

# **Focal Loss based Residual Convolutional Neural Network for Speech Emotion Recognition**

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

**Paper# 308**

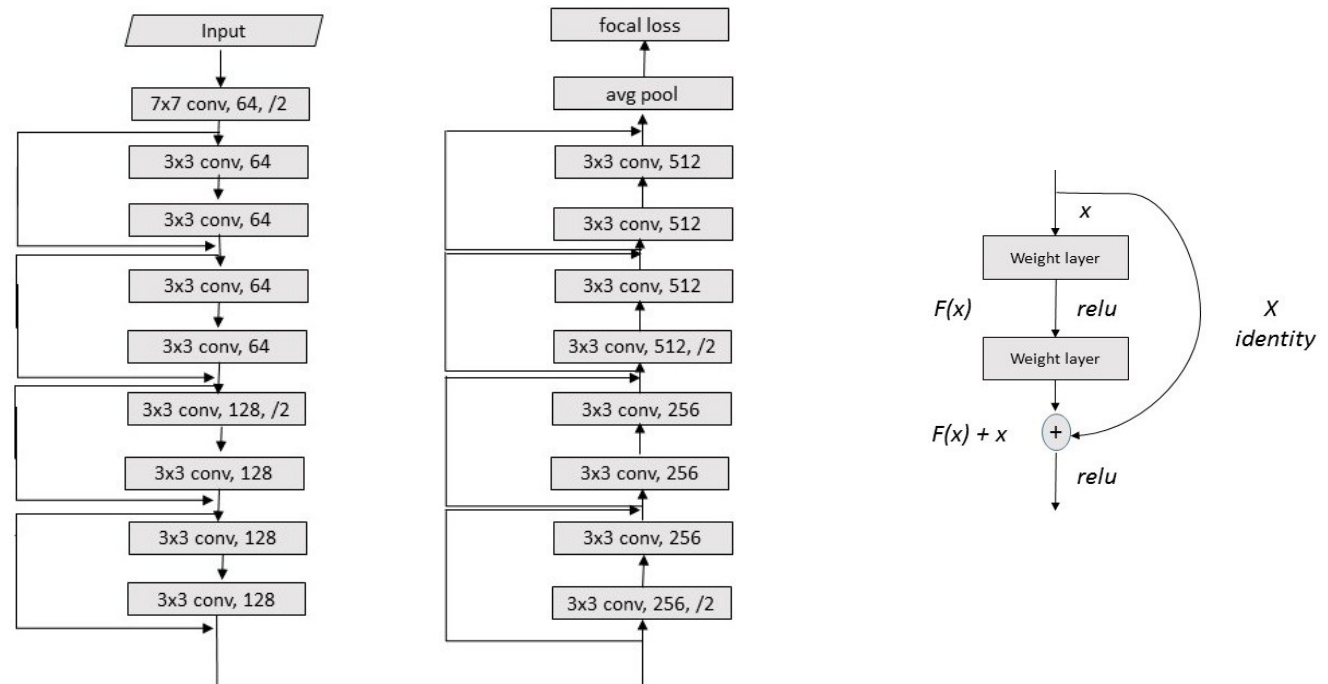
**Presented by  
Abhay Kumar**

# Introduction

- This paper proposes a Residual Convolutional Neural Network (ResNet) based on speech features and trained under Focal Loss to recognize emotion in speech.
- Speech features such as Spectrogram and Mel-frequency Cepstral Coefficients (MFCCs) have been used to characterize emotion better than just plain text.
- Focal Loss has the ability to focus the training process more towards hard-examples and down-weight the loss assigned to well-classified examples, thus preventing the model from being overwhelmed by easily classifiable examples.

$$FL(p_t) = -(1 - p_t)^\gamma \log(p_t)$$

# Proposed Approach



**Fig. 1** Proposed model architecture

- Our best model achieved a 3.4% improvement in overall accuracy and a 2.8% improvement in class accuracy when compared to existing state-of-the-art methods on IEMOCAP dataset.

