INTERSPEECH 2009 BRIGHTON

Using location cues to track speaker changes from mobile, binaural microphones.

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 presents initial developments towards computational hearing models that move beyond stationary microphone assumptions. We present a particle filtering based system for using localisation cues to track speaker changes in meeting recordings. Recording are made using in-ear binaural microphones worn by a listener whose head is constantly moving. Tracking speaker changes requires simultaneously inferring the perceiver's head orientation, as any change in relative spatial angle to a source can be caused by either the source moving or the microphones moving. In real applications, such as robotics, there may be access to external estimates of the perceiver's position. We investigate the effect of simulating varying degrees of measurement noise in an external perceiver position estimate. We show that only limited self-position knowledge is needed to greatly improve the reliability with which we can decode the acoustic localisation cues in the meeting scenario.

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

10.21437/Interspeech.2009-52

1. Introduction

There have been many previous attempts to model the auditory system's ability to localise and track sound sources. Previous models have attempted to account for the robustness of sound localisation in the presence of additive and convolutive noise (see [1] for a review). However, most of these previous systems employ a *stationary* binaural microphone set up and hence overlook one of the auditory system's more remarkable abilities: the ability to track moving sources using sensors (i.e. ears) that are themselves rarely stationary.

The emergence of mobile hearing applications, such as perceptual robotics and wearable listening devices, lends urgency to the development of computational hearing models that can move beyond stationary microphone assumptions. In this paper we make initial steps towards such models. 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 direction cues to track speaker changes in a meeting. We reconsider this problem from the perspective of a meeting participant making natural head movements.

Allowing the acoustic sensors to 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. Note, head rotation can approach speeds of up to 500 degrees/sec, equivalent to 5 degrees per 10 ms analysis window. Rapid rotation thus results in significant 'motion blurring' of location estimates. Second, head motion introduces extra ambiguity [2]. If a single source is dominating the acoustic scene, then a *clockwise* head movement may be hard to distinguish from a movement of the source in an *anticlockwise* direction around the head, and vice versa. In real systems this second problem may be countered by complementary sensory input from other modalities. In the current study we assume we have access to a (more or less) noisy estimate of the true head position such as might be available to biological systems from proprioceptive feedback.

The paper presents a general solution for tracking sources from a moving perceiver which operates by simultaneously modelling and tracking changes in the state of both the external environment and the perceiver. Section 2 presents both the general framework, and the particular speaker turn-taking scenario on which we have evaluated our systems. Section 3 describes our specific tracking implementation based on particle filtering. Results and conclusions follow in Sections 4 and 5.

2. The sound source tracking problem

2.1. The general approach

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 perceiver and sound source position parameters can be described by a state space that evolves over time. We are particularly interested in inferring the sound source position parameters, but may also wish to infer the perceiver's position and orientation.

In order to proceed we will extract standard localisation cues — in the current work we have concentrated on modelling interaural time differences (ITDs). We then employ two statistical models: i) a *measurement model* that describes the distribution of the ITD observations for a given set of perceiver and source position parameters; and ii) a *system model* that captures how the perceiver and source parameters evolve over time. With access to these models we can treat the problem using recursive Bayesian estimation, which in our case we implement using particle filtering [3].

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

This work was funded by the EU Cognitive Systems STReP project POP (Perception On Purpose), FP6-IST-2004-027268.

in a meeting scenario. Data from the CAVA database has been employed [4]. This data was recorded from the perspective of a moving, human head and in a conversational situation with five speakers. The purpose of the CAVA corpus is to enable the investigation of binaural and stereoscopic cues from a humans' perspective in various environments. This was achieved by equipping the 'perceiver' (either a human or a dummy head) with a pair of binaural in-ear microphones. The perceiver was also wearing a helmet on which a pair of stereoscopic cameras were mounted (in this work the visual data stream was not used) and finally, on top of the helmet a 6 degree of freedom head tracking device was fitted so that the true head position and orientation is known. The head tracker information can be used to verify our algorithms, and it also provides a means of simulating potential position feedback information.

