A Segment Level Approach to Speech Emotion Recognition using Transfer Learning

Sourav Sahoo[1], Puneet Kumar[2], Balasubramanian Raman[2], Partha Pratim Roy[2]
[1] Indian Institute of Technology, Madras, [2] Indian Institute of Technology, Roorkee

Highlights

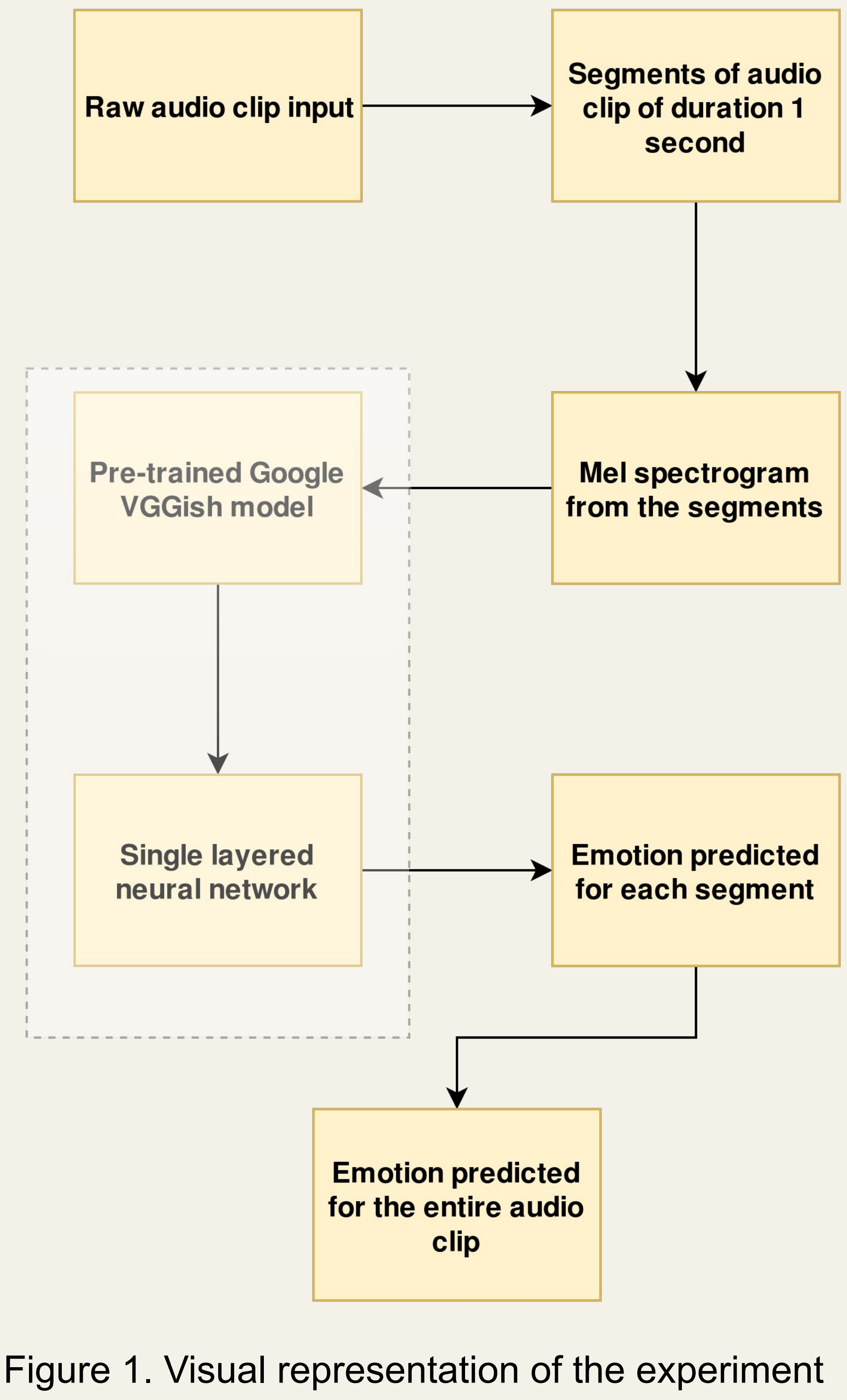
- Proposed a speech emotion recognition system that **predicts emotions for multiple segments** of a single audio clip**.**
- Experimented with both **overlapping and non-overlapping** audio segmentation
- Defined several **new evaluation metrics**
- Achieved **68.7%** accuracy on IEMOCAP audio-only dataset.

Introduction

Speech is the most natural means of communication. Although remarkable advances have been made in speech related tasks such as speech recognition, natural emotion perception is still an unaccomplished capability for the computational systems. Speech emotion recognition is essential in the domains that require substantial man-machine interaction. Building a **robust emotion classifier** is a solution to this problem. We propose a model that uses **transfer learning** to predict emotions (angry, happy, sad and neutral) of a given audio clip.

Proposed Method

- The audio clips are segmented using two different methods: 1) overlapping and 2) non-overlapping audio segmentation. These segments are given as inputs to the proposed system.
- The 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 which passes through a single layered neural network which finally predicts the emotion class. The entire methodology has been shown in Figure 1.



