Speaker Tracking for Hearing Aids

Author: Joachim Thiemann, Jörg Lücke, Steven van de Par

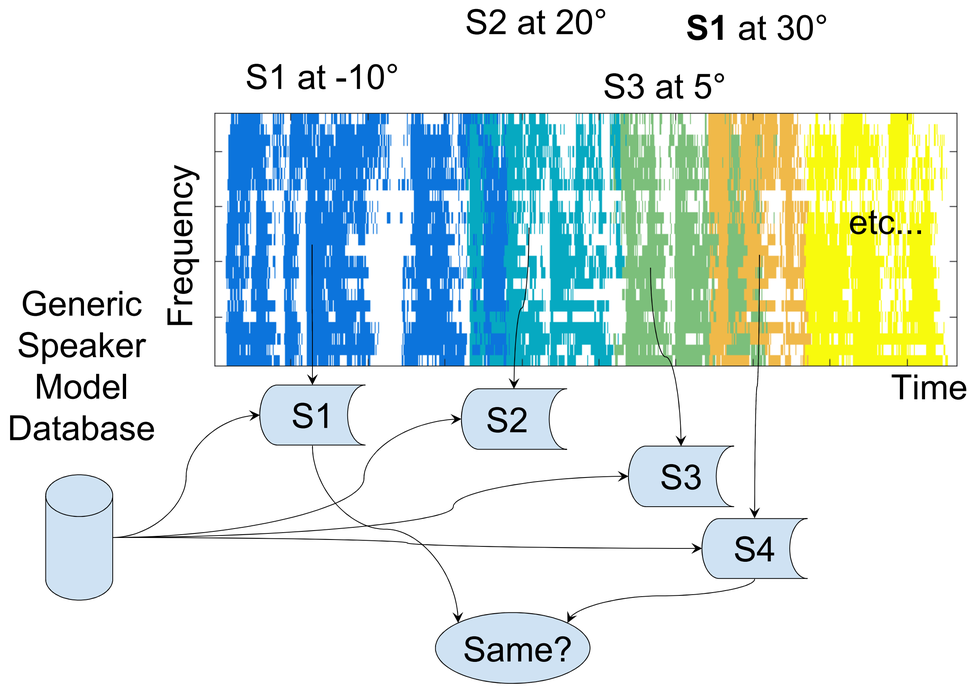
Introduction

Speaker tracking attempts to link speech segments spoken by one individual over pauses and interruptions. In the context of hearing aids, the speakers' positions (relative to the hearing aid user) may have shifted.

This is effectively a speaker recognition or diarization problem with some additional challenges because we have:

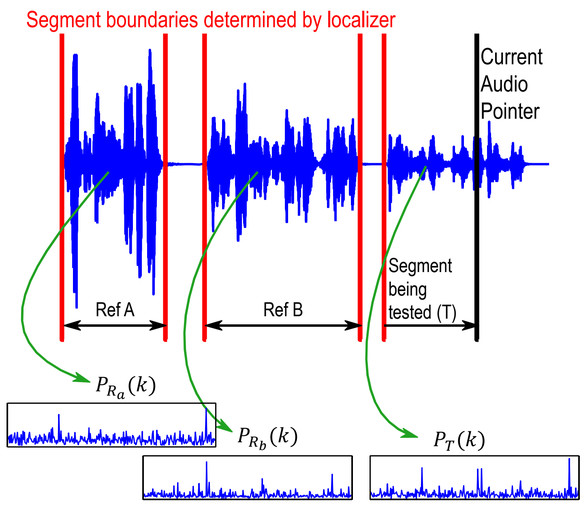
- no a priori knowledge of the speaker (no pre-trained speaker models)
- constantly changing acoustic conditions
- Potentially very short speech fragments
- limited computational resources
- a need for fast, real-time results

Text-independent speaker recognition, that is, recognizing a speaker without constraints on the content of the speech, is a well-established field [Kinnunen, 2010]. However, the typical application is to recognize a speaker that the system was trained on beforehand. Speaker diarization [Anguera, 2012] is the process of segmenting a stream of audio (e.g. a dialogue) such that all segments being spoken by one individual can be grouped together. Usually, this is done off-line, with the entire conversation being available to the algorithm.



System Description

Our proposed algorithm assumes that the speech signal is segmented, that is, the times when there is a change of speaker are known. In the context of hearing aids, this would be inferred from the sound source localization [May, 2011].



Speech is processed in short time frames (20 ms), with a frame advance of 10 ms. Each frame of speech is converted into Mel-Frequency Cepstral Coefficients (MFCCs), which are the feature vectors.

