@gmail.com, promod.y@samsung.com

ABSTRACT

This research proposes a Deep Neural Network architecture for Speech Emotion Recognition called Emoception, which takes inspiration from Inception modules. The network takes speech features like Mel-Frequency Spectral Coefficients (MFSC) or Mel-Frequency Cepstral Coefficients (MFCC) as input and recognizes the relevant emotion in the speech. We use USC-IEMOCAP dataset for training but the limited amount of training data and large depth of the network makes the network prone to overfitting, reducing validation accuracy. The Emoception network overcomes this problem by extending in width without increase in computational cost. We also employ a powerful regularization technique, Multi-Task Learning (MTL) to make the network robust. The model using MFSC input with MTL increases the accuracy by 1.6% vis-à-vis Emoception without MTL. We report an overall accuracy improvement of around 4.6% compared to the existing state-of-art methods for four emotion classes on IEMOCAP dataset.

Index Terms— Speech Emotion Recognition, Inception, Multi-Task Learning, CNN

1. INTRODUCTION

Humans possess prime quality to sense emotions through several modes of communication like facial expression, text, speech etc. Despite significant progress in artificial intelligence, machines still strive to understand emotion from these communication modes especially speech [1] as every individual express emotions with different paralinguistic characteristics such as fundamental frequency, energy, timing and intensity. Recently, Speech Emotion Recognition (SER) has become an interesting field of research. SER aims to identify different types of emotions from speech signal.

Standard SER systems first extract the features from speech signal and then use a classification algorithm to identify the emotion. Numerous studies have been done on feature extraction from speech signals. Generally, MFCC, MFSC and phonemes of speech signal are used as features for voice research. Eyben et al. [2] suggested the Geneva minimalistic acoustic parameter set for voice research, Wu et al. [3] compared speech duration, energy, pitch and MFCC acoustic features, Schmitt et al. [4] proposed Bag-of-Audio-Words of MFCCs feature set to model inherent structure present in speech signal. For efficient SER, various types of supervised machine learning algorithms such as Support Vector Machines and Decision Tree have been exploited. Kim et al. [5] applied unsupervised machine learning algorithm K-Nearest Neighbor (KNN), Schuller et al. proposed Hidden Markov Model (HMM) [6] for SER. In the last few years, deep learning models gave amazing results on speech processing. Han et al. [1] proposed a SER architecture using deep neural network and extreme learning machine. Jin et al. [7] achieved good results on IEMOCAP dataset by early and late fusion of acoustic and lexical features. Later, Lim et al. [8], proposed an architecture using convolution neural networks (CNN) and recurrent neural networks (RNN) with accuracy improvements. Niu et al. [9] trained a very deep convolution neural network with data augmentation algorithm to increase amount of training data and to acquire the different size of spectrogram. Lee et al. [10] used RNN with low-level acoustic features to achieve remarkable results on SER. Satt et al. [11] proposed an architecture using both CNN and LSTM. Yenigalla et al. [12] employed phoneme embedding along with spectrogram in multi-channel CNN to get higher emotion recognition accuracy.

In this paper, we focused on computationally cheap and robust neural network architecture for SER, drawing inspiration from Inception Network [13] predominantly used in computer vision. The proposed neural network is referred to as "Emoception" hereinafter. Emoception Network is validated for different inputs mentioned below:

- Mel-Frequency Cepstral Coefficients (MFCC)
- Mel-Frequency Spectral Coefficients (MFSC)

Further, Emoception is evaluated for the effectiveness of the MTL [14] by performing SER using Single-Task Learning.