We have focused on a particular session from the CAVA database – Panel Meeting 1 (P1). Here the human perceiver and 5 'actors' are sitting around a table. Actors are separated by roughly 25° measured from the perceiver (see Figure 1). The perceiver is blind folded and the actors take turns to speak (counting from 1-5). Perceiver head movements have been induced by giving the perceiver the task of monitoring speaker changes and turning to face the current speaker. The task for our system will be to use ITD cues in the binaural recording to estimate which of the speakers is active at any instant.

2.3. Modelling the turn-taking meeting scenario

The meeting scenario is suitable for this initial study because it allows us to considerably 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 perceiver'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. From the acoustic signal we are extracting localisation cues that indicate θ^O the spatial angle of a sound source *relative* to the rotational angle of the perceiver's head. This perceived angle is the difference between the absolute angle of the perceiver's head, θ^H (i.e. the angle relative to a fixed room axis) and the absolute spatial angle of the active sound source θ^S_{cur} (see Figure 1). 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.

We model the situation with a state space

$$\boldsymbol{\alpha} \triangleq (\theta^H, \theta^S_1, \ldots \theta^S_K, cur), \tag{1}$$

where θ^H is the absolute spatial angle (azimuth) of the head, θ^S_k is the absolute azimuth of speaker k, K is the total number of speakers, and $cur \in \{1, \ldots, K\}$, indicates which speaker is speaking.

This model allows for a fully dynamic setup, where the perceiver's head can be turning, and where each sound source can be moving around independently. Following our assumptions, θ^S_k will be constrained to vary within a small range of a known initial position, $\theta^{S'}_k$.

3. Particle filtering

The task of tracking the azimuth of a set of sound sources lends itself to sequential, recursive filtering approaches where a new

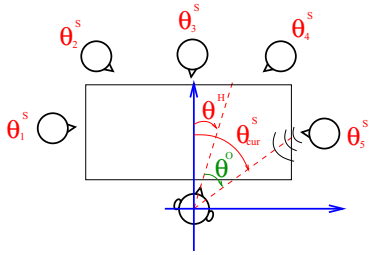

Figure 1: *Illustration of the definition of θ–parameters for P1.*

Bayesian estimate of the state is inferred for each time step by combining the previous estimate with what can be learnt from the incoming set of observations (see [3] for a tutorial). Particle filtering methods attempt to construct a posterior probability density function (pdf) of the state based on all available observations as well as other prior information such as what might be known about the dynamicity of the targets. We will present an overview of a particle filtering formulation suitable for the case of tracking multiple sound sources from the point of view of an active perceiver. This can be seen as an extension of Vermaak and Blake's formulation of the case for a static perceiver [5].

3.1. Observations

The ITD-based observations 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 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 $+90^\circ$ azimuths. The standard procedure of estimating ITDs is to identify one or more peaks in the summed cross-correlogram (e.g. Jeffress' model [6]), 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 summed cross-correlogram looks like for parts of the P1 CAVA session. The underlying 'track' of ITDs was plotted above the image. The sweeps arising from when the perceiver is turning his head 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. We have therefore chosen to extract the lags corresponding to the three largest peaks for each frame; this has proven to be a good compromise between ensuring that a high probability of the true ITD is being observed and that not too much noise is included in the data.

3.2. System model

The system model determines how the state represented by each particle is progressed at each time step: $\boldsymbol{\alpha}_t \rightarrow \boldsymbol{\alpha}_{t+1}$, i.e. a head angle model ($\theta^H_t \rightarrow \theta^H_{t+1}$) and a speaker change model ($\theta^S_{k,t} \rightarrow \theta^S_{k,t+1}$, $cur_t \rightarrow cur_{t+1}$).

The system model assumes very small, i.i.d. Gaussian distributed changes in head angle from frame to frame

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

with $\sigma^H = 1^\circ$, determined empirically.

