Interspeech 2018
2-6 September 2018, Hyderabad

Fully automatic speaker separation system, with automatic enrolling of recurrent speakers

Raphael Cohen, Orgad Keller, Jason Levy, Amit Ashkenazi, Russell Levy, Micha Breakstone

Chorus.ai

{raphael,orgad,jason,amit,russell,micha}@chorus.ai

Abstract

We present a system to enable speaker separation and identification, designed to operate without requiring any effort from the end-user. In the system, single channel conversations are transformed into i-vectors, clustered into speakers and matched to a database of known speakers. Enrollment is automatic and a voice print is constructed for the recording user, taking advantage of the meta-data identifying that user's conversations. Further information is used when available from other information sources such as video and the ASR transcribed content to identify speakers. We describe the system architecture, novel unsupervised enrollment algorithm and describe the difficulties encountered in solving this problem.

Index Terms: speaker separation, diarization, speech recognition

1. Introduction

Sales conversations are a valuable and still underutilized asset for organizations. Recording and analyzing these conversations allow companies to quickly train new representatives, identify optimal behaviour, share and enforce best practices and also propagate customer requests and pain points to other parts of the organization helping product designers prioritize the best features.

Separating the conversation into multiple speakers and identifying them is important for identifying interesting sections in the conversation, *e.g.* an answer to a question, or optimizing behaviour, *e.g.* identify a monologue that goes too long. Furthermore, separating the conversation into multiple channels improves the resulting ASR transcript.

Ideally, conversations would be recorded with one channel per speaker. However, some recording platforms record mono audio, multiple people can join a web conference while sharing a single microphone, and face to face meetings are also recorded using a single microphone.

We present an end to end system which separates speakers for recorded conversations. The system has little information regarding the number of speakers and users are not expected to enroll. Since our users are the ones recording the conversations we can automatically enroll their voices using a novel algorithm. By diarizing all incoming channels the system supports multiple use cases including face to face meetings and overcoming multiple speakers sharing a meeting room.

2. System Architecture

The system is comprised of three parts:

- Embedding module: such as i-vector [1] or x-vector [2]
- Single recording diarization pipeline (see figure 1)
- Unsupervised enrollment and voice print database

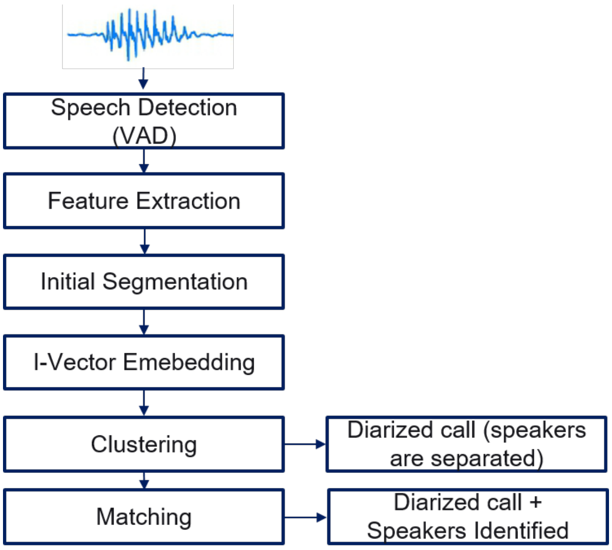

Figure 1: *End-to-end speaker separation and identification for a single conversation*

2.1. Segment Embedding

We use the Kaldi [3] SRE08 recipe for speech detection and i-vector embedding [4] of 1 second segments for voiced segments (using Kaldi VAD [5]). In our domain of sales calls and customer interaction 1 second segments appear a good choice as we want to capture short segments such as one sentence answers (*e.g.* ”yeah, that would work”).

Recent advances in embedding techniques offer a variety of methods for optimizing for speaker embedding. These include x-vectors [2], optimizing overlapping frames and using a siamese network setting of training with random pairs of vectors similar to [6]. Other approaches optimize using triplet loss, such as [7], or triplet loss with improved confuser choice in [8].

