Active binaural distance estimation for dynamic sources

Yan-Chen Lu, Martin Cooke and Heidi Christensen

Department of Computer Science, University of Sheffield, Sheffield, UK

{y.c.lu, m.cooke, h.christensen}@dcs.shef.ac.uk

Abstract

A method for estimating sound source distance in dynamic auditory 'scenes' using binaural data is presented. The technique requires little prior knowledge of the acoustic environment. It consists of feature extraction for two dynamic distance cues, motion parallax and acoustic τ, coupled with an inference framework for distance estimation. Sequential and non-sequential models are evaluated using simulated anechoic and reverberant spaces. Sequential approaches based on particle filtering more than half the distance estimation error in all conditions relative to the non-sequential models. These results confirm the value of active behaviour and probabilistic reasoning in auditorily-inspired models of distance perception.

Index Terms: particle filter, auditory distance perception

1. Introduction

Many previous studies have described human distance perception performance (e.g. see review in [1]) although very few computational models exist. Listeners' estimation of the distance to a sound source is generally much less accurate than the ability to determine the angular direction of a sound source. Listeners can resolve changes in direction of approximately 1˚ for frontal sources [2], but significantly underestimate the distance to faraway sources and typically overestimate the distances to sources closer than 1m.

Potential cues to distance can be classified into relative cues and absolute cues. Relative cues include loudness and source spectrum, but prior information about the sound source is required to estimate absolute distance. For anechoic conditions, the loudness cue can be used to determine changes in the distance of a constant amplitude sound source according to the inverse square law. Differential absorption of frequencies along the propagation path is the major source of spectral cues.

Familiarity, binaural information and reverberation deliver absolute cues. If the listener is sufficiently familiar with the sound source, relative cues can be used to judge absolute distance. Listener familiarity with both the source signals and the acoustic environment is clearly a key factor in any model of auditory distance perception. For near-field listening (distance < 1m), binaural cues based on interaural time and intensity differences provide not only directional but also distance information. Several models [3-5] take advantage of systematic changes in interaural differences. However the utility of these cues for auditory distance perception is doubtful for far-field (distance > 1m) sources because interaural differences are very nearly independent of source distance at such distances.

Another important distance cue is the contribution of reverberant energy relative to direct energy. When sound is produced in a reverberant space, the associated reverberation may often be perceived as a background ambience, separate from the direct energy. The ratio of direct to reverberant energy is greater with nearby objects than it is with distant objects. Thus, distant objects sound more reverberant than close objects. Bronkhost and Houtgast [6] formulated a computational model of auditory distance perception based primarily on the direct-to-reverberant energy ratio. This model requires the listener to have prior knowledge of the reverberation characteristics of the environment.

All the cues described so far assume that the listener is stationary. In the real world, sound sources and listeners are seldom stationary, and their motion has the potential to provide additional cues to auditory distance perception (fig. 1). It has been suggested that motion-induced rate of change of intensity can provide listeners with reliable distance information [7]. This cue, known as acoustic τ (time-to-contact), may also be expressed as a ratio of distance to velocity when velocity is constant. In addition, listener motion creates a changing azimuth, or motion parallax, with respect to a stationary source. This can be used to estimate source distance via the translation distance. The calculation of acoustic τ (fig. 1, right) needs prior distance information from another cue such as motion parallax and hence must be exploited in a framework of multiple, coupled cues. Speigle and Loomis [7] found that dynamic cues of motion parallax and acoustic τ influence an observer's judgment of source distance above and beyond static cues. However, their experiments involved relatively simple auditory scenes and it is an open question as to whether dynamic cues are more or less useful in realistic environments.

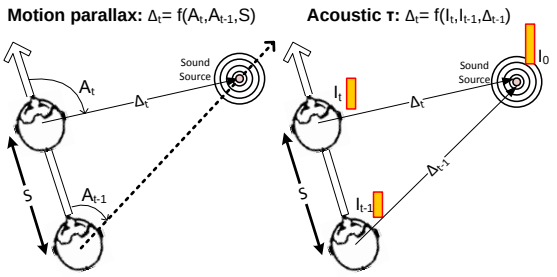

Figure 1: Two potential dynamic auditory cues to distance: motion parallax (left) and acoustic τ (right).

Bronkhorst and Houtgast's model requires acoustic source time-of-onset to separate the contributions of direct and reverberant energy and relies on the existence of a detailed environmental description. By contrast, the current study tests the idea that a listener's active behaviour provides useful cues for distance perception. Consequently, we employ the two dynamic distance cues mentioned above. We further focus on sources beyond 1m since static binaural information can provide salient cues for near-field sources. The central question of the current study is to determine whether it is possible to track, using only dynamic cues, the varying distance from a moving listener to a fixed source. To tackle this tracking problem, we introduce a model based on particle filtering [8] to incorporate a moving listener which exploits the potential of dynamic distance cues, and measure the effectiveness of particle filtering by comparing the models based on 'instantaneous' distance estimates derived from dynamic cues.