*Equal Contribution

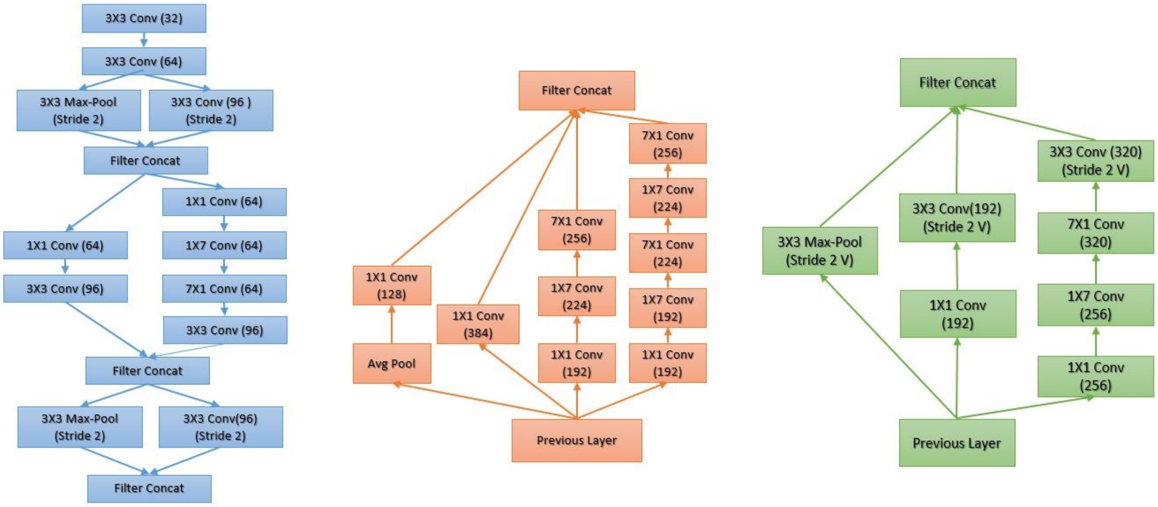

Figure 1: *Architecture of Stem Block (left), Inception Module-B (middle) and Reduction Module-B (right), where number of kernels are shown in parentheses and "V" represents 'valid padding'.*

2. DATASET

We conduct extensive evaluations on University of Southern California's Interactive Emotional Motion Capture (USC-IEMOCAP) [15] dataset to measure the effectiveness of the proposed technique. The entire database is divided into 5 sections, each have both male and female voices. From the improvised speech section and scripted speech section, we only used improvised speech section. In our work, we have focused primarily on Happiness (12.3% of the complete dataset), Sadness (26.9%), Anger (12%) and Neutral (48.8%) state from the original IEMOCAP dataset, as these categories are the most common emotion descriptors and were already used for speech emotion recognition research. We considered Weighted Accuracy (WA) as the classification accuracy of all classes and Unweighted Accuracy (UA) as the average of accuracies of individual classes for evaluation.

3. PROPOSED METHODS

In this section, we first introduce the Audio feature extraction method to generate input for Emoception Network. Then we present the fundamental blocks of Emoception followed by complete architecture of Emoception network.

3.1. Audio Feature Extraction

In our experiments, we used MFSC and MFCC as audio features. IEMOCAP utterances vary between 3-15 seconds. We trim long duration audio files to a duration (6 seconds) which covers 75 percentile of all audio data samples of the dataset. It is assumed that the frequency variations, which could possibly characterize the emotionality of the speech data will be present throughout the dialogue, hence will not be lost by this trimming of audio length.

3.1.1. Mel-Frequency Spectral Coefficients (MFSC)

Mel-frequency spectrogram is a two-dimensional representation of log-magnitude intensity (dB) over frequency and time. The audio signal is sampled at 22050Hz, and windowed using a "hann" window of length 2048. Short Term Fourier Transform (STFT) of windowed audio samples generates spectrograms. We apply Fast Fourier Transform (FFT) on the windowed audio samples, with windows of length 2048 and hop-length equal to 512. Spectrogram magnitudes are obtained from FFT, which are then converted to the Mel-scale to get Mel-frequency spectrum. We use 128 Spectrogram coefficients per window in our experiments. The Mel-frequency scale imitates the human perceptual hearing capabilities by emphasizing small changes at low frequencies over the high frequencies. We used the "librosa" python package, along with the above-mentioned parameters, to compute the Mel-frequency spectrograms.

3.1.2. Mel-Frequency Cepstral Coefficients (MFSC)

Mel Frequency Cepstrum (MFC) is a two-dimensional representation of the Short-Term Power Spectrum of sound. It is based on a linear cosine transform of a log power spectrum on a non-linear Mel-scale of frequency. We use "librosa" python package for MFCC generation as well. The hyperparameters employed for MFCC generation are similar to the ones described for

Spectrogram generation. The only difference is that 40 MFCCs per window are generated compared to the earlier mentioned 128 Spectrogram coefficients per window.

