Pitch Features for Speaker Tracking

Joachim Thiemann and Steven van de Par
CvO Universität Oldenburg, AG Akustik, Cluster of Excellence 'Hearing4all'
26129 Oldenburg, Germany, Email: joachim.thiemann@uni-oldenburg.de

Introduction

Modern multichannel hearing aids (HAs) are capable of using sophisticated signal processing algorithms to enhance acoustic signals for the wearer. One common technique is to use directional filtering (eg. beamforming), that is, the sound from some particular direction is enhanced, whereas sounds from other directions are suppressed.

Such directional filters typically point in a fixed direction to the front of the wearer. However, as the processing power in hearing aids increases, more sophisticated algorithms are being investigated, where the acoustic scene is intelligently evaluated in order to identify all sources the user might be interested in; this could be the case e.g. where the wearer is in a crowd of people. Normal hearing persons can easily deal with this kind of scenario, but it can be challenging for hearing impaired persons. An intelligent hearing aid should be able to mimic this ability for the wearer by analysing the acoustic scene (a process termed Computational Auditory Scene Analysis, CASA), enhancing some sources (speakers of interest) while suppressing others (eg. background babble).

This contribution details research into the aspect of speaker tracking, in this context referring to a form of short-time speaker identification. In particular, the goal is to determine if a speaker observed at a previous instance reappears in the acoustic scene at a different spatial location.

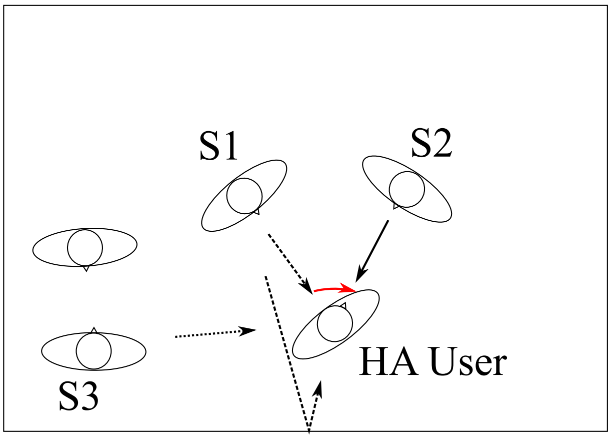

Figure 1: Schematic diagram of a changing acoustic scene. The HA user first attends speaker S1, then turns (red arrow) to pay attention to speaker S2. New speech originating from S1 now originates from a different direction relative to the HA user's head, but should not get suppressed. Speech from interfering speaker S3 should also be suppressed, before and after the head movement.

A sample scenario is depicted in Fig. 1. The HA user faces speaker S1, thus speech arrives from the front. The user then turns to face speaker S2. When speaker S1 resumes speaking the relative direction of S1 has changed since the HA user is still facing S2. A hearing aid should decide that since S1 was being attended to before, the speech coming from the direction of S1 should not be suppressed.

From the perspective of the signal analysis in the HA, this problem is challenging for several reasons. As a speaker recognition problem, the target speaker is in general unknown to the system prior to the current conversation, thus no speaker-specific models can be stored; models must be stored or adapted in real-time. The amount of speech observed from a new or recurring speaker is very limited (in the order of seconds), which is typically insufficient to adapt generic models. Finally, the computational complexity and memory footprint must be kept low since HAs are small devices with limited battery power.

In [1], we proposed a system that could track a speaker in given about 2 or 3 seconds of speech (20 dB SNR), using mel-frequency cepstral coefficients (MFCC) as main features, showing the potential of the proposed "fingerprint" approach. In this paper, parallel to [2], we present an examination of a pitch-based speech feature (based on work in [3]) for improved speaker tracking, which can distinguish speakers with less observed speech and is more robust to noise.

Background

The speaker tracking method proposed in [1] is based on the concept of "fingerprints," that is, a small amount of data extracted from the observed data that will identify the speaker while stripping away information about the speech content and acoustic path. The fingerprint can be extracted from any amount of speech, though the more speech is used to generate the fingerprint, the better it should represent the individual speaker.