2. Computational model

Fig. 2 depicts the proposed model. Dynamic cues are generated from successive measurements of cross-correlation and intensity. Distance inference is based on triangulation for motion parallax and an adaptation of the inverse square law for intensity-based cues.

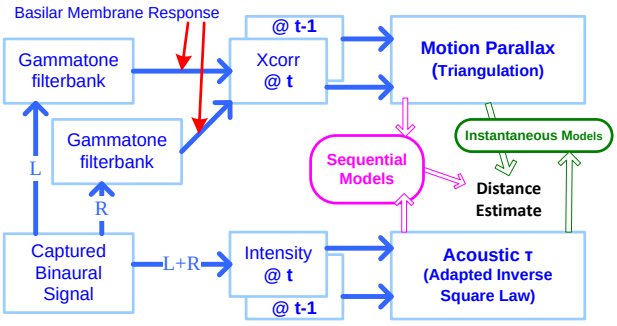

Figure 2: Proposed computational model for distance estimation.

2.1. Azimuth-based cues

Given the estimated azimuth A_{t-1} before moving and the position displacement S due to the motion, one distance hypothesis Δ_i can triangulate one azimuth value A_t and vice versa (see Fig. 1, left). This motion parallax cue needs a robust azimuth estimate to infer the required distance value. Interaural time difference (ITD) has been considered the most reliable acoustic cue in azimuth localisation. The calculation of ITD is based on the cross-correlation of the outputs of auditory filters modelled using a pair of gammatone filterbanks [9]. While the use of ITD information is known to lead to front-back confusions [2], dynamic cues resulting from head movements or listener motion may be important in their resolution.

2.2. Intensity-based cues

Intensity-based cues to distance stem from an adaptation of the inverse square law relating intensity to distance:

$$I = P\left(a \times \Delta^{-2} + b\right) \quad (1)$$

where P is sound source power and Δ is the source-listener distance. In anechoic space, eqn. 1 becomes the inverse square law with a and b equal to $1/4\pi$ and zero respectively. In reverberant space, the term b represents the contribution of diffused reverberant energy which is independent of distance and is assumed to be fixed for static environments. Reverberant energy also decreases in distance [10]. As a consequence, a takes on a larger value than for an anechoic space.

Since acoustic τ is based on successive estimates, there is no need to know P to estimate the current perceived sound intensity I_t. Given estimates of distance and intensity, Δ_{t-1} and I_{t-1}, at the previous time step, and the current distance hypothesis Δ_t, I_t can be computed from (1) as follows:

$$I_t = I_{t-1}(a\Delta_{t-1}^2 + b\Delta_{t-1}^2\Delta_t^2 / a\Delta_t^2 + b\Delta_{t-1}^2\Delta_t^2) \quad (2)$$

2.3. Sequential models

The sequential models we employ – particle filters (PFs) – use probabilistic inference to monitor a target. It has been shown that PFs are an effective way of tracking sound sources in reverberant environments [11]. Here, we estimate source distance using observations accumulated over time by a moving listener. The target states are modelled using a collection of N particles (h^i, ω^i), at each time point t:

$$h_t^i = (\Delta_t^i, A_t^i, I_t^i)\ ,\ i = 1,\ldots, N \quad (3)$$

where Δ^i, A^i and I^i are random variables representing hypotheses for distance, azimuth and intensity for the i_{th} particle. A probability distribution function characterizes these state variables and particles are drawn from it to provide a sample-based representation. A weight ω is associated with each particle during this sampling process.

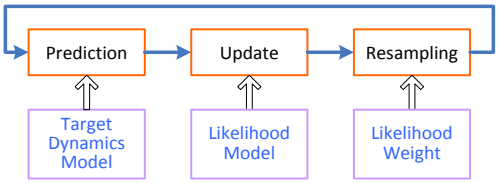

Figure 3: Standard particle filtering process.

Particle filtering is the iterative application of the operations depicted in Fig. 3. Each operation alters the state variables and associated particle weights based on models of the sound source dynamics and the likelihood of the current observations. Estimates are provided by the weighted mean across all particles. Each iteration of the PF algorithm has three stages: prediction, update and resampling. Each particle is first modified based on the prediction made by the model of target dynamics. Next, particle weights are updated according to a likelihood function derived from current observations. Finally, particles with low weights are eliminated and replaced using a resampling mechanism which maintains a proper sample-based representation of the true pdf.

The above description is referred to as a sampling importance resampling (SIR) particle filter [8]. In our implementation, noise is added to state variables at the prediction stage to differentiate the duplicated particles at the re-sampling step. To avoid the potential loss of diversity among particles caused by performing the resampling at every time step, we calculate an effective sample size N_{eff} [8] which helps monitor the insignificance of particles and resampling schedule.

Particle redistribution during resampling using both the current observation and that available at the next time-step leads to a variant of SIR particle filtering known as auxiliary sampling importance resampling (ASIR) [12] which allows the state space to be explored by considering the on-going dynamics. Consequently, the generated particle distribution will be more likely to be closer to the true pdf.