We recommend bootstrapping the training of the i-vector embedder using an out of domain model and then choosing reliable training data from clusters created for recordings in domain, after manual annotation marking clusters containing multiple speakers.

2.2. End-to-end analysis of a single recording

Each recording is broken into 1 second segments which are embedded into i-vectors of length 100.

The vectors are transformed using Linear Discriminant Analysis (LDA). Vectors are then clustered using agglomerative clustering (Sklearn implementation), the number of clusters is guessed using a heuristic based on the number of invited attendants on the calendar invite. This heuristic inflates the number of speakers, as we have found out that it is hard to recover from a mistake in this stage that combines two speakers into a single cluster. We experimented with replacing the agglomerative clustering with K-means, the results do not differ in a significant manner. We have also experimented with using *probabilisitc latent discriminant analysis* (PLDA) [9] as the distance metric for the agglomerative training, which also yielded no differences.

Clusters are then joined using PLDA distance combined with temporal information (proportion of their i-vectors which are consecutive) using a neighbor joining scheme. The resulting clusters are matched to the voice print database (described in the next section) using PLDA for identifying the known users. We assume each customer facing conversation includes a known user or users and an unknown user or users.

Clusters are further tagged as company-representative or customer using a text classifier. This classifier is trained automatically per company to identify its vernacular based on the voice-matched sections. If the recording source is a video conferencing tool the clusters are further enriched with information about the speaker based on the video indication of the speaker name.

2.3. Unsupervised Enroll

To enroll a new user the system chooses 3-5 candidate conversations recorded by the same user, the metadata of the conversation (calendar and CRM) is used to ascertain that the conversations do not share any other speakers to avoid enrolling the wrong user.

Clusters from all the candidate conversations are then compared using PLDA, and a similarity graph is created, adding an edge between clusters if the similarity is greater than a threshold. Connected components are then extracted and compared to a data base of common conference announcements and voice mail announcements. If a component includes clusters from multiple conversations, and doesn't match any known confounder it is used as the enroll voice print for that user.

After enrolling the user, that person's conversations are rediarized with clusters being compared to the new voice print. If the voice print doesn't match a speaker in enough conversations the process is automatically repeated with another set of conversations.

3. Results

The resulting system requires no end user effort, and results in high quality speaker separation, separating company representatives from customers and identifying the recording users. See figure 2 for example view of the output.

4. Conclusions

We have presented a tool which is used for analyzing conversations. It has been designed to be robust and requires no user effort. It combines multiple learning models combined in a framework to provide a full solution for speaker separation.

The data produced by the system allows further testing of hypotheses about the impact of different conversation cues to the outcome of conversations on large data sets, such as long monologues or entrainment.

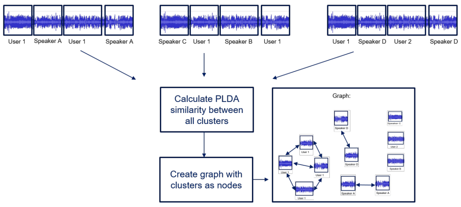

Figure 2: *Automatic enrolling based on recording participant metadata. The input in this example is three recordings sharing a single user in the data set, 1 other known user (user 2) and 3 distinct unknown speakers. First, each recording is clustered, yielding multiple clusters for each speaker. Secondly, clusters are compared using PLDA and a graph is created with edges between pairs of clusters which pass a similarity threshold. Connected components from multiple calls are then used as the voice print.*

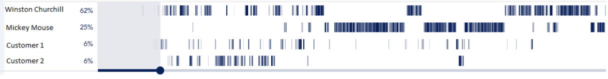

Figure 3: *Speaker separation view*

5. References

[1] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1695–1699.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *ICASSP, Calgary*, 2018.

[3] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[4] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[5] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[6] K. Chen and A. Salman, "Extracting speaker-specific information with a regularized siamese deep network," in *Advances in Neural Information Processing Systems*, 2011, pp. 298–306.

[7] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[8] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," *arXiv preprint arXiv:1710.10467*, 2017.

[9] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007.* IEEE Computer Society, 2007, pp. 1–8. [Online]. Available: https://doi.org/10.1109/ICCV.2007.4409052