A Segment Level Approach to Speech Emotion Recognition using Transfer Learning

Sourav Sahoo[1][0000−0002−1956−2834], Puneet Kumar[2][0000−0002−4318−1353],
Balasubramanian Raman[2][0000−0001−6277−6267] and
Partha Pratim Roy[2][0000−0002−5735−5254]

[1] Dept. of Electrical Engineering, Indian Institute of Technology Madras, Chennai - 600036, India
ee17b040@smail.iitm.ac.in
[2] Dept. of Computer Science and Engineering, Indian Institute of Technology Roorkee, Roorkee - 247667, India
pkumar99@cs.iitr.ac.in,{balarfcs,proy.fcs}@iitr.ac.in

Abstract. Speech emotion recognition (SER) is a non-trivial task considering that the very definition of emotion is ambiguous. In this paper, we propose a speech emotion recognition system that predicts emotions for multiple segments of a single audio clip unlike the conventional emotion recognition models that predict the emotion of an entire audio clip directly. The proposed system consists of a pre-trained deep convolutional neural network (CNN) followed by a single layered neural network which predicts the emotion classes of the audio segments. The predictions for the individual segments are finally combined to predict the emotion of a particular clip. We define several new types of accuracies while evaluating the performance of the proposed model. The proposed model attains an accuracy of 68.7% surpassing the current state-of-the-art models in classifying the data into one of the four emotional classes (*angry, happy, sad* and *neutral*) when trained and evaluated on IEMOCAP audio-only dataset.

Keywords: Emotion Recognition · Affective Computing · Deep Learning · Mel Spectrograms · Computational Paralinguistics

1 Introduction

Speech is one of the most natural means of communication. The semantics as well as the emotional prosody of speech are both essential for conveying any information through it. Despite the remarkable advances made in speech related tasks such as speech recognition [1] and text-to-speech synthesis [20], *natural* emotion understanding is still an unaccomplished capability for the computational systems. Speech emotion recognition is essential in the domains that require a significant amount of man-machine interaction. In the recent years, conversational interfaces or voice assistants have become ubiquitous through smartphones and home automation [3]. These systems will perform better in certain situations if they can capture and process both the semantics as well as the emotional content of speech. Speech emotion recognition is challenging due to a number of reasons. It is difficult to strictly categorize different emotions because the very definition of emotion is obscure [13]. Large scale annotated emotional datasets are required for the training of complex emotion recognition systems. However, creating such large datasets is cost prohibitive due to the extensive human efforts involved, which is another significant challenge.

In this work, we attempt to address the data insufficiency challenge. Following two ideas are presented to overcome the same: 1) train and test the classification model on *multiple segments* of the audio clip rather than the entire audio clip as a whole and 2) use *transfer learning* to improve the model performance. In the recent years, transfer learning has successfully tackled data insufficiency challenge up to a great extent. In this study, we specifically use *inductive transfer learning*, a transfer learning method in which the horizon of possible models is reduced by implementing a model trained on a different but related task [19]. We propose a new emotion recognition model which uses Google VGGish [10], a deep convolutional neural network followed by a single layered neural network for classification. We conducted multiple experiments to investigate various architectures and hyperparameters. The model when trained and evaluated on overlapping segments, achieves an accuracy of 68.7% and outperforms the current state-of-the-art model [30] by 6.3% relative (4.1% absolute) accuracy in speech emotion recognition on IEMOCAP audio-only dataset.

The rest of the paper is organized as follows. Existing techniques in the context of speech emotion recognition have been reviewed in Section 2. Section 3.1 depicts two methods to partition the audio clips in the dataset into multiple segments. The model predicts the emotion class for multiple segments of a single audio clip rather than predicting the emotion class of the entire audio clip at once. The individual predictions are finally incorporated for predicting the emotion class of the entire clip. Several new types of accuracies are defined in Section 3.3 to evaluate the performance of the model. In Section 4.4, we discuss the outcome of our experiments and compare the performance of the proposed model with the existing models in speech emotion recognition and finally conclude in Section 5.

