SPEAKER TURN TRACKING WITH MOBILE MICROPHONES: COMBINING LOCATION AND PITCH INFORMATION.

Heidi Christensen, Jon Barker

Department of Computer Science, University of Sheffield, United Kingdom
h.christensen@dcs.shef.ac.uk, j.barker@dcs.shef.ac.uk

ABSTRACT

This paper considers the problem of using binaural microphones to track speakers in a situation where the microphones are themselves in motion (i.e. due to listener head movement). We present a general framework that applies particle filtering to combine sequential interaural time difference (ITD) cues with noisy sensor motion data. The framework is demonstrated in a meeting scenario applied to a moving-listener version of a speaker-diarization task. The paper extends previous work by investigating two potentially complementary ways of exploiting pitch track estimates in this framework, either, i) informing the time points at which speaker turn changes may occur, or ii) improving the ITD estimates by allowing integration over spectro-temporal regions grouped by pitch. Experiments using real meeting scenario recordings, made with in-ear binaural microphones, show that the latter approach leads to large and significant reductions in diarization error rate.

Index Terms: speaker change tracking, binaural hearing, pitch extraction, particle filtering, active listening

1. INTRODUCTION

Acoustic signals provide one of the simplest and most reliable means for localising and tracking moving objects. Audio-based tracking systems using arrays of two or more microphones are being researched within a wide range of application scenarios, including intelligent meeting rooms and smart houses (see e.g. [1, 2]). However, in the vast majority of cases algorithms and methods are developed with the underlying assumption that the microphones are located in fixed positions. This assumption is a particularly attractive simplification in sound source tracking applications where any microphone motion introduces added complexity and ambiguity in the cues. However, enforcing and relying on stationary sensors assumptions introduces a constraint that makes the technology unsuitable in many situations (e.g. wearable listening devices, hearing robots, vehicle sensors).

Accepting that acoustic sensors may move, significantly increases the difficulty of the sound source tracking problem. First, the quasi-stationary assumptions that are used in window-based extraction of source location cues (i.e., interaural time and level differences) are not compatible with rapid head rotations. Head rotation can approach speeds of up to 500 degrees/sec, equivalent to 5 degrees per 10 ms analysis window [3]. Rapid rotation thus results in significant 'motion blurring' of location estimates. Second, head motion introduces extra ambiguity [3]. For example, if a single source is dominating the acoustic scene, then a *leftwards* head movement may be hard to distinguish from a *rightwards* movement of the source, and vice versa. Note, in biological systems this second problem may be countered by complementary sensory input from other modalities such as vision or proprioceptive feedback.

The human auditory system has clearly developed solutions to the moving sensor problem. Take for example the ease with which we can interpret a complex acoustic scene such as a busy street without having to stop and listen: in general, your ears will be subject to constant movement confounding the tracking of the absolute position of external sound sources. The emergence of mobile hearing applications, such as perceptual robotics and wearable listening devices, lends urgency to the development of machine listening technology with comparable 'on-the-go' capabilities

In this paper we present initial steps towards mobile machine listening: in particular, we consider the additional complexity that microphone motion introduces to a problem that has been well studied from a stationary microphone perspective – the problem of using auditory localization cues to track speaker changes in a meeting (i.e. diarization). We reconsider this problem from the perspective of a meeting participant (human or robotic!) making natural head movements and propose a model which operates by simultaneously modelling changes in both the state of the external environment and of the listener.

Solving the moving-listener diarization tasks require explaining changes in the observed, binaural localisation cues (which provide information relative to the listener's head position) by 'decoding' them in terms of changes in the listeners's head position and changes in the currently active speaker. This is illustrated in Figure 1. From the acoustic signal we are extracting localisation cues, θ^O, that indicate the spatial angle of a sound source *relative* to the rotational angle of the listeners's head. This perceived angle is the difference between the absolute angle of the listener's head, θ^H (i.e. the angle relative to fixed room axis) and the absolute spatial angle of the active sound source θ^S_{cur}. It is these underlying angles, θ^H and θ^S_{cur}, that we wish to track in order to recover a full description of the scenario.