We begin by training (offline) a database of K generic speaker models, $\mathcal{S}_k, k = 1, ..., K$, and assume that given speaker in the scene is equally likely to match any speaker model $\mathcal{S}_k$ in the database. Given a frame of speech signal $\mathbf{y}[m]$, where m is the time index, for each speaker model $\mathcal{S}_k$ and each m in which speech is detected, we compute the *a posteriori* probability

$$P(\mathcal{S}_k \,|\, \mathbf{y}[m]) = \frac{P(\mathbf{y}[m] \,|\, \mathcal{S}_k) P(\mathcal{S}_k)}{\sum_{k'} P(\mathbf{y}[m] \,|\, \mathcal{S}_{k'}) P(\mathcal{S}_{k'})} \quad (1)$$

$$\sim \frac{1}{Z_m} P(\mathbf{y}[m] \,|\, \mathcal{S}_k), \quad (2)$$

where Z_m is the normalization factor for the particular observation (time frame) m. In [1], these probabilities are computed by using Gaussian mixture models for $\mathcal{S}_k$, whose *a posteriori* probability can be computed efficiently and integrated over multiple frames.

Assuming we now have a segment of speech $\mathbf{y}[m], m \in$ Ref from one speaker, and we designate this the reference speaker. We compute a reference pattern $\mathcal{P}^{\mathrm{Ref}}$ by summing the *a posteriori* probabilities $P(\mathcal{S}_k \,|\, \mathbf{y}[m])$ for each speaker model k over time, with

$$\mathcal{P}_k^{\mathrm{Ref}} = \sum_{m \in \mathrm{Ref}} P(\mathcal{S}_k \,|\, \mathbf{y}[m]) = \sum_{m \in \mathrm{Ref}} \frac{1}{Z_m} P(\mathbf{y}[m] \,|\, \mathcal{S}_k), \tag{3}$$

where "Ref" indicates the time frames of the reference speaker speech segment. Thus $\mathcal{P}^{\mathrm{Ref}} = (\mathcal{P}_1^{\mathrm{Ref}}, \ldots, \mathcal{P}_K^{\mathrm{Ref}})^T$ is a K-dimensional vector that can be regarded as a "fingerprint" of the reference speaker. Note that as more speech segments (from potentially other speakers) are seen by the system, they can be stored as multiple reference patterns $(\mathcal{P}^{\mathrm{Ref}_A}, \mathcal{P}^{\mathrm{Ref}_B}, \ldots)$.

Now assume that new speech is observed by the system and we try to determine is the newly observed speech is from the reference speaker or a different speaker. As in (3), the fingerprint of the current speech segment can be obtained by summing the *a posteriori* probabilities of the speech frames being from the generic speaker models,

$$\mathcal{P}_k^{\mathrm{current}} = \sum_{m=\mathrm{B}}^{t} \frac{1}{Z_m} P(\mathbf{y}[m] \,|\, \mathcal{S}_k), \tag{4}$$

where B is the frame index of the beginning of the current speech segment and t is the current time frame index. An example of fingerprinting speech segments is shown in Fig. 2.

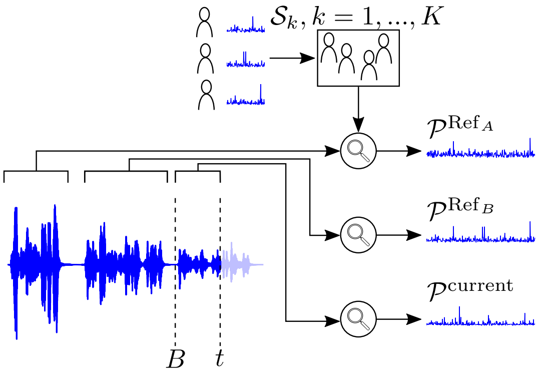
$\mathcal{S}_k, k = 1, ..., K$
$\mathcal{P}^{\mathrm{Ref}_A}$
$\mathcal{P}^{\mathrm{Ref}_B}$
$\mathcal{P}^{\mathrm{current}}$
B
t

Figure 2: Tracking speakers by fingerprinting segments. The first segment results in fingerprint $\mathcal{P}^{\mathrm{Ref}_A}$, the second in $\mathcal{P}^{\mathrm{Ref}_B}$. The current speech segment beginning at time B until the current frame t yields fingerprint $\mathcal{P}^{\mathrm{current}}$.