Raw audio clip input

Segments of audio clip of duration 1 second

Mel spectrogram from the segments

Pre-trained Google VGGish model

Single layered neural network

Emotion predicted for each segment

Emotion predicted for the entire audio clip

Figure 1. Visual representation of the experiment procedure

Dataset and Evaluation Metrics

Dataset: IEMOCAP[1] audio-only dataset. We use only four emotion classes i.e. **angry, happy, sad and neutral**. Utterances labelled as excited are merged with those labelled as happy for consistency with previous works [2,3]. The final dataset contains **5531 utterances**. The dataset is split into training,validation and test sets in the ratio 8:1:1.

Types of accuracies defined:

- **Segment Level Accuracy:** Percentage of test segments predicted correctly.
- **Absolute Clip Accuracy**: Percentage of clips for which all the segments were predicted correctly.
- **Standard Clip Accuracy**: The predicted class of a clip is the class which is predicted for maximum number of segments. Standard Clip Accuracy is the percentage of clips classified using this criterion.
- **Average Logits Clip Accuracy**: We compute the average value of logits over all the segments of a particular clip and the argument of the maximum value in the average logits array is the predicted class. Percentage of clips classified correctly using this criteria is Average Logits Clip Accuracy.
- **Best Clip Accuracy**: If the actual class of the clip is present in the list of the classes that were predicted maximum number of times, the model is said to have predicted correctly. Percentage of clips classified correctly using criteria is Best Clip Accuracy.

Experiments

- **Non-overlapping Segmentation:**
  - We extract non-overlapping segments of one-second duration.
  - By applying this process for the entire dataset, we get ~28K segments.
- **Overlapping Segmentation:**
  - We extract overlapping segments of one-second duration.
  - The overlapping duration is 0.5 seconds for all the segments.
  - By applying the process for the entire dataset, we get ~51K segments.