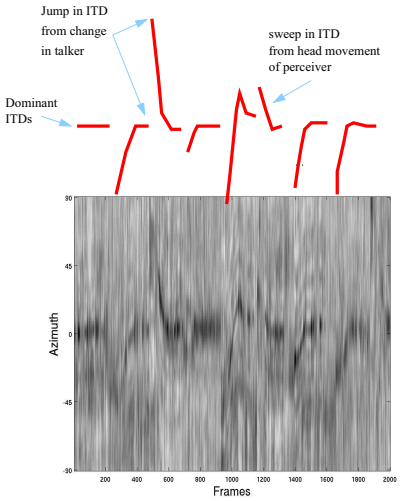

Figure 2: *Illustration of summed cross-correlogram for 2000 frames of the P1 CAVA session. The underlying ITD 'track' has been manually drawn above.*

The speaker changes controlled by cur are modelled by a two-state model with a probability q of continuing with the same speaker and a probability $(1-q)$ of changing speakers. The propagated $\theta^S_{k,t+1}$ will be drawn from a Gaussian distribution

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

where $\theta^{S'}_k$ and σ_S are the known mean position and standard deviation of the speaker. Parameters were estimated from data as $q = 0.9953$, $\sigma_S = 2$.

3.3. Measurement model

The measurement model provides the model with our belief about the likelihood of observations conditioned on the current state. The sensory system consists of a pair of in-ear microphones and the observation vector is extracted by identifying three peaks in the cross-correlogram. These are transformed into azimuth, $\boldsymbol{D} \triangleq (D_1, \ldots, D_N)$, where $0 \leq N \leq N_{max}$ is the number of candidate azimuth measurements. 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 source state $\boldsymbol{\alpha}$ is given by

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

where $\theta_{\boldsymbol{\alpha}}$ is the true location of the current speaker relative to the perceiver's head.

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, $p(\boldsymbol{D}|\boldsymbol{\alpha})$. We note that as Eq. 4 defines a deterministic mapping, the likelihood satisfies $p(\boldsymbol{D}|\boldsymbol{\alpha}) = p(\boldsymbol{D}|D_{\boldsymbol{\alpha}})$, which we will base our development on. We assume that each of the peak locations observed are independent, so that

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

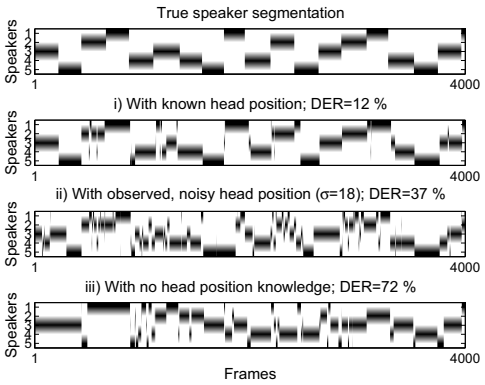

Figure 3: *Examples of speaker change segmentations from separate runs of the system. The top panel is the true segmentation and the following panels are system outputs for: i) when the head position is known, ii) the head position is observed with some noise, and iii) no head position information is available.*

Following the approach in [5] we develop a description for each $p(D_i|D_{\boldsymbol{\alpha}})$ 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

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

where $\mathcal{D} \triangleq [-D_{max}, D_{max}]$ is the set of admissible azimuth values for the microphones, and $c_{\boldsymbol{\alpha}}$ 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

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

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

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

where $\boldsymbol{H} \triangleq (\theta^H_{obs})$ and $H_{\boldsymbol{\alpha}} \triangleq (\theta^H_{\alpha})$ and we have assumed that $\boldsymbol{D}$ and $\boldsymbol{H}$ are independent given the state, $\boldsymbol{\alpha}$. We take the observation noise of the head position measurements to be normally distributed

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

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

4. Results

The task is to estimate which speaker is active at each frame. It was evaluated using the diarization error rate (DER) as defined