Fingerprints are compared to each other using the normalized correlation,

$$D(\mathcal{P}^R, \mathcal{P}^T) = \frac{\sum_k \mathcal{P}_k^R \mathcal{P}_k^T}{\|\mathcal{P}^R\| \|\mathcal{P}^T\|}, \tag{5}$$

which can be compared to a threshold τ. A correlation above this threshold indicates that the current speech segment is assumed to originate from the reference speaker. The choice of τ controls a trade-off: a low τ leads to more "false accepts," (FA) that is, other speakers mistaken for the reference, whereas a high τ means that "false rejects" (FR) become more likely, where one speaker is mistaken for many individuals.

Pitch-based features

As reported in [1], MFCCs allowed for a FA/FR rate of about 30% with 3s of speech in clean conditions. While this demonstrated the feasibility of the proposed approach, that performance is deemed insufficient for a real system, and the main course for improving performance we investigate is the use of pitch features. In particular, as the goal is to integrate speaker tracking with a a more comprehensive CASA system [2], we focus on the "periodicity degree" (PD) feature as proposed in [3], which was developed for speech enhancement, but in [2] is used for glimpse segregation. This feature has shown itself to be very robust to noise, and below we show that sufficiently speaker-specific features can be extracted from PD to improve speaker tracking performance.

To compute PD, the signal is analysed by a perceptually motivated complex-valued gammatone filterbank with subband filter center frequencies spaced on a nonuniform ("ERB") scale. As with MFCCs, the signal is processed in discrete time frames, and the division of the signal into time and frequency gives a set of time-frequency (t-f) bins, which we denote by $\mathbf{y}(f, m) = y(f, m, n), n = 1, \ldots, N$, where f is the subband index, m the time index as above, and n is the sample index within each t-f unit. The PD is computed in each t-f bin by combining two methods, the normalized autocorrelation (NAC) and the comb filtering ratio (CFR). In the lower-frequency subbands (F_L) we use $y(f, m, n)$ directly, while in the subbands with centre frequencies > 1.5 kHz (F_H) we use the envelope $y_E(f, m, n) = |y(f, m, n)|$. Using

$$\mathrm{NAC}(\cdot, p) = \begin{cases} \frac{\sum_{n=0}^{N-1-p} [y(\cdot,n) y(\cdot,n+p)]}{\sqrt{\sum_{n=0}^{N-1-p} y(\cdot,n)^2} \sqrt{\sum_{n=0}^{N-1-p} y(\cdot,n+p)^2}}, \\ \hfill f \in F_L \\ \frac{\sum_{n=0}^{N-1-p} [y_E(\cdot,n) y_E(\cdot,n+p)]}{\sqrt{\sum_{n=0}^{N-1-p} y_E(\cdot,n)^2} \sqrt{\sum_{n=0}^{N-1-p} y_E(\cdot,n+p)^2}}, \\ \hfill f \in F_H \end{cases} \tag{6}$$

and

$$\mathrm{CFR}(\cdot, p) = \begin{cases} \frac{\sum_{n=0}^{N-1-p} [y(\cdot,n) + y(\cdot,n+p)]^2}{\sum_{n=0}^{N-1-p} [y(\cdot,n) - y(\cdot,n+p)]^2}, & f \in F_L \\ \frac{\sum_{n=0}^{N-1-p} [y_E(\cdot,n) + y_E(\cdot,n+p)]^2}{\sum_{n=0}^{N-1-p} [y_E(\cdot,n) - y_E(\cdot,n+p)]^2}, & f \in F_H \end{cases}, \tag{7}$$

where $\cdot$ is substituted for the f, m pair, and p is the pitch candidate index. From the above, PD is computed as the product of the NAC and CFR, where

$$\mathrm{PD}(f, m, p) = \max[0.01, \mathrm{NAC}(f, m, p) \cdot \mathrm{CFR}(f, m, p)]. \tag{8}$$