2 Related Work

Speech emotion recognition is a well studied research area in which several architectures, techniques and approaches have been deployed. In this section we briefly review the existing work in this domain.

2.1 Traditional Machine Learning Approaches

Traditional machine learning methods such as hidden markov models (HMM), support vector machines (SVM) and decision-trees etc. have been utilized for speech emotion recognition problems [24, 25, 15]. A recent work by Sahu [22] has shown that an ensemble of multiple traditional machine learning methods can achieve performance as good as the latest models in emotion recognition. All these methods extensively explored various features that determine the emotion contained in speech. However, a major drawback of traditional machine learning techniques is that a prior knowledge of all the necessary features that influence emotion recognition like fundamental frequency (F0), energy etc. is required.

2.2 Deep Learning Approaches

Deep neural networks were deployed for automatic extraction of high-level features from audio and were shown to be successful for speech emotion recognition [9]. Since then, several neural network architectures have been deployed for this task. Zheng et al. [32] did an experimental study on the use of convolutional neural network (CNN) for speaker independent emotion recognition system. Their system determined that deep learning methods outperform traditional machine learning techniques for SER. Variants of recurrent neural networks (RNN) like bidirectional long-short term memory (BLSTM) have proven to be successful in emotion recognition [16]. In an another work, Trigeorgis et al. [29] deployed a combination of CNN and RNN to efficiently recognize emotions in speech samples.

2.3 Audio Segmentation based Approaches

It was demonstrated that a speech segment longer than 0.25 seconds carries sufficient information for detecting the emotion present in it [21]. Since then, various research attempts have been made to detect emotion from multiple segments of audio clip instead of processing the clip at once. A natural advantage of using segments instead of clips is that the model learns the salient features that determine the presence of a particular emotion in speech in a more elaborated manner. On the other hand, a potential downside of using segments is that slicing the clip into non-overlapping segments causes loss of correlation and flow of the speech. In this context, a study by Shami and Kamel [26] combined the use of segment level and utterance level features for emotion recognition. Satt et al. [23] presented a system which detected emotion at segment level whose performance was comparable to the state-of-the-art model in SER.

2.4 Transfer Learning in SER

Transfer learning has been applied in SER in multiple ways. One of the approaches is learning features from one emotion dataset and applying it on another emotion dataset. Since many paralinguistic tasks are closely related, a different approach is learning the features from other paralinguistic tasks such as - speaker or gender recognition and applying it on emotion recognition [8]. Both the approaches were successful but the idea relies on paralinguistic datasets which are currently very limited. Badshah et al. [2] made an attempt to use pre-trained model in emotion recognition where they compared the performance of fine-tuned AlexNet [14] and a CNN trained from scratch. However, the freshly trained CNN outperformed the fine-tuned AlexNet which was not very surprising considering that AlexNet is trained on ImageNet [6], a large-scale image database.

In this paper, segments are used rather than entire clips for emotion detection. To investigate the loss of correlation between segments, both overlapping and non-overlapping segmentations are tried. We use mel spectrograms as opposed to numerous low-level hand-crafted features which were practiced in traditional machine learning approaches. Instead of using models trained on image dataset, a neural network trained on audio dataset is utilized for better transfer of knowledge.

3 The Proposed Method

The audio clips are segmented using two different methods i.e. overlapping and non-overlapping audio segmentation. These segments are given as inputs to the proposed system. The proposed system comprises of a generator which generates mel spectrogram from raw audio input which is passed into a pre-trained deep CNN. The CNN produces a 128-dimensional embedding from the mel spectrogram which passes through a single layered neural network which finally predicts the emotion class. The entire methodology has been shown in Figure 1 and elaborated in the following subsections.

