CUNY Speech Diarization System for the CHiME-6 Challenge

Zhaoheng Ni[1], *Michael I Mandel*[2]

[1]The Graduate Center, City University of New York
[2]Brooklyn College, City University of New York
zni@gradcenter.cuny.edu, mim@sci.brooklyn.cuny.edu

Abstract

In this paper, we present a speech diarization system for the second track of the CHiME-6 challenge. Different from using the Agglomerative Hierarchical Clustering (AHC) algorithm in the baseline system, we apply the spectral clustering algorithm on the similarity matrices generated by probabilistic linear discriminant analysis (PLDA). To overcome the speech overlap problem, we apply a post-processing stage which detects the overlap in the segment and assign the segment to two speakers. The results show that our system reduces the word error rate to 76.04% for the development set and 72.74% for the evaluation set.

Index Terms: speech recognition, speech diarization, speech processing

1. Introduction

We participate in track 2 of the CHiME-6 challenge, which only provides the raw audio and the number of speakers in the audio. We didn't change the settings of the acoustic model or language model, thus all reported results belong to Category A of the track.

2. CUNY Diarization System

We use BeamformIt [1] as the front-end system to get the single-channel enhanced audio out of the multi-channel recordings. It is applied a single array, U06, which is the default array in the baseline system. Then we apply the pre-trained SAD model to detect speech segments. For each segment we extract the x-vectors with a 1.5-second window and a hop size of 0.75 seconds. Then we apply PLDA to get similarity matrices for all segment pairs in the same audio session. We use spectral clustering with the K-Means algorithm for diarization. Each segment is assigned to a cluster. To deal with regions of speech overlap, we predict which segments contain multiple speakers using a logistic regression classifier and assign these segments to the two temporally closest clusters.

2.1. Spectral Clustering

Spectral Clustering [2] is a graph-based clustering algorithm which has been applied in speech diarization [3, 4, 5, 6].

2.2. Overlap Post-processing

The main issue of clustering-based diarization systems is that a speech segment can only be assigned to one speaker. When there are multiple speakers talking at the same time, diarization error is thus unavoidable. To address this, we train a logistic regression classifier to predict whether a segment has overlap (i.e., multiple speakers) or not. We use x-vectors as the input feature for the classifier.

To create the labels for training, we use a gated function defined as follows:

$$l = \begin{cases} 1, & \text{if} \frac{N_O}{N_S} > \theta \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where N_O is the duration of overlapped region, N_S is the duration of the whole segment (i.e. 1.5 seconds for every segment). θ is the threshold that can be fine-tuned. We choose 0.5 and 0.67 as empirical values for θ in the experiments.

After we get the overlap prediction, we apply K-Means clustering on the spectral embeddings of those segments that are predicted to be a single talker and compute the cluster centers. We assign the segments that are predicted to contain multiple speakers to the two temporally closest clusters. Figure 1 visualizes the diarization result of the CUNY diarization system which uses spectral clustering with post-processing (0.67 threshold). The upper 4 rows represent the 4 speakers in the original RTTM reference provided by the challenge, the lower 4 rows represent the 4 speakers in our diazrization result.

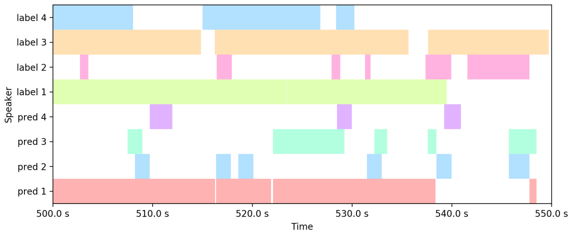

Figure 1: *Visualization of the diarization prediction by spectral clustering with post processing (0.67 threshold).*

3. Results

We follow the rules of category A of track 2: using the same acoustic model and language model as the baseline system for decoding lattices. There are three settings of our experiments:

- Using spectral clustering for clustering x-vectors (SPC)
- Using spectral clutsering and post-processing with 0.5 threshold. (SPC+PP (0.5))
- Using spectral clutsering and post-processing with 0.67 threshold. (SPC+PP (0.67))

Table 1 and Table 2 show the DER and JER results of the methods. In the baseline system there are two RTTM references for evaluating diarization performances. One is generated from the original CHiME-5 transcription annotations, the other is generated by running forced-alignment on the first channel of the binaural audio using the tri3 acoustic model in the baseline system. We reported DER and JER scores for both references.