For each time frame we sum the PD vector over frequency, $\mathrm{PD}_A(m, p) = \sum_f \mathrm{PD}(f, m, p)$, from which we obtain the *pitch candidate* $P_0(m) = \arg_p \max \mathrm{PD}_A(m, p)$.

Experiment

To evaluate the new feature, the setup as used in [1] was reused with modifications to accommodate the modified front-end. The sampling rate of the system is 16 kHz, and a frame length of 20 ms is used with 50% overlap for a frame rate of 10 ms. We compute ERB frequency cepstral coefficients,

$$\mathrm{EFCC}(m,k) = \sum_{f=1}^{F} \log \|\mathbf{y}(f,m)\|^2 \cos\left[k\left(f-\frac{1}{2}\right)\frac{\pi}{F}\right], \tag{9}$$

which are functionally equivalent to MFCCs but use the subband decomposition from the filterbank described above.

All speech samples are taken from the TIMIT [4] database, but we restrict ourselves to male speakers only to remove any bias due to the unequal size of the number of male and female speakers in the database. The generic speaker models $\mathcal{S}_k$ are trained using the 326 male speakers from the training set of TIMIT; using a set of binaural room impulse responses (BRIR) taken from [5], all speech from these speakers is rendered at several positions relative to the head, and all feature vectors from this data is used to train a GMM with 4 Gaussians for each speaker.

To test tracking performance, 1000 sentence triplets are created with the structure A-B-A, that is, a sentence from speaker A, followed by a sentence from speaker B, followed by a sentence from speaker A again. All three sentences are rendered at different locations (which includes head rotation of the receiver), also using the BRIRs from [5]. For each sentence triplet, speaker A and B are chosen randomly from the test portion of TIMIT. Spatially uncorrelated speech-shaped noise is added to the signal at fixed SNR.

Results

We show experimental results using detection error tradeoff (DET) curves, that show the achievable performance for various settings of τ.

Figure 3 shows the performance with 20 dB SNR, similar to the experiment in [1]. The panels show the performance of different feature sets: panel A using 5 selected EFCCs (1, 2, 3, 6 and 8), panel B those same EFCCs plus P_0. Panel C is the performance using the full EFCC set ($k = 1, \ldots, 12$), without P_0, and panel D the performance using the full feature set. The different lines on each plot show the performance depending on how much of the test speech is used to make the reference/not-reference decision.

The plots show that the performance improves significantly with the inclusion of the P_0 feature, with a 6-dimensional feature vector slightly exceeding the performance of using the 12-dimensional feature vector of the full EFCC set. Best performance is gained by using the full EFCC set plus P_0. We also see that using more than 2 s of speech does not gain much more in performance, no matter what features are used.