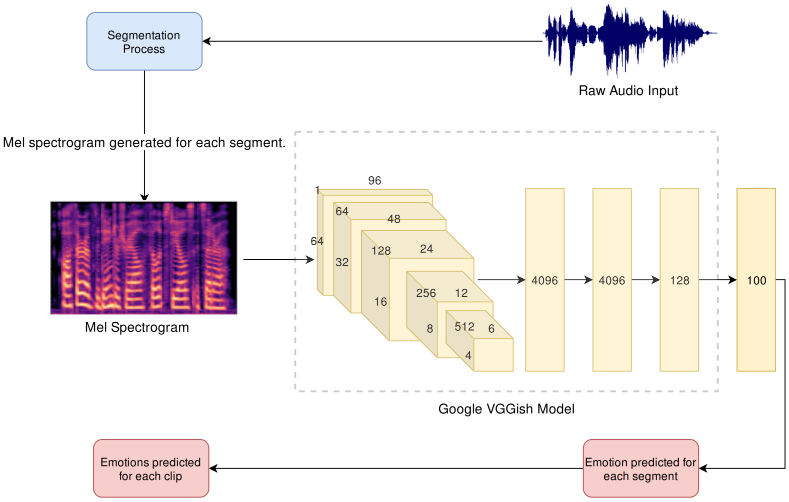


Fig. 1. Visual representation of the proposed method

3.1 Audio Segmentation

Non-overlapping Segmentation We extract non-overlapping segments of one-second duration and pad the last segment with silence to also make its length one second. By applying this process for the entire dataset, we get a total of 49795 segments out of which we use 27935 segments that belong to the classes that are relevant in this context. Table 1 may be referred for detailed sample distribution of the dataset after non-overlapping segmentation.

Overlapping Segmentation We extract overlapping segments of one-second duration and pad the last segment with silence to also make its length one second. The overlapping duration is 0.5 seconds for all the segments. Overlapping segmentation serves two main purposes: 1) it captures better correlation among the segments of the clip 2) it increases the number of data points. By applying the process for the entire dataset, we get a total of 91017 segments out of which we use 51180 segments that belong to the classes that are relevant in this context. The detailed sample distribution of the dataset after overlapping segmentation has been presented in Table 1.

For both types of segmentations, we hypothesize that if an utterance belongs to class X, then each segment of the utterance also belongs to the same class X. A visual representation of the segmentation of a toy example has been shown in Figure 2.

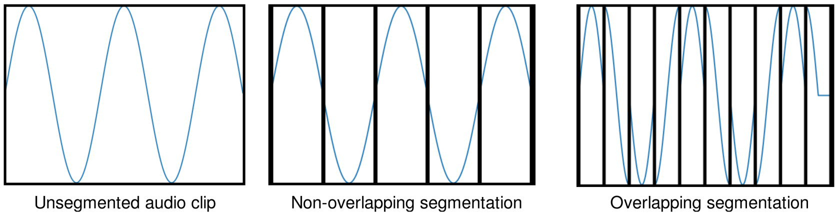


Fig. 2. Segmentation process. The first image shows an unsegmented audio clip. The second image shows the five segments obtained after non-overlapping segmentation. The third image shows the segments obtained after overlapping segmentation. The duration of each segment is *same* in both the cases. The zero padding which is done during segmentation process is visible in the last segment in case of overlapping segmentation.

3.2 Spectrogram Generator

Mel spectrogram generation method, as adopted from Google VGGish paper [10] has been described as follows. For each 1000 ms segment, Short-Time Fourier Transform (STFT) magnitude is computed using a 25 ms length window, 10 ms hop and Hann window function. Mathematically, the expression for Hann window is:

$$w[k] = 0.5\left[1 - cos\left(\frac{2\pi k}{N}\right)\right] \tag{1}$$

where N = window length and $k = 0, 1, 2 \dots N-1$. The log mel-spectrogram of 96×64 patches has been obtained from the resulting spectrogram by integrating it into the 64 mel-spaced frequency bins. A small offset of 0.01 has been added to avoid numerical issues and then the magnitude of each bin is log transformed.