In [4] we presented our initial mobile speaker turn tracking system based on a particle filtering formulation using binaural localisation cues. This paper describes further improvements to the system through the introduction of *pitch-track* information. Through pitch tracking we can identify spectro-temporal regions (*fragments*) which are likely to come from the same sound source. This information can be of potential use in several ways. First, periods spanned by a single fragment are unlikely to contain speaker changes, i.e. speaker changes will be commonly indicated by a break in voicing or a pitch track discontinuity. Second, in [5] we showed how integrating localisation cues across pitch-based fragment regions improved the accuracy of the localisation cues as well

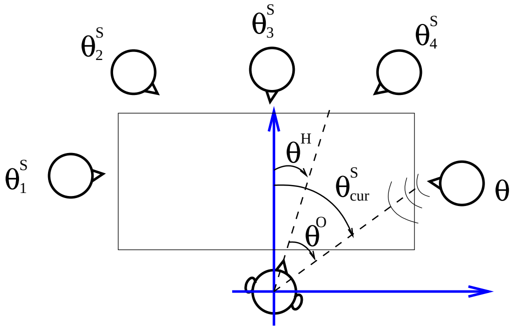


Figure 1: *Tracking the perceived direction of sound for a typical meeting scenario and from a moving perspective. Tracking the observed spatial angle (θ^O) composes into simultaneously tracking the absolute angle of the head (θ^H) and the absolute angle of the active sound source (θ^S).*

as their robustness to reverberation. In this paper we set out to investigate whether these potential benefits can be realised in practise and whether they can be usefully combined.

Section 2 presents both a general statement of the problem, and a description of the particular speaker turn-taking scenario on which we have evaluated our systems. Section 3 describes our particle filtering implementation of a sequential Bayesian approach to the solution. Results and conclusions follow in Sections 4 and 5.

2. THE SOUND SOURCE TRACKING PROBLEM

2.1 The general problem

The general approach to the tracking problem can be described as follows: We assume that we observe the acoustic mixtures arriving at a pair of microphones set in a binaural configuration. The microphones are fixed to a head that can in general move with 6 degrees of freedom (translation and rotation). The environment contains a number of potentially moving sound sources which may also switch between being active or inactive. The listener and sound source position parameters can be described by a state space that evolves over time. Our belief about the state space is informed by localisation cues extracted from the microphone data and potential self-position information originating from other modalities. We are particularly interested in inferring the sound source position parameters, but may also be interested in the listener's position.

2.2 The turn-taking meeting scenario

For this initial work we have concentrated on a constrained case of the general tracking problem: tracking speaker turns in a meeting scenario. Data from the CAVA database has been used [6]. This data was recorded from the perspective of a moving 'listener' in a conversational situation with five speakers. The 'listener' was fitted with a pair of binaural in-ear microphones and was also wearing a helmet-mounted tracking device so that the true head position could be recorded. Using the head tracker information allows us to model systems with access to self-position information.

We have focused on a particular session from the CAVA database – Panel Meeting 1 (P1). Here the human listener and 5 'actors' are sitting around a table[1]. The listener is blind folded and the actors take it in turn to speak. Listener head movements have been induced by giving the listener the task of monitoring speaker changes and always turning to face the current speaker. The task for our system will be to use the ITD cues in the binaural recording to estimate which of the five speakers is active at any instant. (Note, although the listener turns towards the active speaker, we do not make use of this information in solving the task – our solution is designed to work with arbitrary head motion).

2.3 Modelling the turn-taking meeting scenario

The meeting scenario was chosen for this initial study because it allows us to reduce the complexity of the general model described in Section 2.1. We will model the scenario with three main assumptions: i) that there are a fixed and known number of speakers seated at fixed, known positions and making only small scale movements around this position, ii) that the listener's head movement is mainly head rotation in the horizontal plane, i.e. from -90° to $+90^\circ$ azimuths, and iii) that one and only one person is speaking at a time.

Given the above assumptions, the CAVA meeting scenario can be described by a relatively simple state space (see Figure 1), modelled as

$$\alpha \triangleq (\theta^H, \theta_1^S, \dots \theta_K^S, cur), \tag{1}$$

where θ^H is the absolute spatial angle (azimuth) of the head, θ_k^S is the absolute azimuth of speaker k, K is the total number of speakers, and $cur \in \{1, \dots, K\}$, indicates which speaker is speaking. This model allows for a fully dynamic setup, where the listener's head can be turning, and where each sound source can be moving around independently. Following our assumptions, θ_k^S will be constrained to vary within a small range of a known initial position, $\theta_k^{S'}$.