We compute the a posteriori probability of each observation matching speaker models in a generic speaker database, which are represented by Gaussian mixture models (GMMs). The mixture models are deliberately simple, using few Gaussians per GMM. The models are trained by rendering the training speech in many different acoustic conditions. **The set of probabilities that a given speaker matches the speakers in the database forms a "fingerprint"** P.

$$P_{\mathrm{k}}^{R} = \sum_{\mathrm{m}\in R} \mathrm{P}(\mathbf{y}[\mathrm{m}] \mid \mathrm{S_k}), P_{\mathrm{k}}^{T} = \sum_{\mathrm{m}\in T} \mathrm{P}(\mathbf{y}[\mathrm{m}] \mid \mathrm{S_k}),$$

where R and T represent the set of time frames of the reference and the test segment respectively.
The difference between two fingerprints P^R (the reference) and P^T (test) is computed using the normalized correlation coefficient,

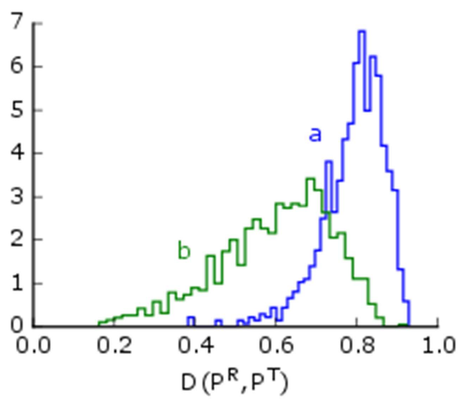
$$D(P^R, P^T) = \frac{\sum_k P_k^R P_k^T}{\|P^R\|\|P^T\|}$$

Any previously seen speech segment can be used as the reference. The fingerprint can be updated with little computational effort as new frames are recorded.

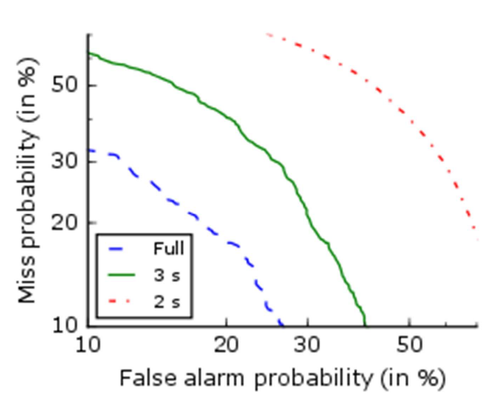
Results

Evaluation was performed by creating a generic speaker database of 326 male speakers from the training set if the TIMIT database. Using speech from the test set of the same database (no overlap of speakers), sentences were rendered in an office environment at different locations.

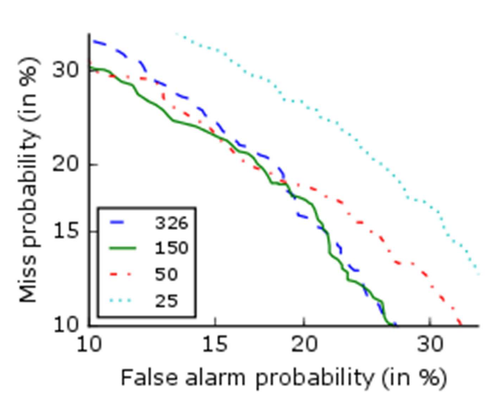
A threshold can then be selected to make a trade-off between the miss probability (same speaker mistaken for different speaker) and the false alarm probability (differing speaker mistaken for reference speaker)



Histograms of $D(P^R, P^T)$ for speech items spoken by differing speakers (a) and the same speaker (b)



Detection error trade-off (DET) showing the effect of truncating the test sentence. As more speech becomes available, the performance increases.



DET showing the effect of reducing the number of speakers in the generic model database.

Conclusion

Speaker tracking is useful for intelligent hearing aids to ensure that speakers of interest to the hearing aid user are not supressed accidentally. In this work, we present an algorithm that uses a generic speaker database to obtain fingerprints of speakers. This method is robust to the speech content and varying conditions, yet is also low in computational complexity, an important cosideration for hearing aids.

References

[Kinnunen, 2010] T. Kinnunen and H. Li, „An overview of text-independent speaker recognition: From features to supervectors," Speech Communication, vol. 52, no. 1, pp. 12-40, 2010

[Anguera, 2012] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland and O. Vinyals, „Speaker Diarization: A Review of Recent Research," IEEE Trans. Audio, Speech, and Signal Processing, vol. 20, no. 2, 2012

[May, 2011] T. May, S. Van de Par, and A. Kohlrausch, „A probabilistic model for robust localization based on a binaural auditory front-end," IEEE Trans. Audio, Speech, and Signal Processing, vol. 19, no. 1, pp. 1-13, 2011