3.3 Evaluation Metrics

The unsegmented dataset is represented as $\mathcal{D}_{\mathcal{US}} = \{(c_0, y_0), \dots, (c_{N-1}, y_{N-1})\}$ where c_i is the i^{th} clip, y_i is the corresponding emotion class and N is the number of relevant utterances in total i.e. 5531 in this case. Let $\mathcal{N} = \{n_0, \dots n_{N-1}\}$ where n_i is the number of segments of clip c_i for $i = 0, 1, \dots, N-1$. After segmentation, the dataset is $\mathcal{D}_{\mathcal{S}} = \left\{(s_{0,0}, y_0), \dots (s_{0,n_0}, y_0), (s_{1,0}, y_1), \dots, (s_{N-1,n_{N-1}}, y_{N-1})\right\}$ where $s_{i,j}$ represents the j^{th} segment of the i^{th} clip and y_i represents the corresponding emotion class. Let the segmented test dataset be $\mathcal{T}_{\mathcal{S}} \subset \mathcal{D}_{\mathcal{S}}$ and $\mathcal{Y}_{\mathcal{S}} = \{y_0, \dots, y_{T-1}\}$ be the correct emotion classes of the segments in $\mathcal{T}_{\mathcal{S}}$, where $T = |\mathcal{T}_{\mathcal{S}}|$. Suppose a clip $c \in \mathcal{T}_{\mathcal{S}}$ consists of k segments $s_0, \dots s_{k-1}$ and $\mathcal{P} = \{p_0, \dots p_{k-1}\}$ be the corresponding prediction, where p_i is the predicted emotion class for s_i for $i = 0, 1, \dots, k-1$. Let $\mathcal{M}$ be the set of emotions that are predicted for maximum number of segments of c i.e. the set of elements that appear maximum number of times in $\mathcal{P}$. Explanation of various types of accuracies is provided in the subsequent sections and more details about their calculation procedure is available in the supplementary material.

Segment Accuracy (SA) Segment Accuracy is the percentage of the test segments predicted correctly.

Absolute Clip Accuracy (ACA) It is stated that the model has classified the clip correctly if the model predicts correct emotion class for *each* s_i in c for $i = 0, 1, \dots, k-1$. The percentage of clips that are classified correctly using the aforementioned criterion is defined as Absolute Clip Accuracy.

Standard Clip Accuracy (SCA) If $|\mathcal{M}| = 1$ and the emotion in $\mathcal{M}$ is the correct emotion class of the clip, then it is stated that the model has classified correctly. Standard Clip Accuracy is the percentage of clips that are classified correctly using the aforementioned criterion.

Average Logits Clip Accuracy (ALCA) The final layer of classification model, also called the logits layer, gives an n-length array of floating point values called logits, where n is the number of classes (which is 4 in this context). We compute the average value of logits over all the segments of a particular clip and state the argument of the maximum value in the average n-length array as the predicted class. The percentage of clips that are classified correctly using the aforementioned criterion is called Average Logits Clip Accuracy.

Best Clip Accuracy (BCA) If the correct emotion of the clip $e \in \mathcal{M}$, then we state that the model has classified the clip correctly. The percentage of clips that are classified correctly using this criterion is termed as Best Clip Accuracy.

4 Experiments and Results

4.1 Experimental Setup

The experiments have been performed on NVIDIA Quadro P5000 graphics processing unit (GPU) and Intel Xeon central processing unit (CPU) using TensorFlow Deep Learning library [1]. The model training utilized mini-batch size of 32 and Adam optimizer [12] with learning rate of 10^{-6}. The learning rate is kept very low as compared to the default value of 10^{-3} because we are fine-tuning a pre-trained model instead of training it from scratch. The model is prone to overfitting due to lack of sufficient data as well as very high number of parameters. So, early stopping [5] is used to counter overfitting by fixing the value of patience to be 50.

4.2 Dataset

The proposed model is trained and evaluated on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [4]. The dataset consists of five recorded sessions of conversations, each containing utterances from two speakers (one female and one male). The dataset contains audio, audio+video and corresponding transcriptions. In this paper, the audio-only dataset is used. The audio clips are sampled at 16 kHz. Each of 10039 utterances is classified into one of the following classes - angry, happy, sad, neutral, frustrated, excited, fear, surprise, disgust and others. We use only four emotion classes i.e. *angry, happy, sad* and *neutral* for consistent comparison with the previous works that used IEMOCAP dataset [30, 31] and utterances labelled as *excited* are merged with those labelled as *happy*. So, the final dataset contains 5531 utterances. The dataset is divided into train, validation and test sets in the ratio 8:1:1. Table 1 may be referred for detailed sample distribution of the dataset.