3.2. Emoception Blocks

Emoception Network consists of three fundamental blocks: Stem, Simplified Inception Modules and Multi-Task Learning regularizer. In the following sections, the motivations and usage of each block are discussed.

3.2.1. Stem Block

Stem Block is Emoception's first block which is fed with audio features as input. It is designed to provide low-level features for the successive Simplified Inception Modules without any significant loss of input information. We do this by maintaining the Stem Block depth low. This offers an additional advantage of reduced over-fitting due to less trainable parameters.

Figure-1 shows detailed architecture of Emoception Stem Block. The convolution layers with different kernel sizes have been stacked up as shown in the figure. Convolutions done in parallel to Max pool layer have a stride of two, making their outputs of same shape. Remaining convolutions in the Stem Block use a stride of one. The output shape obtained from Stem Block is one-fourth of the input shape. This output is then fed to Simplified Inception Modules (discussed in section 3.2.2).

3.2.2. Simplified Inception Modules

Inception Modules are novel in the way that they use Network-in-Network approach introduced by Lin et al. [16] to enhance the representational power of the network. These Modules use convolution with different kernel sizes in parallel, along with Max pool layer to capture multi-scale features of the input. Padding for these convolutions is kept "same" in order to concatenate parallel outputs together. This technique eliminates the need to select appropriate kernel sizes for the input and propagates only those features forward which help in better discrimination of the input. Unlike the Inception Network [17], we use Inception Modules without repetition. This simplification reduces computations and increases efficiency of the network. The different convolutions in the Modules are preceded by 1×1 convolutions that serve as bottleneck layer reducing the computational cost significantly and introducing non-linearity in the network. Reduction Modules A and B [17] are used to downsample the features provided by Inception blocks.

Emoception Network uses three Inception Modules, namely Inception-A, Inception-B and Inception-C [17]. These Inception Modules A, B and C are stacked with intermediate Reduction Modules A and B as shown in Figure-2. The detailed architecture of Inception Module-B and Reduction Module-B is shown in Figure-1(middle) and Figure-1(right) respectively.

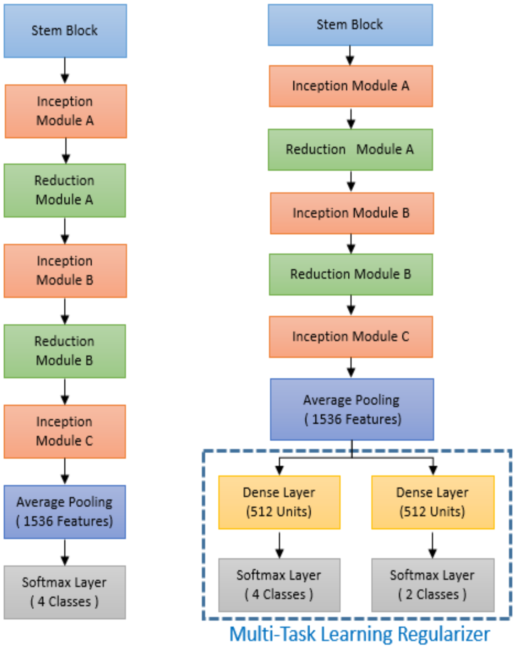

Figure 2: *Architecture of proposed Model-1 (left) and Model-2 (right).*

3.2.3. Multi-Task Learning Regularizer

Emoception exploits representational power of the wide network due to its ability to handle multiple scales efficiently. However, the advantages of Inception Modules are rendered moot with a large number of trainable parameters that make network prone to overfitting and a uniform increase in computational cost. The problem of overfitting gets worse if the training data corresponding to each label is limited and unbalanced. This requires the implementation of a powerful regularization technique to train the network. Hence, Multi-Task Learning is employed as a regularizer to improve generalization ability of the network by learning a shared representation of multiple related tasks in feature layers. The two independent classifiers used in MTL are hereinafter referred to as Main task and Auxiliary task. The auxiliary task introduces an inductive bias [18,19] in hypotheses space, which accommodates novel future input samples and reduces over-fitting. The auxiliary task chosen in MTL should be related to the main task to have an added benefit of implicit data augmentation. These tasks, trained jointly in a single model, also reduce the computation cost compared with individually trained tasks.