3. A PARTICLE FILTERING SOLUTION

The task of tracking the state of the meeting scenario lends itself to a sequential Bayesian filtering approach, and in particular to a particle filtering implementation (see [7] for a tutorial and see Vermaak and Blake [8] for application of particle filtering to source tracking with a *static* listener.) In such approaches, estimates of the system state (Eq. 1) are updated at each time step by combining the previous state estimates with new information learnt from the incoming set of observations. The update is governed by two statistical models: a *system model* which describes our prior belief about how the system state evolves through time; a *measurement model* which describes our belief about the observations we are likely to make given the state of the system. In our case, the observations are of two types, i) interaural time difference (ITD) estimates extracted from the microphone signals, and ii) potentially noisy self-position estimates.

Section 3.1 describes the ITD observations and explains how they may be enhanced through the use of pitch information. Section 3.2 describes the system model and how it can be informed by pitch track information. Section 3.3 describes the measurement model which remains essentially the same as in our previous work [4] but is included here for the sake of completeness.

[1] The distances between the listener and the speakers are around 90 cm. The room is a typical large with $T_{60} = 300$ ms.

3.1 Observations

Both the pitch and localisation cues are extracted from an auditory front-end simulating the cochlear frequency analysis of the human ear. The model is implemented using a filterbank consisting of 64 overlapping bandpass gammatone filters, with centre frequencies spaced uniformly on the equivalent rectangular bandwidth (ERB) scale [9] between 50 Hz and 8000 Hz. The output of the filterbank is used to generate *cross*-correlograms on lags corresponding to the range -90° to $+90^\circ$ azimuth and *auto*-correlograms corresponding to a pitch period of up to 15 ms.

The pitch-based fragments are generated from a signal produced by averaging the left and right ear signals. After averaging, the fragment generation procedure follows that of the system designed for monaural signals presented by Ma *et al.* [10]. Briefly, from analysis of the auto-correlation delay patterns, multiple local pitch estimates are computed, and a simple rule-based tracker is used to form potentially overlapping pitch track segments that extend through time. Each pitch track is then used to recruit a spectro-temporal fragment (see Ma *et al.* for details).

The standard procedure of estimating ITDs (e.g. Jeffress' model [11]; and more recently [12, 13]) is to identify one or more peaks in the summary cross-correlogram (i.e. the cross-correlogram summed over frequency channels). However, the data are often very noisy and spurious peaks may arise due to reverberation in the room or competing sound sources. Figure 2 illustrates what the summary cross-correlogram looks like for a 20 second portion of the P1 CAVA session. The underlying 'track' of ITDs is plotted below the image. The sweeps arising from when the listener's head is turning towards a new speaker are clear. However, it is also evident that the data is challenging and that the largest peak in each frame would not always capture the active speaker location.

Two strategies are employed to handle the summary cross-correlogram noise. First, following [4], observations are obtained by extracting the lags corresponding to the *three largest peaks* for each frame rather than just the largest. The measurement model (Section 3.3) then accounts for the fact that two of these peaks are due to noise. Second, computing a summary cross-correlogram by summing the cross-correlogram across *time-frequency* fragment regions – rather than just across frequency – significantly reduces the degree of noise [5]. So, when no fragments are present, the peaks are extracted from the standard cross-correlogram integrated over all 64 frequency channels, but when a fragment *is* present we extract peaks from a summary computed across both the frame and the fragment.

3.2 System model

The system model determines how the state is progressed at each time step: $\alpha_t \rightarrow \alpha_{t+1}$, i.e. a head angle model ($\theta^H_t \rightarrow \theta^H_{t+1}$) and a speaker change model ($\theta^S_{k,t} \rightarrow \theta^S_{k,t+1}$). The system model assumes very small, i.i.d. Gaussian distributed changes in head angle from frame to frame

$$\theta^H_{t+1} \quad \sim \quad \theta^H_t + \mathscr{N}(0, \sigma^2_H), \tag{2}$$

with $\sigma_H = 1$ determined empirically. The speaker propagation component of the system model is an obvious place to exploit pitch information (e.g. pitch tracks). In [4] the speaker changes controlled by *cur* were modelled by a two-state model with a probability q of staying in the same