4.3 Model Architecture

A baseline model is defined against which we compare our results. The proposed architecture has two components: a) VGGish, which is a deep CNN and b) a single layered neural network, also called the classification model. The detailed architecture of the models is discussed in the following sections.

[1] https://www.tensorflow.org/

Table 1. Sample distribution of IEMOCAP dataset before and after segmentation

	Before Seg.	**Non-overlapping Seg.**	**Overlapping Seg.**
Total	10039	49795	91017
Relevant	5531	27935	51180
Rel. Frac.	55.10%	56.10%	56.23%
Angry	19.94%	19.80%	19.75%
Happy	29.58%	30.04%	30.06%
Sad	19.60%	23.25%	23.76%
Neutral	30.88%	26.91%	26.44%

where Seg. = Segmentation, Rel. Frac. = Relevance Fraction i.e. percentage of the entire dataset that we are considering for this paper. It is to be noted that *happy* includes both *happy* and *excited* data.

Baseline Model The input to the baseline model is 96 × 64 dimensional mel spectrogram of non-overlapping segments of audio clips in the training set. A fully connected neural network of N layers with M units in each layer is considered. We try $N = [2, 3, 4]$ and $M = [100, 200]$ and choose the best performing model. A dropout [28] layer with $p = 0.5$ and batch normalization [11] layer is used between every fully connected layer. All the fully connected layers use ReLU [17] activation function. We get the best results with $N = 2$ layers, $M = 200$ units which has been considered as the baseline model.

Proposed Architecture

- **VGGish Model:** The Google VGGish [10] model is a deep convolutional neural network which has an architecture very similar to that of VGG [27] model that was designed for large scale image classification. The VGGish model has been pre-trained on AudioSet [7], a collection of ∼2M human labelled ten second length audio clips from YouTube videos spread over ∼600 sound classes. The VGGish model takes a 96 × 64 dimensional mel spectrogram as input. The VGGish architecture comprises of four blocks of two dimensional convolution and max-pooling layers. The final max-pooling layer is followed by 2 fully connected layers each comprising of 4096 units and finally a fully connected layer of 128 units which generates the embedding vector. All the convolution and fully connected layers use ReLU activation function. The model has ∼72 million parameters. The architecture of VGGish model has been described in Table 2.

- **Classification Model:** The 128-dimensional embedding vector passes into to a single layered neural network comprising of a fully connected layer with N units followed by the final fully connected layer, the logits layer, which predicts the emotion class of each segment of the audio clip. We have tried different values of $N = [100, 200, 400]$ to find the best performing model.

Table 2. VGGish Network Architecture

Layer	Activation Size
input	1 × 96 × 64
64 × 3 × 3 conv, stride 1	64 × 96 × 64
2 × 2 maxpool, stride 2	64 × 48 × 32
128 × 3 × 3 conv, stride 1	128 × 48 × 32
2 × 2 maxpool, stride 2	128 × 24 × 16
256 × 3 × 3 conv, stride 1	256 × 24 × 16
256 × 3 × 3 conv, stride 1	256 × 24 × 16
2 × 2 maxpool, stride 2	256 × 12 × 8
512 × 3 × 3 conv, stride 1	512 × 12 × 8
512 × 3 × 3 conv, stride 1	512 × 12 × 8
2 × 2 maxpool, stride 2	512 × 6 × 4
flatten	1 × 12288
fully connected I	1 × 4096
fully connected II	1 × 4096
output	1 × 128

where $C \times H \times W$ conv denotes a 2D convolutional layer with C filters of size $H \times W$. $H \times W$ maxpool denotes a max-pooling layer of pooling size $H \times W$.