3.3. Model 1: Emoception without MTL

Model-1 architecture is shown in Figure-2. It consists of the Stem Block and three Inception Modules A, B and C [17] stacked serially. Reduction Module A and B follow both Inception Module A and B respectively. The model uses Single-Task Learning to classify emotions into four categories of Neutral, Happiness, Sadness and Anger. This is regarded as "Main Task" of the Emoception Network. The average-pool layer takes output from Inception Module-C to generate 1536 features. These extracted features are flattened and fed to softmax layer to classify into four emotion classes.

3.4. Model 2: Emoception

MTL learns the common and rich feature representation between different tasks successfully and leads to generalization of the main task. Model-2 is similar to Model-1 except the introduction of auxiliary task. The auxiliary task for proposed model is to classify the input into two categories namely "emotion" and "non-emotion". The three classes Happiness, Sadness and Anger are aggregated to form "emotion" category while the Neutral class is denoted as "non-emotion".

As main task and auxiliary task chosen are closely related, features from Inception-C module can be used by both the tasks. The average-pool layer extracts these features, which are flattened and fed to two fully connected layers of size 512 in parallel. The parallel dense layers are then fed to parallel soft-max layers corresponding to main task and auxiliary task. Total loss for Emoception is calculated as the weighted sum of loss from main task and auxiliary task.

$$L_{total} = W_1 L_1 + W_2 L_2 \tag{1}$$

Where L_1 and L_2 correspond to cross entropy loss incurred in main task and auxiliary task respectively. W_1 and W_2 corresponds to the weight given to the main task and auxiliary task respectively. We experimentally tune the weights W_1 and W_2 with 1.0 and 0.5 respectively.

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1. Experimental Setting

We investigated Model-1 and Model-2 with MFCC and MFSC as input for speech emotion recognition. MFCC and MFSC inputs are of shapes 40×256 and 128×256 respectively. We use dropout of 0.5 and 0.1 in Average Pooling layer and Dense Layers in both Models. Batch-normalization is applied to reduce sensitivity towards LeCun [20] normal initialized convolution kernel weights. We use "AdaDelta" optimizer with initial learning rate of 1.0 and decay factor of 0.95 to train our models.

4.2. Results

We have aggregated our experimental results in Table-1 along with state-of-the-art benchmark results on IEMOCAP dataset. Emotion classification accuracy is obtained from 5-fold cross validation results and class accuracy is evaluated as the mean of the accuracies obtained in individual class.

For Model-1, the overall accuracy of the MFCC fed network is 73.5%, while the accuracy improved to 74.3% with the MFSC input. Model-2 employing MTL improves the overall accuracy to 75.3% and 75.9% for MFCC and MFSC respectively. The accuracy is 4.6% more than the state-of-the-art accuracy of 71.3%.

Table 1: *Accuracy Comparison*

Model	**Input**	**Overall Accuracy**	**Class Accuracy**
Lee [10]	MFSC	62.8	63.9
Satt [11]	MFSC	68.8	59.4
Promod [12]	MFSC	71.3	61.6
Model-1	MFSC	**74.3**	62.7
Model-1	MFCC	**73.5**	62.1
Model-2	MFSC	**75.9**	**68.0**
Model-2	MFCC	**75.3**	**69.1**

4.3. Discussion

The proposed Emoception network is efficient due to the reduced number of computations achieved by bottleneck layers in Inception Modules and jointly training two related tasks using MTL. The network performs better as it effectively extracts multi-scale features with the Inception modules. The training process is regularized by MTL to avoid overfitting in Model-2.

Table-2 shows confusion matrix evaluated on Model-1 (MFCC), with 73.5% as overall accuracy and 62.1% class accuracy.

Table 2: *Confusion Matrix in Percentage on Model-1.*

Class Labels	**Neutral**	**Happiness**	**Sadness**	**Anger**
Neutral	96.74	1.86	0.93	0.46
Happiness	68.75	29.17	0.0	2.08
Sadness	41.18	2.94	55.88	0.0
Anger	31.37	1.96	0.0	66.67

We evaluate Model-2 (MFCC) with an overall accuracy comparable to Model-1 (MFCC) for a fair comparison of class accuracy. The auxiliary task accuracy for Model-2 (MFCC) is 75.5%. Confusion matrix for main task is represented in Table-3.