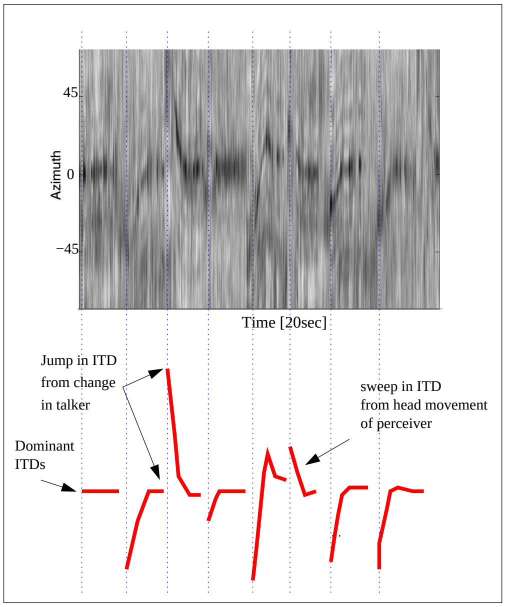


Figure 2: *Illustration of summed cross-correlogram for 20 seconds of data from the P1 CAVA session. The underlying ITD 'track' has been manually drawn below.*

sate, i.e. continuing with the same speaker, and a probability $(1-q)$ of changing state/speaker. In this paper we have augmented this model such that the probability q takes on a different value depending on whether the current frame is in a pitch track. Specifically we use $q_{t\in\mathscr{P}} = 0.998$[2] and $q_{t\in\mathscr{P}'} = 0.995$[3] where $\mathscr{P}$ is the set of frames associated with a pitch track, and $\mathscr{P}'$ is the set of frames without an associated pitch track. This 'tightens' up the speaker duration model and inhibits particles changing speaker mid-track. As previously, the propagated $\theta^S_{k,t+1}$ will be drawn from a Gaussian distribution

$$\theta^S_{k,t+1} \sim \mathscr{N}(\theta^{S'}_k, \sigma^2_{S'}) \tag{3}$$

where $\theta^{S'}_k$ and σ'_S are the known mean position and standard deviation of the speaker. σ'_S, was estimated from the data to be about 2.

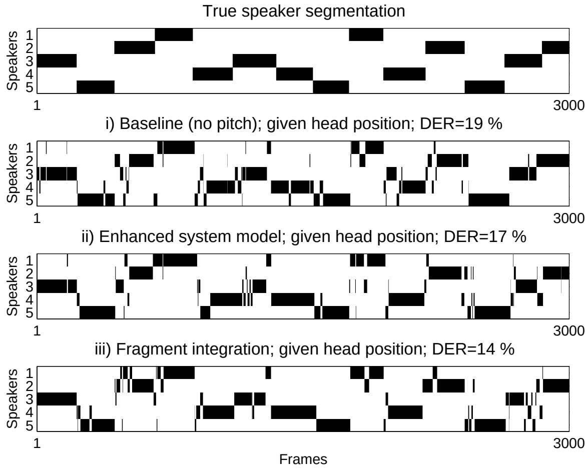
3.3 Measurement model

The measurement model expresses our belief about the likelihood of observations conditioned on the current state of the system. The set of three cross-correlogram peak positions that have been observed are mapped into azimuth estimates (i.e. time delay is mapped onto angle), $D \overset{\triangle}{=} (D_1, D_2, D_3)$. We assume that at most one of the candidate measurements corresponds to the true peak and that the rest are due to spurious peaks, 'clutter' peaks. The true azimuth associated with the system state α, i.e. the true location of the current speaker relative to the listener's head, is given by

$$D_\alpha \overset{\triangle}{=} (\theta_\alpha) = (\theta^S_{\alpha,cur} - \theta^H_\alpha), \tag{4}$$

[2] Determined in pilot experiments.

[3] Estimated from the data.



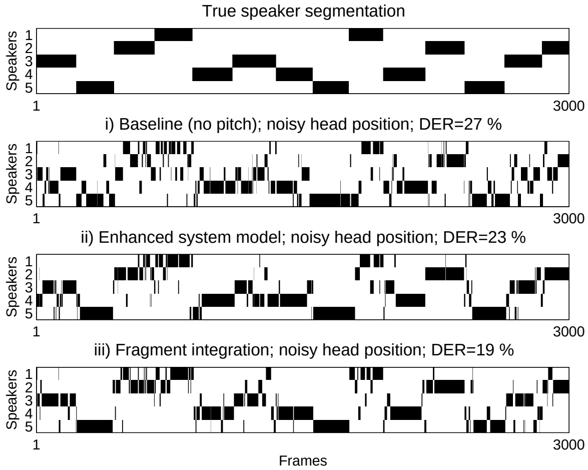
True speaker segmentation

i) Baseline (no pitch); given head position; DER=19 %

ii) Enhanced system model; given head position; DER=17 %

iii) Fragment integration; given head position; DER=14 %

Figure 3: *Examples of speaker change segmentations in the case of given, known self-position for different systems; the top panel shows the true speaker segmentation.*