4.4 Results and Discussion

The model is trained and evaluated six times on IEMOCAP audio-only dataset and the mean accuracy and standard deviation is reported. The single layered neural network that follows the VGGish model consists of $N = 200$ units because the value of N corresponding to the model with best performance is observed to be 200. The performance of the proposed system has been compared for overlapping and non-overlapping segments using the evaluation metrics mentioned in Section 3.3 in Table 3.

Table 3. Performance of model with overlapping and non-overlapping segmentation

	Non-overlapping Seg.	Overlapping Seg.
SA	$\mathbf{0.564 \pm 0.006}$	0.561 ± 0.011
ACA	$\mathbf{0.196 \pm 0.010}$	0.125 ± 0.010
SCA	0.559 ± 0.009	$\mathbf{0.621 \pm 0.019}$
ALCA	0.668 ± 0.014	$\mathbf{0.687 \pm 0.019}$
BCA	$\mathbf{0.707 \pm 0.011}$	0.703 ± 0.017

where SA: Segment Accuracy, ACA: Absolute Clip Accuracy, SCA: Standard Clip Accuracy, ALCA: Average Logits Clip Accuracy, BCA: Best Clip Accuracy.

In Table 4, we compare the performance of the model when overlapping segments are given as input to three different models with $N = 100, N = 200$ and $N = 400$ units in the penultimate layer of the model. The performance of the proposed model is compared with various existing state-of-the-art models in SER in Table 5. We have calculated multiple evaluation metrics as defined in Section 3.3. However, *Average Logits Clip Accuracy* has been primarily used for the comparison because existing studies on segment level approach in SER have used similar metrics to compare their models with the conventional models [23].

Table 4. Comparison of models with different values of N on overlapping segments

	$\boldsymbol{N = 100}$	$\boldsymbol{N = 200}$	$\boldsymbol{N = 400}$
SA	0.560 ± 0.012	$\mathbf{0.561 \pm 0.011}$	0.559 ± 0.017
ACA	$\mathbf{0.132 \pm 0.012}$	0.125 ± 0.010	0.131 ± 0.010
SCA	0.616 ± 0.017	$\mathbf{0.621 \pm 0.019}$	0.617 ± 0.022
ALCA	0.684 ± 0.020	$\mathbf{0.687 \pm 0.019}$	0.674 ± 0.018
BCA	$\mathbf{0.703 \pm 0.013}$	0.703 ± 0.017	0.699 ± 0.020

where SA: Segment Accuracy, ACA: Absolute Clip Accuracy, SCA: Standard Clip Accuracy, ALCA: Average Logits Clip Accuracy, BCA: Best Clip Accuracy.

Table 5. Comparison of different state-of-the-art models on IEMOCAP dataset

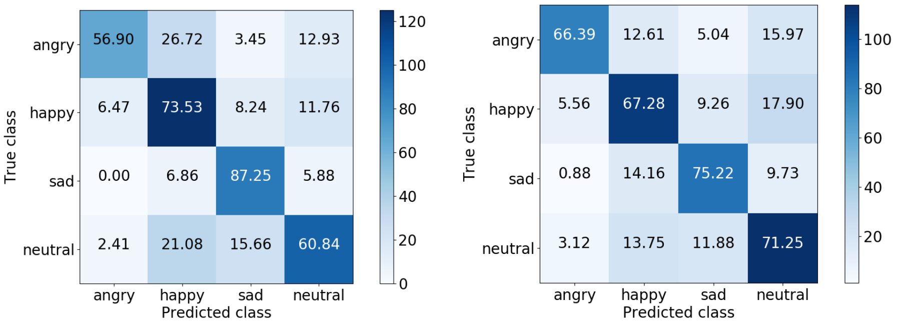
Model Name	**Modality**	**Accuracy**
Baseline	A	0.511
ARE [31]	A	0.546
ACNN [18]	A	0.561
Ensemble [22]	A	0.562
audio-BRE [30]	A	0.646
Ensemble [22]	T	0.631
TRE [31]	T	0.635
Proposed (Non-overlap)	A	0.668 ± 0.014
Proposed (Overlap)	A	$\mathbf{0.687 \pm 0.019}$

where ARE: Audio Recurrent Encoder, ACNN: Attentive CNN, Ensemble: Ensemble of six traditional machine learning methods, BRE: Bidirectional Recurrent Encoder, TRE: Text Recurrent Encoder. A = Audio-only, T = Text-only