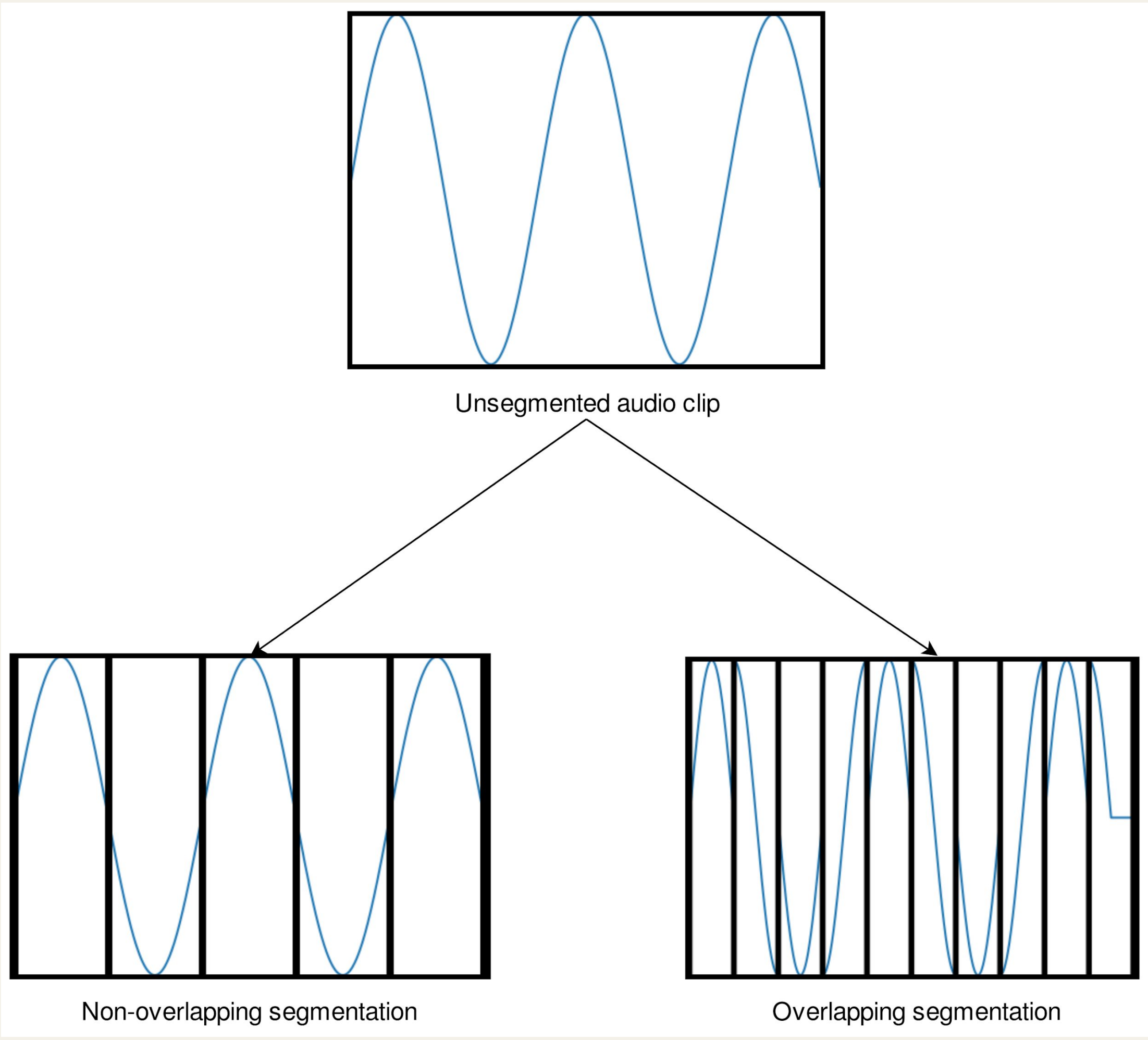


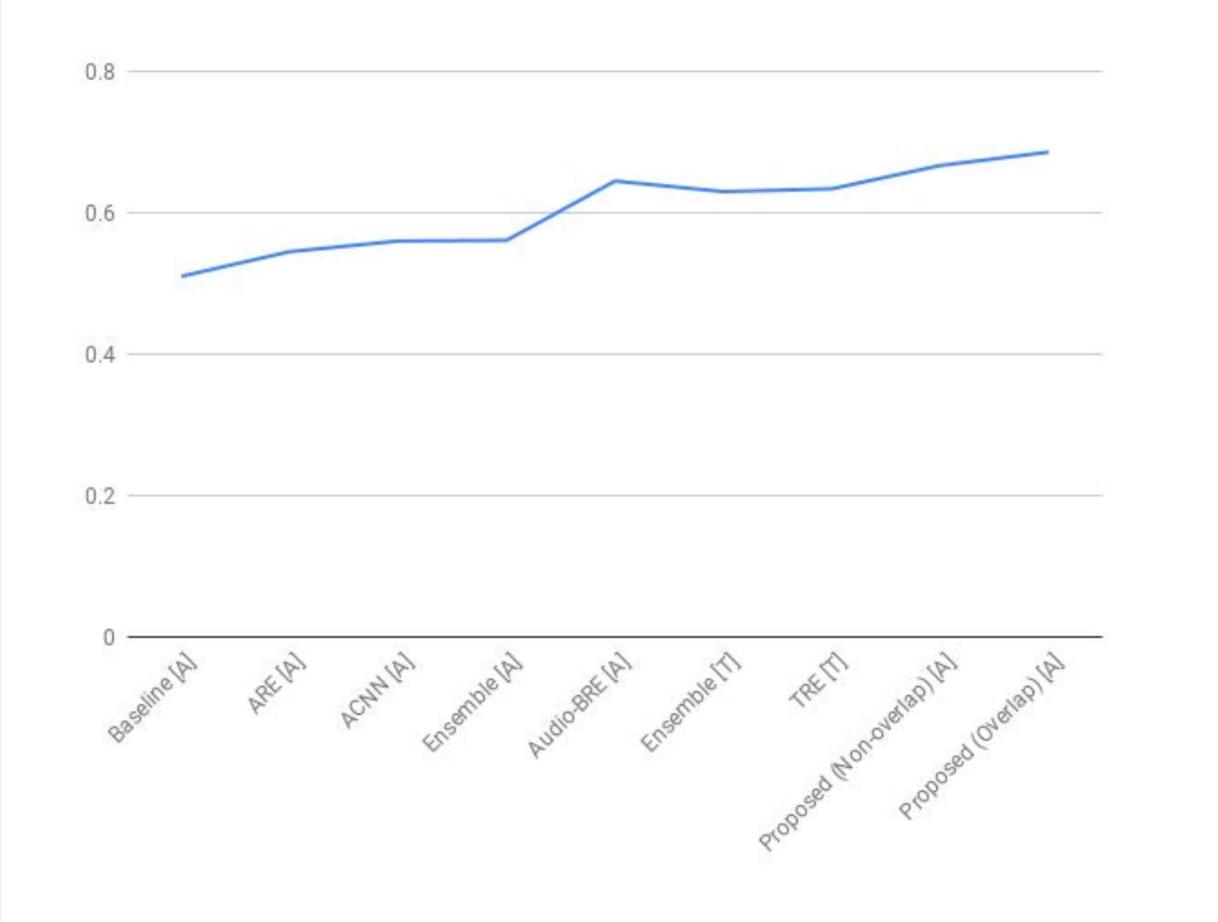
Figure 2. Segmentation process. The first image shows an unsegmented audio clip. The duration of each segment is same in both the cases.

Results

Table 1. Performance of model with overlapping and non-overlapping segmentation.

	Non-overlapping Segments	Overlapping Segments
SA	**0.564**	0.561
ACA	**0.196**	0.125
SCA	0.559	**0.621**
ALCA	0.668	**0.687**
BCA	**0.707**	0.703

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



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

Conclusion and Future Work

- 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] 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)

[2] 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)

[3] 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)

Poster template by ResearchPosters.co.za