**Table 1.** Comparison of accuracies

Methods	Input	Overall Accuracy	Class Accuracy
Lee et al. [1]	Spectrogram	62.8	63.9
Satt et al. [2]	Spectrogram	68.8	59.4
Yenigalla [3]	Spectrogram	71.2	61.9
Proposed Model	Spectrogram	74.2	64.3
Proposed Model	MFCC	<b>74.6</b>	<b>66.7</b>

**Table 2.** Ablation study of the effectiveness of Focal Loss

Input Features	Loss functions settings	Overall Accuracy	Class Accuracy
Spectrogram	Softmax Loss	70.2	55.8
Spectrogram	Focal Loss	74.2	64.3
MFCC	Softmax Loss	70.7	56.9
MFCC	Focal Loss	74.6	66.7


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. Yenigalla, P., Kumar, A., Tripathi, S., Kar, S., Vepa, J.: Speech Emotion Recognition using Spectrogram & Phoneme Embedding, In: Interspeech, 2018.

# Poster Screenshot

Title & Authors  
 Introduction  
 Proposed Approach  
 Results  
 Poster Screenshot

## Focal Loss based Residual Convolutional Neural Network for Speech Emotion Recognition

Suraj Tripathi, Abhay Kumar, Abhiram Ramesh,  
 Chirag Singh, Promod Yenigalla  
 Samsung R&D Institute India – Bangalore



**SAMSUNG**

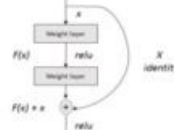
### Introduction

This paper proposes a Residual Convolutional Neural Network (ResNet) based on speech features and trained under Focal Loss to recognize emotion in speech.

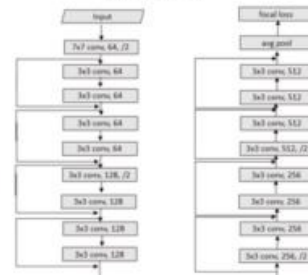
**Proposed Solution:**

- Speech features such as Spectrogram and Mel-frequency Cepstral Coefficients (MFCCs) have been used to characterize emotions better than just plain text.
- Focal Loss has the ability to focus the training process more towards hard-examples and down-weight the loss assigned to well-classified examples, thus preventing the model from being overwhelmed by easily classifiable examples.
- Our best model achieved a 3.4% improvement in overall accuracy and a 2.8% improvement in class accuracy when compared to existing state-of-the-art methods on IEMOCAP dataset.

### Methods



**Fig. 1. Residual block**

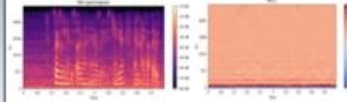


**Fig. 2. Proposed model architecture**

Focal loss is formulated as:

$$F(\mathcal{X}) = -\sum_i (1 - p_i)^{\gamma} \log(p_i)$$

### Speech Features



### Results

Methods	Input	Overall Accuracy	Class Accuracy
Lee et al. [1]	Spectrogram	62.8	63.9
Sain et al. [2]	Spectrogram	68.8	59.4
Yenigalla [3]	Spectrogram	71.2	61.9
Proposed Model	Spectrogram	74.2	64.3
Proposed Model	MFCC	74.6	66.7

**Table 1. Comparison of accuracies**

Input Features	Loss function settings	Overall Accuracy	Class Accuracy
Spectrogram	Softmax Loss	70.2	55.8
Spectrogram	Focal Loss	74.2	64.3
MFCC	Softmax Loss	70.7	56.9
MFCC	Focal Loss	74.6	66.7

**Table 2. Ablation study of the effectiveness of Focal Loss**

### Conclusions

- The use of Spectrogram and MFCC provides low-level features, which when combined with ResNets has allowed us to extract very deep features boosting the model performance.
- Our best model (MFCC) outperforms the benchmark results by 3.4% and 2.8% for overall and class accuracies respectively.
- With the help of Focal Loss, we have significantly improved recognitions of the rare emotion classes (Anger, Sadness and Happiness) as shown in our confusion matrices.
- Focal Loss helps to scale the standard cross-entropy loss to down-weight loss corresponding to easily classifiable examples dynamically and focus more on hard examples to make the system perform better on hard examples as well.

### References

- Lee, J., Taylor, J. High-level feature representation using recurrent neural network for speech emotion recognition. In: INTERSPEECH (2011)
- Sain, A., Sreenivasy, S., Murty, S. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In: INTERSPEECH, Stockholm (2017)
- Yenigalla, P., Kumar, A., Tripathi, S., Rao, S., Naga, J. Speech Emotion Recognition using Spectrogram & Dynamic Embedding. In: INTERSPEECH, 2018

**SAMSUNG Research**