By comparing with the original reference, spectral clustering with post-processing (0.67 threshold) achieves the best performance on the development set. However, we see a huge difference between the two reference DER scores if we consider the overlap in the diarization. For the baseline system and the spectral clustering method the difference is marginal. Table 2 shows the same trend as Table 1. Spectral clustering method achieves the best performance if we choose the original reference.

Reference	CHiME-5		Forced alignment	
Method	DER	JER	DER	JER
Baseline	61.56	63.42	69.75	70.83
SPC	57.15	61.77	57.55	61.18
SPC + PP (0.5)	54.60	52.53	78.83	57.79
SPC + PP(0.67)	51.67	54.45	63.81	57.20

Table 1: *The DER and JER scores of the development set by comparing the diarization predictions with the original CHiME-5 human-annotated reference (CHiME-5) and the reference generated by forced-aligning the binaural recordings (Forced alignment).*

Reference	CHiME-5		Forced alignment	
Method	DER	JER	DER	JER
Baseline	61.96	71.40	68.20	72.54
SPC	60.64	65.59	66.29	65.48
SPC + PP (0.5)	70.18	59.72	96.71	63.60
SPC + PP (0.67)	61.51	60.51	77.75	62.75

Table 2: *The DER and JER scores of the evaluation set by comparing the diarization predictions with the original CHiME-5 human-annotated reference (CHiME-5) and the reference generated by forced-aligning the binaural recordings (Forced alignment).*

Table 3 reports the WER results for all experiment settings. Different from the DER scores, spectral clustering with post-processing (0.67 threshold) achieves the lowest WER for both development and evaluation sets. This shows that the diarization performance may not be consistent with the speech recognition performance. To validate our hypothesis, we report the WER using the RTTM references as the diarization predictions and run lattice decoding. The scores can be regarded as the lower bound for category A. The results show the forced alignment reference achieves 63.33% WER compared with 67.46% for the original CHiME-5 reference. This indicates the forced aligned RTTM reference may not be a good standard reference for diarization, but it is a better reference for speech recognition.

To figure out the reason why there are such large differences in DER scores in the two RTTM references, we visualize excerpts from the two RTTM references in Figure 2. The forced alignment reference has more small segments while the CHiME-5 reference is more continuous. The blank regions between the small segments increase false alarm errors and reduce missed speech errors in the DER computation that makes the DER higher in the final result. Since we don't focus on speech separation and enhancement in our system, we believe removing those blank regions from the decoding stage helps improve the ASR performance, while it may not be true for the diarization. We would like to do more analysis by using more advanced front-end systems and retraining the acoustic model.

Method	Dev	Eval
Baseline	84.25	77.94
SPC	76.48	73.31
SPC + PP (0.5)	77.79	74.49
SPC + PP (0.67)	76.04	72.74
CHiME-5 reference	67.46	61.08
Forced align. reference	63.33	59.58

Table 3: *Word Error Rate results on the development and evaluation sets, including using both the CHiME-5 and binaural forced alignment references.*

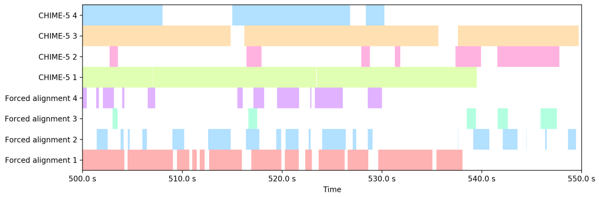

Figure 2: *Visualization of the two RTTM references provided by the challenge. The upper 4 and lower 4 rows are from the original CHiME-5 reference and the binaural forced alignment reference, respectively.*

4. Conclusions

We present our CUNY system for diarizing speech given only the raw audio and the number of speakers. The results show that a spectral clustering method achieves better performance than agglomerative hierarchical clustering. Post-processing to identify regions of speech overlap helps improve the speech recognition performance.

5. Acknowledgements

This material is based upon work supported by a Google Faculty Research Award and the National Science Foundation (NSF) under Grant IIS-1409431. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

6. References

[1] X. Anguera, "Beamformit, the fast and robust acoustic beamformer," 2006.

[2] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[3] H. Ning, M. Liu, H. Tang, and T. S. Huang, "A spectral clustering approach to speaker diarization," in *Ninth International Conference on Spoken Language Processing*, 2006.

[4] S. Shum, N. Dehak, and J. Glass, "On the use of spectral and iterative methods for speaker diarization," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[5] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, “Speaker diarization with lstm,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5239–5243.

[6] Q. Lin, R. Yin, M. Li, H. Bredin, and C. Barras, “Lstm based similarity measurement with spectral clustering for speaker diarization,” *arXiv preprint arXiv:1907.10393*, 2019.