An Attention based Neural Network on Multiple Speaker Diarization

Shao Wen Cheng[1], Kai Jyun Hung[2], Hsie Chia Chang[2], Yen Chin Liao[2]
[1]National Yang Ming Chiao Tung University, College of Electrical and Computer Engineering, Taiwan
[2]National Yang Ming Chiao Tung University, Institute of Electronics, Taiwan

Abstract

Speaker diarization is a task to label audio or video recordings with classes that correspond to speaker identity for each point in time, which can be used in a multispeaker conversation environment, such as a meeting or interview. Moreover, speaker diarization can be used to improve the performance of auto speech recognition. This paper presents an endtoend diarization model based on an attention mechanism with data augmentation, several data pre-processing, and post-processing. In the CALLHOME data set, the case of two speakers reached a 9.12% diarization error rate.

Our work combines the speaker diarization model, and auto speech recognition model and implement the transcript conversion system on an edge device. By using proposed speaker diarization as preprocessing to segment recording according to different speakers, then get the transcript of each segmented utterance by ASR model to fulfill the transcript conversion on the edge device. Experiment shows that our model also performs well in the scenario with two people on edge devices with both accuracy and inference time.

Introduction

The traditional diarization framework splits diarization into multiple tasks, voice activity detection, feature extraction, and clustering. It is time-consuming because those models should be trained separately. Most papers consider VAD as a separate task and usually ignore this part. Moreover, most of the papers ignored the overlapping speech issue due to using the clustering method. In 2019, Hitachi published " End to End Neural Diarization" (EEND). Different from previous works, EEND doesn't have multiple separated models. It took the speaker diarization task as a multilabel classification problem, so it can deal with overlapping speech and the function of VAD is also included. In this paper, our model is based on an attention mechanism to achieve an end-to-end model with more effective data pre-processing, post-processing, and layer design to increase the performance on speaker diarization tasks.

Result

We evaluate the proposed model on real diarization data, called CALLHOME, and compare it with other works SAEEND and SA-EEND-EDA. CALLHOME is disk 8 of NIST Speaker Recognition Evaluation which includes 2 to 7 speakers in each recording. To compare with SAEEND and SA-EEND-EDA, we only adopt recordings with 2 speakers then split these recordings into finetune set and testing set, which include 155 and 148 recordings respectively. The recordings in the fine-tune set and testing set are totally identical to the com pared works. Our proposed models, SA-EEND and SA-EEND-EDA use the same speaker set and noise set of simulated mixtures mentioned in section 3.5 as the training set. The training set of SAEEND is the same as ours, but the training set which SA-EEND-EDA used is 4 times larger than ours. All the models use the same fine-tune set to fine-tune.

Model	DER
Proposed convolution model of 2 speakers	9.47%
Proposed linear model of 2 speakers	9.12%
SA-EEND	9.54%
SA-EEND-EDA	8.07%

Table 2. Speaker diarization on CALLHOME

Conclusion

Our work successfully implement a speaker diarization application with some pre-processing and post-processing method, such as pre-emphasis, short-term Fourier transform, log Mel filter bank, utterance, and threshold choosing. The proposed diarization model can achieve 9.12% of DER on CALLHOME testing set. Achieving better multi-people speaker diarization or even fulfilling online speaker diarization will be our goal in the future.

Our work also combines our proposed diarization model and auto speech recognition model and implements the transcript conversion system on an edge device. Experiment shows that our model also performs well in the scenario with two people on edge devices with both accuracy and inference time.

Proposed Method

- **Data Preprocessing**

Pre-emphasis is implemented by subtracting each time's value from the previous time's value by a constant ratio α which is usually set as 0.97. Pre-emphasis increases the amplitude of high-frequency bands and decreases the amplitudes of lower bands.

Method	Diarization Error Rate (DER)
Without pre-emphasis	9.77%
With pre-emphasis	9.57%

Table 1. Comparison between using pre-emphasis or not

- **Utterance Reconstruction**

Our work segments the input recording into several utterances which are able to feed into the model individually. Each segmented utterance has an overlap part with the previous utterance. After inference, we will have the hypothesis of each utterance. Then use the overlap part of the hypotheses to determine the combination of utterances by calculating the binary cross-entropy loss of the overlap between the permuted hypotheses and the first hypothesis. The permuted hypothesis with the least loss is concatenated with the first hypothesis. If the input recording is split into k utterances, this procedure should be repeated k−1 times. The overlap factor in this paper is 100.

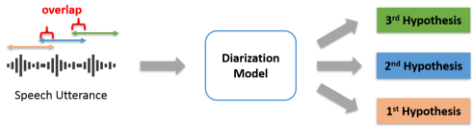

Fig 1. Segment the input recording

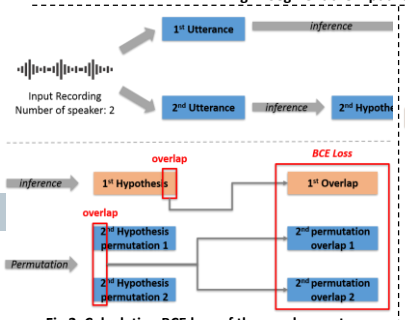

Fig 2. Calculating BCE loss of the overlap part

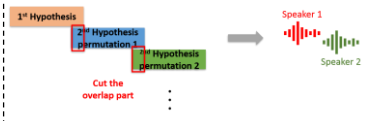

Fig 3. Concatenate the split utterance

- **Threshold Choosing**

Our work uses binary search to adjust the threshold to be closer to the best value for finding the lowest point of DER. The binary search iteration will be executed in the training stage and the DER of every epoch will be compared to the previous DER. If the DER is lower, the threshold will be recorded as the best threshold. If the best threshold remains unchanged for 10 epochs, the iteration will be stopped and set this value as the best threshold then adopted in the inference stage.

- **Proposed Diarization Model**

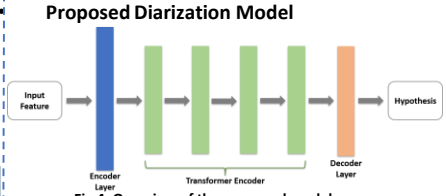

Fig 4. Overview of the proposed model

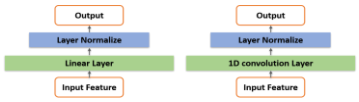

Fig 5. Two kinds of Encoder Layer

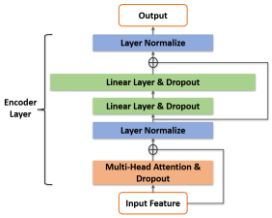

Fig 6. Transformer encoder layer.

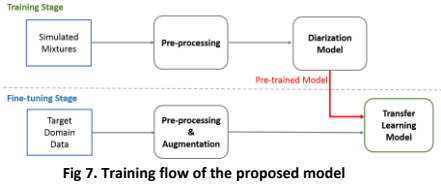

Fig 7. Training flow of the proposed model