Discussion Even though SA of overlapping segments is 0.3% lower than that of non-overlapping segments, the SCA of the model trained on overlapping segments is 6.2% better than the model trained on non-overlapping segments. Similarly, the ALCA in case of overlapping segments is 1.9% better than non-overlapping segments. This is expected because overlapping segments capture the correlation between the different segments of the clip which is not present in non-overlapping segments. The sharp difference in SCA of overlapping and non-overlapping segments proves that the model learns to recognize the emotions precisely in case of overlapping segments because by definition, SCA is the percentage of clips that are said to be classified correctly when the model predicts *exactly one* emotion for majority of the segments of a clip. The difference between SCA and BCA in case of overlapping segments (8.2%) is much lower than that of non-overlapping segments (14.8%). This observation shows that instead of predicting multiple classes with equal frequency for the segments of a particular clip, the model can pick out a *single* class with maximum frequency when trained on overlapping segments. There is a 7.1% decrease in ACA in case of overlapping segments which is anticipated due to the increased number of segments (almost 2x) of the clip because of which it is difficult to predict the correct class for *every* segment of a clip.

In Table 4, we observe the performance of model when trained on overlapping segments as N varies from 100 to 400. As we increase N from 100 to 400, we do not notice any strict trend in the performance of the model. However, all the experiments resulted in a slightly better ALCA of the model with $N = 200$ units which we use in Table 5 for comparison with the existing models in SER. Apart from ALCA, SA and SCA are also better for $N = 200$ than $N = 100$ or $N = 400$.

As per the surveyed literature, the current state-of-the-art model for SER on IEMOCAP audio-only dataset is audio-BRE, a bidirectional recurrent encoder. The proposed model shows higher performance than the audio-BRE by 6.3% relative (0.646 to 0.687 absolute) accuracy. Most of the existing works on IEMOCAP dataset have shown that emotion recognition systems show better performance on transcriptions than the audio-only dataset [22, 30, 31]. Although our model is trained on audio-only dataset, it shows better performance than some of the models trained on text-only dataset, as demonstrated in Table 5.



(a) Non-overlapping Segmentation

(b) Overlapping Segmentation

Fig. 3. Confusion Matrix

We compute the confusion matrix for both overlapping and non-overlapping cases which is presented in Figure 3. In Figure 3a, we observe that model incorrectly classifies most examples of *angry* as *happy* (26.72%). The true positives in case of non-overlapping is low for *angry* (56.9%) and *neutral* (60.84%) as compared to the other emotions. These numbers are improved when we observe the case of overlapping segments i.e. 66.39% for *angry* and 71.25% for *neutral*. Although Neumann and Vu [18] and Yoon et al. [31] have shown that most of the emotions are gotten confused with *neutral* class during SER because it lies in the centre of activation-valence space, the proposed model shows much less confusion when trained on overlapping segments.

5 Conclusion

We present a segment level approach for speech emotion recognition using transfer learning in this paper. The proposed approach consisting of a single layered neural network on top of a pre-trained CNN outperformed the current state-of-the-art emotion classification model on IEMOCAP audio-only dataset by a relative accuracy of 6.3%. The improved performance proves the applicability of transfer learning for SER. In future, we will focus on incorporating transcriptions and audio-visual data to design a model with better performance in emotion recognition. A different research could focus on developing an intelligent segmentation process instead of using fixed segment length or fixed overlapping duration.