True speaker segmentation

i) Baseline (no pitch); noisy head position; DER=27 %

ii) Enhanced system model; noisy head position; DER=23 %

iii) Fragment integration; noisy head position; DER=19 %

Figure 4: *Examples of speaker change segmentations in the case of noisy self-position for different systems; the top panel shows the true speaker segmentation.*

The measurement model is used in the 'update' state of the particle filtering algorithm, where the particles are updated with the knowledge we can gain from the new observations. Hence, we are interested in the likelihood function, $p(D|\alpha)$. We note that as Eq. 4 defines a deterministic mapping, the likelihood satisfies $p(D|\alpha) = p(D|D_\alpha)$, which we will base our development on. We assume that each of the peak locations observed are independent, so that

$$p(D|D_\alpha) = \prod_{i=1}^{N} p(D_i|D_\alpha). \tag{5}$$

Following the approach in [8] we develop a description for each $p(D_i|D_\alpha)$ based on the hypothesis that at most one of the observed peaks will have arisen as a result of the true state space and the remaining peaks are clutter. This is described below by using the indicator variable c_i, such that $c_i = T$ if D_i is associated with the true source, and $c_i = C$ if D_i is associated with clutter. The likelihood for a measurement from the true source is taken to be

$$p(D_i|D_\alpha, c_i = T) = c_\alpha \mathcal{N}(D_i; D_\alpha, \sigma_D^2) \quad for \quad \mathscr{D}(D_i), \tag{6}$$

where $\mathscr{D} \overset{\triangle}{=} [-D_{max}, D_{max}]$ is the set of admissible azimuth values for the microphones, and c_α is a normalising constant. Thus, a true source peak is assumed to be normally distributed around the true relative azimuth. The likelihood of a clutter peak is assumed to be uniformly distributed within the admissible interval, independent of the true relative azimuth

$$p(D_i|c_i = C) = \mathscr{U}_{\mathscr{D}}(D_i). \tag{7}$$

The overall likelihood is found by summing over the possible hypotheses of true and clutter peaks [8].

In certain applications, information about the listener's position might be available and hence should be included in the measurement model; Eq. 5 is thus expanded

$$p(D, H|\alpha) = \prod_{i=1}^{N} p(D_i|D_\alpha) \cdot p(H|H_\alpha) \tag{8}$$

where θ^H_{obs} is the observed self-position angle, $H \overset{\triangle}{=} (\theta^H_{obs})$ and $H_\alpha \overset{\triangle}{=} (\theta^H_\alpha)$; we have assumed that D and H are independent given the state, α. We take the observation noise of the head position measurements to be normally distributed

$$p(H|H_\alpha) \sim \mathcal{N}(H; H_\alpha, \sigma_H^2). \tag{9}$$

The σ_H is set to match the variance used for generating the simulated, observed head tracks.

4. RESULTS

At each frame the system outputs the value of the current speaker, *cur*, which has the maximum posterior probability, i.e. *cur*, is chosen as the value $k \in 1:K$ which has the largest total particle weight associated with it. The system is evaluated by comparing the against the correct active speaker (as given by the CAVA corpus' manual annotation) and computing the diarization error rate (DER) as defined by [14]:

$$DER = \frac{\text{Number of frames incorrectly assigned}}{\text{Total number of frames}} \times 100. \tag{10}$$

DER was measured on systems without access to self-position information and on systems with access to self-position information corrupted by varying degrees of noise. The noisy self-position observations were obtained by adding Gaussian noise with increasing standard deviation to the true head tracks.

Examples of speaker segmentations output are shown in Figure 3 (no access to the true self-position) and 4 (access to noisy self-position, $\sigma^2 = 20°$). The true current speaker segmentation (top panels) is compared against outputs by the system for different usages of pitch information. Comparing Figure 3 to 4 it is visually clear that results deteriorate as self-position information is reduced, but that pitch information can improve system performance.

