Blind Speaker Clustering Using Phonetic and Spectral Features in Simulated and Realistic Police Interviews

Anil Alexander[1] *and Oscar Forth*[1]
[1]*Research and Development Oxford Wave Research Ltd, Oxford*
{anil|oscar}@oxfordwaveresearch.com

In this study, we present a novel approach to the blind automatic segmentation of speakers in police interviews, combining the use of phonetic features like pitch and the statistical pattern recognition of short-term power spectrum features like Mel Frequency Cepstral Coefficients (MFCCs). This approach requires minimal user intervention and allows for easy segmentation of the speech of separate speakers from multi-speaker recordings. This approach can have significant benefit in the harvesting of the speech of a single speaker for use in phonetic and automatic speaker recognition as well as gleaning quick intelligence in surveillance recordings. We propose a two-tiered approach to speaker segmentation, the first using discontinuities in the pitch trajectories to identify potential speaker clusters, and the second using an iterative speaker assignment and training method based on Gaussian mixture models. This approach will be demonstrated using realistic and simulated police witness interviews.

Proposed approach and test databases

The pitch tracks for the voiced segments are extracted from the interview recording using the autocorrelation-based pitch tracker in Praat (Boersma, 1993). Based on discontinuities in the pitch track, we extract 'zones of reliability' for the identity of a speaker. A continuous 'run' of similar values in the pitch track provides such a zone of reliability and any significant discontinuities in the pitch track, either in time or frequency, is used to define a candidate transition point between speakers. These candidate transition points are then used to define clusters as illustrated in Figure 1a. We use the clusters with sufficient information to model potential speakers. A statistical model of each cluster is then compared to all other segments in order to get the most divergent pair of segments.

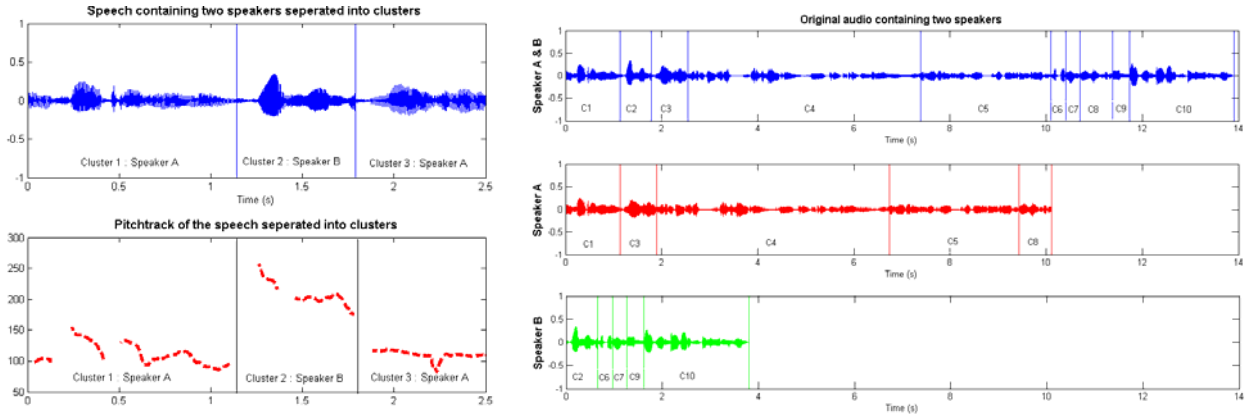

Figure 1a: Using discontinuities in the pitch track to define candidate speaker transitions and clusters of speech **1b:** The original audio with identified clusters and the results of blind speaker segmentation with clusters assigned to Speaker A and Speaker B

At this stage, the statistical model consists of lower complexity Gaussian mixture models as the

cluster sizes considered are generally between 0.5 and 1s. After the two most divergent cluster models are identified, they are compared against all the clusters in the recording and the first pass of label assignment as Speaker A and Speaker B is done based on the highest scoring model. The segments newly labelled as belonging to Speaker A and Speaker B are then used to create new models (of higher complexity) for Speakers A and B which are compared against all clusters (Narayanaswamy et al 2006). This process is repeatedly iteratively till convergence, where no new assignments take place. Thus, with little or no user interaction, the system can automatically identify and separate the two speakers in the recording (Figure 1b).

In order to test this approach we used two interview databases; one a simulated police interview from the DyVIS database (Nolan et al 2009), and a realistic police interview recorded in a police station in London. The first test database consisted of a subset of twenty speakers from the DyVIS database from the simulated police interviews in Task 1. This task was developed 'to elicit spontaneous speech in a situation of cognitive conflict' (Nolan et al 2009). Since the recordings were stereo recordings with each channel containing the speech of both speakers at different levels we mixed the two channels (50% from each) into a single channel waveform. Although the DyVIS database is recorded in relatively ideal circumstances with high quality microphones, in a noise-controlled environment it provides a good approximation of a police interview, albeit one of unusually high quality. The second database was recorded in a witness interview room at a police station in London, using a new digital recording system that was under evaluation. This recording set-up was considered as typical for witness recordings made at the London Metropolitan Police Service. There was a small amount of background noise, but this did not render the speech inaudible and was considered representative of real-world conditions. The database consisted of speech samples in two conversational styles, namely free speech and scripted speech consisting of simulated police interviews conducted by trained, serving police officers as interviewers. The data used in both the DyVIS and the police interview recordings were 16bit, 44,100 Hz uncompressed mono files in Microsoft WAV file format.

Results and conclusions

With both these test databases we observe that the blind speaker separation approach is able to accurately extract the speech of individual speakers with minimal mislabeled speaker assignments. This approach shows robustness to noise and works well even with the voices of speakers with close pitch ranges. The most challenging problem we have encountered is that of over-talking between speakers. This capability of being able to collect quantities of the speech of individual speakers from a multi-speaker conversation would not only be of use to automatic speaker identification systems and phonetic analysis but also in phonetic research in areas such as long-term formant analysis and vocal profiling.

References

P. Boersma (1993), Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, *in Proc. of the Institute of Phonetic Sciences*, University of Amsterdam, vol. 17, pp. 97 – 110.

B. Narayanaswamy, R. Gangadharaiah and R. Stern (2006) Voting for Two-Speaker Segmentation, *Proceedings of Interspeech, (ICSLP)*, pp. 2086–2089.

Nolan, F., McDougall, K., de Jong, G., Hudson, T. (2009) The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law 16(1)*, pp. 31-57.