References

1. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al.: Deep speech 2: End-to-end speech recognition in english and mandarin. In: International Conference on Machine Learning (ICML). pp. 173–182 (2016)
2. Badshah, A.M., Ahmad, J., Rahim, N., Baik, S.W.: Speech emotion recognition from spectrograms with deep convolutional neural network. In: International Conference on Platform Technology and Service (PlatCon). pp. 1–5. IEEE (2017)
3. Braun, M., Mainz, A., Chadowitz, R., Pfleging, B., Alt, F.: At your service: Designing voice assistant personalities to improve automotive user interfaces. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. p. 40. ACM (2019)
4. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation **42**(4), p. 335 (2008)
5. Caruana, R., Lawrence, S., Giles, C.L.: Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In: Advances in Neural Information Processing Systems. pp. 402–408. (2001)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: Computer Vision and Pattern Recognition (CVPR). IEEE (2009)
7. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780. IEEE (2017)

8. Gideon, J., Khorram, S., Aldeneh, Z., Dimitriadis, D., Provost, E.M.: Progressive neural networks for transfer learning in emotion recognition. arXiv preprint arXiv:1706.03256 (2017)
9. Han, K., Yu, D., Tashev, I.: Speech emotion recognition using deep neural network and extreme learning machine. In: Fifteenth annual conference of the International Speech Communication Association (INTERSPEECH). pp. 223–227. ISCA (2014)
10. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: CNN architectures for large-scale audio classification. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 131–135. IEEE (2017)
11. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Kleinginna, P.R., Kleinginna, A.M.: A categorized list of emotion definitions, with suggestions for a consensual definition. Motivation and Emotion **5**, pp. 345–379 (1981)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105 (2012)
15. Lee, C.C., Mower, E., Busso, C., Lee, S., Narayanan, S.: Emotion recognition using a hierarchical binary decision tree approach. Speech Communication **53**(9-10), pp. 1162–1171 (2011)
16. Lee, J., Tashev, I.: High-level feature representation using recurrent neural network for speech emotion recognition. In: Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH). pp. 1537–1540. ISCA (2015)
17. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML). pp. 807–814 (2010)
18. Neumann, M., Vu, N.T.: Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. arXiv preprint arXiv:1706.00612 (2017)
19. Pan, S.J., Yang, Q.: A survey on transfer learning. Transactions on knowledge and data engineering **22**(10), pp. 1345–1359. IEEE (2009)
20. Ping, W., Peng, K., Gibiansky, A., Arik, S.O., Kannan, A., Narang, S., Raiman, J., Miller, J.: Deep voice 3: Scaling text-to-speech with convolutional sequence learning. arXiv preprint arXiv:1710.07654 (2017)
21. Provost, E.M.: Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3682–3686. IEEE (2013)
22. Sahu, G.: Multimodal speech emotion recognition and ambiguity resolution. arXiv preprint arXiv:1904.06022 (2019)
23. Satt, A., Rozenberg, S., Hoory, R.: Efficient emotion recognition from speech using deep learning on spectrograms. In: Eighteenth Annual Conference of the International Speech Communication Association (INTERSPEECH). pp. 1089–1093. ISCA (2017)
24. Schuller, B., Rigoll, G., Lang, M.: Hidden markov model-based speech emotion recognition. In: Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. II–1. IEEE (2003)

25. Seehapoch, T., Wongthanavasu, S.: Speech emotion recognition using support vector machines. In: 5th International Conference on Knowledge and Smart Technology (KST). pp. 86–91. IEEE (2013)
26. Shami, M.T., Kamel, M.S.: Segment-based approach to the recognition of emotions in speech. In: International Conference on Multimedia and Expo. pp. 4–pp. IEEE (2005)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
28. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research **15**(1), pp. 1929–1958 (2014)
29. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., Zafeiriou, S.: Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5200–5204. IEEE (2016)
30. Yoon, S., Byun, S., Dey, S., Jung, K.: Speech emotion recognition using multi-hop attention mechanism. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2822–2826. IEEE (2019)
31. Yoon, S., Byun, S., Jung, K.: Multimodal speech emotion recognition using audio and text. In: Spoken Language Technology Workshop (SLT). pp. 112–118. IEEE (2018)
32. Zheng, W., Yu, J., Zou, Y.: An experimental study of speech emotion recognition based on deep convolutional neural networks. In: International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 827–831. IEEE (2015)