The overall results from measuring DER on segmentations based on localisation and noisy self-position measurements are presented in Figure 5. The systems have either

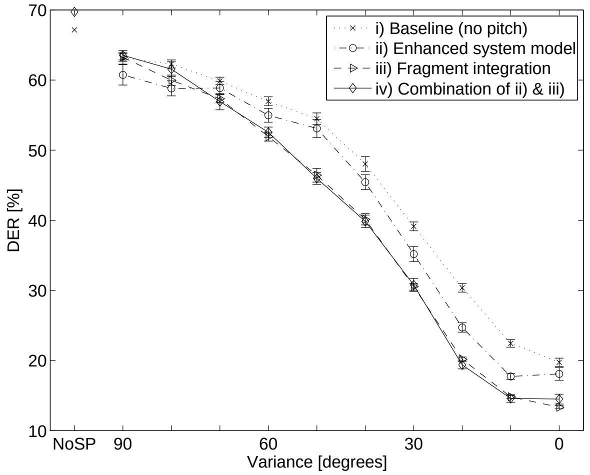


Figure 5: *DER scores for different systems using pitch and different degrees of simulated measurement noise in the self-position observations. 'NoSP' denotes not supplying any self-position information. Results are averaged over 10 runs and using* $40,000$ *particles. The error bars indicate the standard error of the mean.*

access to no self-position information (indicated by 'NoSP' on the plot) or some measurements of self-position with a noise variance varying from 90° down to 0° (that is, the true head position is given to the system). For all the different systems the DER values decreases from around 70 % down to only 15-20 %. When the variance of the noise drops to below 45 degrees the improvement in DER is very noticeable. Regarding pitch, both using pitch to inform the system model (system ii) and improving the ITD-based observations by integrating across pitch-based fragments (system iii) provide significant improvements over the ITD-only baseline of our previous system for all but the severest of self-position noise settings. However, the two systems do not appear to be complementary as combining them (system iv) does not provide any significant additional benefit.

5. CONCLUSIONS

It has been demonstrated how pitch and location cues can be usefully combined with noisy head-position estimates in a particle filtering framework to track speaker changes from the perspective of a moving listener. We have proposed two different methods for using pitch; i) by enhancing the system model to discourage speaker changes during voiced-speech segments, and ii) by improving ITD observations through the integration across pitch-based speech fragments. Although overall performance decays rapidly as head-position noise increases, adding pitch information reduces DER by about 10% absolute over a wide range of operating conditions. Using pitch to extract more reliable ITD observations bought the biggest gains.

Acknowledgements

The authors would like to thank Ning Ma of Sheffield University for kindly making his fragment generation software available online.

REFERENCES

[1] D. Imseng and G. Friedland, "An adaptive initialization method for speaker diarization based on prosodic features," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, USA, March.

[2] K. Ishizuka, S. Araki, K. Otsuka, T. Nakatani, and M. Fujimoto, "A speaker diarization method based on the probabilistic fusion of audio-visual location information," in *Proc. of int. conf. on Multimodal Interfaces*, Cambridge, USA, 2009, pp. 55–62.

[3] J. Leung, D. Alais, and S. Carlile, "Compression of auditory space during rapid head turns," *PNAS*, vol. 105, pp. 6492–7, 2008.

[4] H. Christensen and J. Barker, "Using location cues to track speaker changes from mobile, binaural microphones," in *Proc. INTERSEPEECH'09*, Brighton, UK, Sep 2009.

[5] H. Christensen, N. Ma, S. N. Wrigley, and J. Barker, "A speech fragment approach to localising multiple speakers in reverberant environments." in *Proc. of ICASSP'09*, Taipei, Taiwan, April 2009.

[6] E. Arnaud, H. Christensen, Y.-C. Lu, J. Barker, V. Khalidov, M. Hansard, B. Holveck, H. Mathieu, R. Narasimha, E. Taillant, F. Forbes, and R. Horaud, "The CAVA corpus: Synchonised stereoscopic and binaural datasets with head movements." in *Proc. of Internation Conference on Multimodal Interfaces*, Crete, Greece, 2008.

[7] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussianbayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[8] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. ICASSP'01*, Salt Lake City, Utah, US, 2001, pp. 3021–3024.

[9] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 44, pp. 99–122, 1990.

[10] N. Ma, P. Green, J. Barker, and A. Coy, "Exploiting correlogram structure for robust speech recognition with multiple speech sources," *Speech Commun.*, vol. 49, no. 12, pp. 874–891, 2007.

[11] L. A. Jeffress, "A place theory of sound localization," *Comparative Physiology and Psychology*, vol. 41, pp. 35–39, 1948.

[12] F. Talantzis, A. Constantinides, and L. Polymenakos, "Estimation of direction of arrival using information theory," *IEEE Signal Processing Letters*, vol. 12, pp. 561–564, Aug 2005.

[13] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech and Audio Proc.*, vol. 11, pp. 549–557, 2003.

[14] "The 2009 (trt-09) rich transcription meeting recognition evaluation plan," http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-pl%an-v2.pdf, 2009.