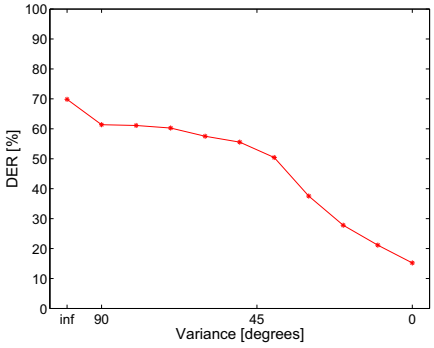

Figure 4: *DER averaged over 5 runs for different degrees of simulated measurement noise in the self-position observations.*

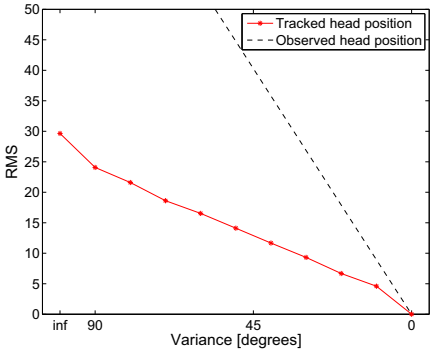

Figure 5: *RMS error (in degrees) for the inferred head orientation when using initial head orientation estimates with additive noise of a given variance. Results averaged over 5 runs.*

by [7]:

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

For each frame, the system's judgement of the active speaker was taken by finding the value of cur that was assigned the most weight when summed across all particles. The 'correct' value was obtained from a hand segmentation of the data. DER was measured either for systems which used no external self-position estimate, or systems which employed self-position estimates with variable degrees of noise. The self-position estimates were generated by adding Gaussian noise of known standard deviation to the true head orientation tracks.

Figure 3 shows the true speaker change segmentation (top panel) against three examples of segmentations as output by the system for different degrees of self-position noise: i) known head position (i.e. no noise), ii) observed, noisy head position and iii) no head position knowledge. The segmentation deteriotates as the head orientation noise increases.

The overall results for the system's DER, averaged over multiple runs, are presented in Figure 4. Note first that if the head position is precisely known (i.e. variance equals zero) then DER is around 15%. From Figure 3 it can be seen that most of these errors occur as isolated frames and often correspond to short pauses at speaker boundaries or within an utterance. Most of these errors could be fixed with trivial post-processing. As noise is added to the head position estimates performance deteriotates slowly for sigma up to 20° and then more rapidly. Once the noise is greater than 50° the head track information provides little advantage over having no initial estimate at all.

Figure 5 plots the RMS error in the inferred estimates of perceiver's head orientation (θ^H) given initial head orientation estimates with varying degrees of noise. As one would expect, the head tracking accuracy improves as soon as any self-position information (however noisy) is provided. For reference, the dashed line on the plot shows the RMS error value of the initial noisy estimate. The difference between the two lines indicates the improvement in head position certainty, when localisation cues extracted from the acoustic signal are taken into account in the algorithm. Note, diarization performance deteriorates rapidly at about the point when the tracked head error becomes comparable to the half angular separation between speakers (i.e. allowing confusions to occur).

5. Conclusions

We have re-examined the task of tracking speaker changes using localisation from the perspective of a listener sitting in a meeting performing natural head movements. This is a challenging problem as the effect of the perceiver turning his head and the effect of a speaker change can hard to distinguish using noisy binaural cues. A particle filtering solution is seen to work well when perceiver head orientation is known a priori and to degrade reasonably gracefully as noise is added to the head orientation estimate. Future work will look at expanding the model to handling multiple, simultaneous speakers.

6. References

[1] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis : Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[2] J. Leung, D. Alais, and S. Carlile, "Compression of auditory space during rapid head turns," in *Proc. National Academy of Sicences of the USA*, vol. 105, no. 17, April 2008, pp. 6492–6497.

[3] 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.

[4] 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.

[5] 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.

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

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