Table 3: *Confusion Matrix in Percentage on Model-2.*

Class Labels	Neutral	Happiness	Sadness	Anger
Neutral	78.60	4.18	3.25	13.95
Happiness	56.25	33.33	2.08	8.33
Sadness	24.26	5.14	68.38	2.21
Anger	15.69	1.96	1.96	80.39

The confusion matrix obtained for Model-1 shows that the class accuracy for three emotion classes (Happiness, Sadness and Anger) are hampered the most by Neutral emotion. The inductive bias introduced in the hypotheses space by auxiliary task lowers this “non-emotion” class interference in Model-2. Thus, Table-3 shows decrease in Neutral class percentage present in each “emotion” class. Model-2 (MFCC) class accuracy improved by 7% than Model-1(MFCC) as shown in Table-1. Class accuracy on Model-2 (MFSC) increased by 5.3% than Model-1(MFSC) further validating the use of MTL as a powerful measure to improve generalization.

5. CONCLUSIONS

In this paper, Emoception is demonstrated to be efficient for speech emotion recognition by virtue of bottleneck layers in Inception Modules. The extended width of the network extracts multi-scale features from input which helps to better recognize speech emotion. The robustness of the network is further improved by introducing auxiliary task, related to Main task, assisting in positive transfer of learning. Multi-Task Learning used as a regularization technique creates implicit data augmentation which overcomes the imbalance and limited availability of the labeled dataset. Emoception increases overall accuracy to 75.9%, which is 4.6% better than existing state-of-the-art accuracy. The class accuracy is also improved by around 6% than state-of-the-art class accuracy.

6. ACKNOWLEDGEMENT

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

7. REFERENCES

[1] K. Han, D. Yu, and I. Tashev. "Speech emotion recognition using deep neural network and extreme learning machine." In *Fifteenth annual conference of the international speech communication association*. 2014.

[2] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers et al. "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing." *IEEE Transactions on Affective Computing* 7, no. 2 (2016): 190-202.

[3] D. Wu, T. D. Parsons, and S. S. Narayanan. "Acoustic feature analysis in speech emotion primitives estimation." In *Eleventh Annual Conference of the International Speech Communication Association*. 2010.

[4] M. Schmitt, F. Ringeval, and B. W. Schuller. "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech." In INTERSPEECH, pp. 495-499. 2016.

[5] Y. Kim, and E. M. Provost. "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions." In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3677-3681. IEEE, 2013.

[6] B. Schuller, G. Rigoll, and M. Lang. "Hidden Markov model-based speech emotion recognition." In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 2, pp. II-1. IEEE, 2003.

[7] Q. Jin, C. Li, S. Chen, and H. Wu. "Speech emotion recognition with acoustic and lexical features." In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4749-4753. IEEE, 2015.

[8] W. Lim, D. Jang, and T. Lee. "Speech emotion recognition using convolutional and recurrent neural networks." In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1-4. IEEE, 2016.

[9] Y. Niu, D. Zou, Y. Niu, Z. He, and H. Tan. "A breakthrough in speech emotion recognition using deep retinal convolution neural networks." *arXiv preprint arXiv:1707.09917* (2017).

[10] J. Lee, and I. Tashev. "High-level feature representation using recurrent neural network for speech emotion recognition." in *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.

[11] A. Satt, S. Rozenberg, and R. Hoory. "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms." In INTERSPEECH, pp. 1089-1093. 2017.

[12] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa. "Speech Emotion Recognition Using Spectrogram & Phoneme Embedding." Proc. INTERSPEECH 2018 (2018): 3688-3692.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going deeper with convolutions." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1-9. 2015.

[14] R. Caruana, "Multitask learning." Machine learning 28, no. 1 (1997): 41-75.

[15] 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.

[16] M. Lin, Q. Chen, and S. Yan. "Network in network." *arXiv preprint arXiv:1312.4400* (2013).

[17] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. "Inception-v4, inception-resnet and the impact of residual connections on learning." In *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[18] J. Baxter, "A model of inductive bias learning." *Journal of artificial intelligence research* 12 (2000): 149-198.

[19] A. Kumar, and H. Daume III. "Learning task grouping and overlap in multi-task learning." *arXiv preprint arXiv:1206.6417* (2012).

[20] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. "Self-normalizing neural networks." In *Advances in neural information processing systems*, pp. 971-980. 2017.