Not a Lone Voice: Automatically Identifying Speakers in Multi-Speaker Recordings

Anil Alexander, Oscar Forth, Alankar Atreya, Samuel Kent, and Finnian Kelly
Research and Development, Oxford Wave Research Ltd., Oxford, U.K.
{anil|oscar|alankar|sam|finnian}@oxfordwaveresearch.com

Law enforcement audio recordings such as interviews, telephone intercepts and surveillance recordings often contain speech from more than one speaker. Identifying speakers of interest within these multi-speaker recordings first involves editing to extract the speech of a single speaker. This editing process, in which extraneous noises and other speakers are removed, can either be performed manually or assisted using speaker diarisation software. However, if a large number of such files need to be analysed in a short period of time, it may not be practical to involve a human in the loop. In this paper, we attempt to address the challenging task of efficiently and accurately spotting certain target speakers from large volumes of multi-speaker recordings automatically.

We have tried to address this problem using a simple but effective approach, in which short overlapping segments of the multi-speaker recording are extracted and modeled within an i-vector framework. The i-vector approach converts a recording into a fixed length, low-dimensional representation of the speaker's voice. The i-vectors for each overlapping segment (e.g. 10s segments, with 5s overlap) are compared with the i-vector for the target speaker file. The match scores obtained across all overlapping segments are first smoothed to reduce the effect of outliers, and then an average of the three maximum scoring segments provides a match score for the file.

We tested our approach with controlled laboratory data as well as real telephone intercept data. We used a multi-speaker modified version of the VOCALISE speaker recognition software (Alexander et al, 2014). For our experiments with laboratory data, we used interview and intercept recordings in same- and cross-channel conditions from the DyVIS database (Nolan et al, 2009). For 'single target' cross-channel comparisons, we used 51 files containing two speakers from the intercept task and compared them with 59 single speaker files from DyVIS Task 3 (report and report recall). For each multi-speaker recording, the majority (94.1%) of corresponding target speakers were identified at rank one or two of the match score list (Figure 1). The equal error rate (EER) of this comparison was 3.90%. For uncontrolled real telephone intercept data, we have worked with a subset of the FRITS database (van der Vloed et al, 2014). All tests were conducted by and at the Netherlands Forensic Institute (NFI). This subset consisted of 11 multi-speaker conversations (mostly two, and in some cases, more speakers) and a set of 32 target speakers. For each multi-speaker recording the majority of corresponding target speakers were identified at rank one or two of the match score list (76.1%) (Figure 1). Conversely, for each target, a matching multi-speaker file containing that speaker was identified at rank one or two, 80% percent of the time.

We observe that the total duration of speech and the relative speaker mix for each target in a multi-speaker file are important for accurate recognition. Despite these challenges, this approach shows promise for automatically processing large volumes of real-world multi-speaker files.

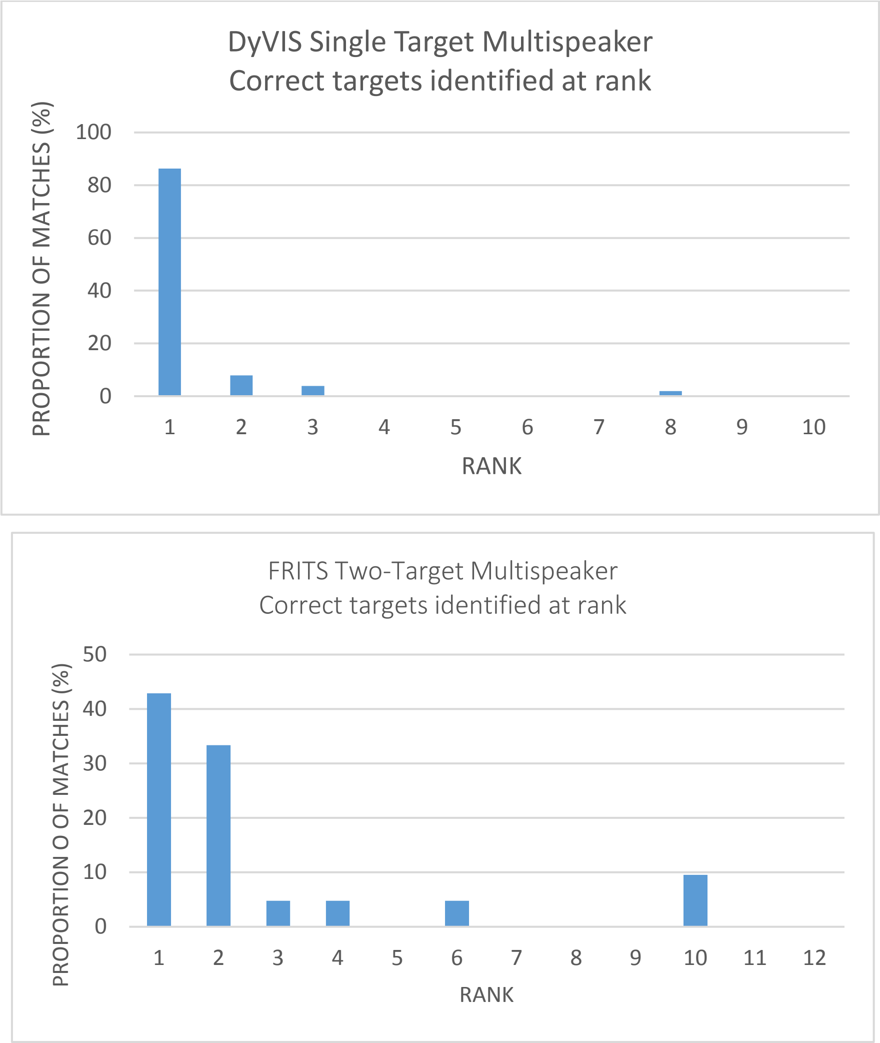

Figure 1 The proportion of correct targets identified at a certain rank for single-target DyVIS database and two-target FRITS database.

References

A. Alexander, O. Forth, A. A. Atreya, and F. Kelly (2016). "VOCALISE: A forensic automatic speaker recognition system supporting spectral, phonetic, and user-provided features", *Odyssey 2016 Speaker and Language Recognition Workshop*, Bilbao, Spain, 2016.

D. L. Van der Vloed, J. S.Bouten, and D.A. Van Leeuwen (2014). NFI-FRITS: A forensic speaker recognition database and some first experiments, *Proceedings of Odyssey Speaker and Language Recognition Workshop 2014*, Joensuu, Finland, pp. 6-13, 2014.

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

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