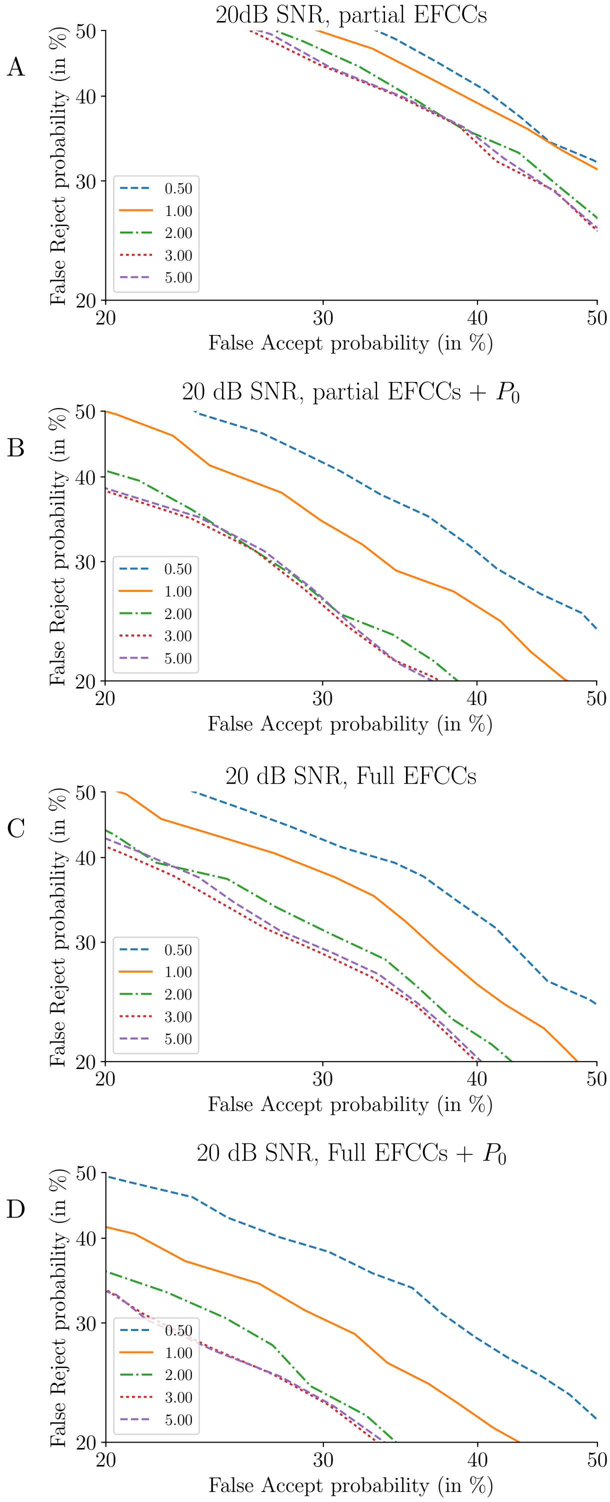

Figure 3: Detection Error Tradeoff curves, at 20 dB SNR. Legends show the amount of speech of the test segment observed. Panel A shows the DET using the 5 most salient EFCCs, panel B the same plus the P_0 feature. Panel C shows the DET using all EFCCs but without P_0, and panel D using all EFCCs and P_0.

Figure 4 shows the same experiment as above, but this time at a SNR of 10 dB. We note that compared to Fig. 3, all curves are shifted to the right indicating degraded performance as expected; however, the inclusion of P_0 makes

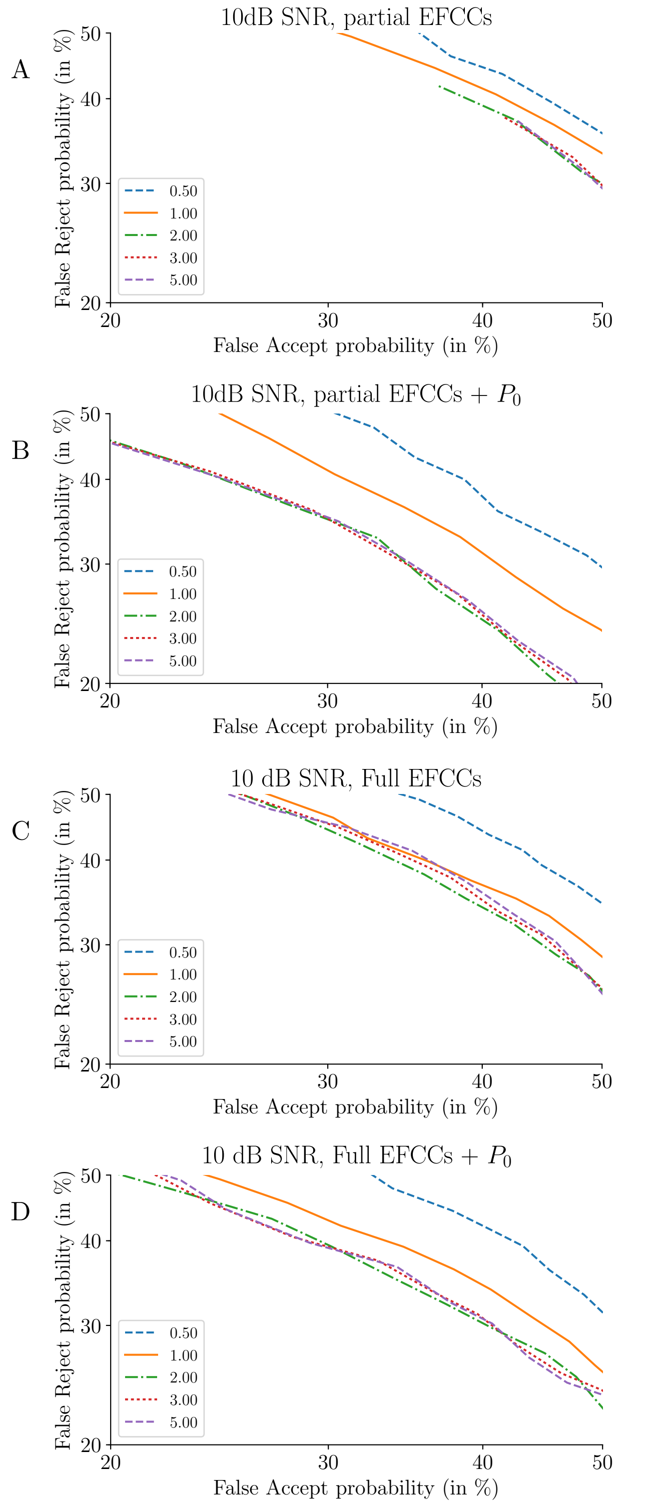

Figure 4: Detection Error Tradeoff curves as in Fig. 3, at 10 dB SNR.

the speaker tracking far more robust to noise. Interestingly, the full EFCC set plus P_0 seems to be marginally worse than the partial EFCC set plus P_0; probably the additional EFCC features add too much noise at this SNR. Against expectations, the performance still reaches a plateau after 2 s of speech have been observed.

Discussion

In this article, we examine the use of a pitch feature for tracking speakers in a dynamic acoustic scene as received by a hearing aid. The pitch feature we investigate is available to the hearing aid as part of the initial signal analysis and for speech enhancement; however we expect that it is also useful for recognizing individual speakers within small groups over short time intervals.

The assumption can be made that speakers have an individual pitch range that is being used in normal conversational speech. However, speech pitch and pitch dynamics can vary due to many factors, and thus relying on pitch as only feature would probably not be reliable, and should be always be used in conjunction with more traditional speaker recognition features (eg. the MFCC-like features used here).

Overall, a reliable and robust pitch extraction method is a useful addition to speech-oriented CASA algorithms. Further research could look into modulating the weighting of pitch features, since the speaker-identifying ability of pitch may vary depending on the phoneme being spoken.

References

[1] J. Thiemann, J. Lücke, and S. van de Par, “Speaker tracking for hearing aids,” in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, September 2016.

[2] S. Sutojo, S. van de Par, and J. Thiemann, “Combination of monaural and binaural auditory cues for source segregation,” in *Proc. DAGA 2017*, Mar. 2017.

[3] Z. Chen and V. Hohmann, “Online monaural speech enhancement based on periodicity analysis and a priori snr estimation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1904–1916, Nov 2015.

[4] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, “Timit acoustic-phonetic continous speech corpus,” National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep. NISTIR 4930, 1993.

[5] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, “Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses,” *EURASIP J. on Applied Sig. Proc.*, 2009.