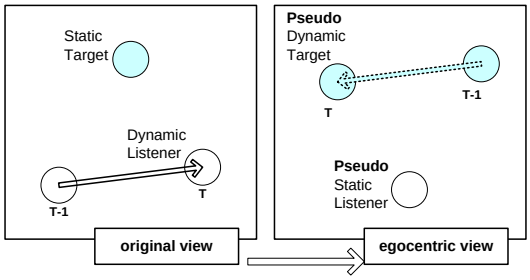

Figure 4: Conversion from dynamic listening to a tracking scenario.

The PF framework for tracking sound sources requires a model of sound source dynamics. PF is normally used in target tracking applications. Here, the target is assumed to be static and the listener accounts for the dynamics (fig. 4, left). However, we can also adopt an egocentric view in which the motion is that of the source relative to a static listener (fig. 4, right). The source is viewed as dynamic with motion determined by listener movement. The distance hypotheses of all particles are

transformed according to this movement. A noise term is applied to some particles (with probability 0.5) to effect an additional shift in both azimuth and distance. Particle weights ω^i are linearly rescaled to occupy the full range 0~1 (i.e. the minimum particle weight is 0 and the maximum 1 after rescaling). The noise term is weighted by $(1-\omega^i)$, so that weaker particles make larger movements. The derived distance hypothesis Δ_t is used to update the state parameters A_t and I_t as described in section 2. These two updated parameters are later evaluated with the likelihood models described below to generate an associated likelihood weight.

2.3.1. Likelihood model

The weight ω^i of each particle i is updated according to functions derived from the current binaural acoustic observation Z, represented by Z_{cc}, the summary cross-correlation (i.e. the individual channel cross-correlations averaged across frequency), and Z_{intens}, the summed intensity from the binaural signals.

We used the pseudo-likelihood approach from [11] to treat the cross-correlation function as the likelihood function directly in evaluating the hypothesized azimuth of each particle:

$$p(Z_{cc} \mid h_t) = f_{CC}(\theta, B_t), -\pi/2 < \theta \le \pi/2 \quad (4)$$

where f_{cc} is the summary cross-correlation function, B_t is the binaural input from the gammatone filters at the current time step, and θ is the azimuth angle represented by a given cross-correlation lag. Sample lags of the cross-correlation function are transformed into azimuth angle θ by applying Woodworth's spherical model [13].

A one-dimensional Gaussian distribution whose mean is the intensity estimate I_t in eqn. 2 is used as the likelihood function for intensity. In addition, if the P in eqn. 1 is calculated by replacing Δ with the previous distance estimate $\hat{\Delta}_{t-1}$ (weighted over all particles) and I with the previous intensity measure, $Z_{intens,t-1}$, we arrive at a new intensity-based estimate I_t' which can be viewed as cue based on acoustic power:

$$I_t' = Z_{intens,t-1}(a\hat{\Delta}_{t-1}^2 + b\hat{\Delta}_{t-1}^2\Delta_t^2 / a\Delta_t^2 + b\hat{\Delta}_{t-1}^2\Delta_t^2) \quad (5)$$

The product of Gaussian distributions (N_p and N_τ for the acoustic power and acoustic τ cues respectively) whose means are the intensity estimates I_t' and I_t is used as the likelihood function for intensity:

$$p(Z_{intens} \mid h_t) = N_s(\mu_p, \sigma_p^2) \cdot N_\tau(\mu_\tau, \sigma_\tau^2) \quad (6)$$

An accurate distance estimate can improve the reliability of this cue and help stabilize the PF performance after reaching a certain confidence level. However, it is worth noting that inaccurate distance estimates and intensity measures may degrade I_t'. A larger σ will lead to a smoother likelihood function and generate weights with smaller contrast. During pilot trials, we noted that the acoustic power cue is better underemphasized relative to the acoustic τ. The values $\sigma_p = 4$ and $\sigma_\tau = 0.4$ were chosen here. The final likelihood function was the product of individual functions for azimuth and intensity:

$$p(Z \mid h_t) = p(Z_{cc} \mid h_t) \cdot p(Z_{intens} \mid h_t) \quad (7)$$

Both (6) and (7) make the assumption that the estimates are independent. This is unlikely to be the case in practice since the calculation of distance and acoustic τ are coupled.

2.3.2. Robust distance estimation

Given a particle distribution, several different methods can be used to obtain a distance estimate. Apart from the global weighted mean, it is possible to choose the best particle or the weighted mean in a small window around the best particle (also called the robust mean). The weighted mean derived from multi-modal distributions can lead to a biased estimate if there is no proper weighting between individual measurements, while the best particle introduces a discretization error. The robust mean is considered the best method and is used in the evaluations presented here, but it is also the most computationally expensive because of the determination of the window.

2.4. "Instantaneous" (non-sequential) models

In the instantaneous model, we assume no prior knowledge from the previous states and there is no probabilistic particle filter model to support the current state estimate. Given the current position, motion parallax derives a distance estimate from azimuth measurements of two successive time steps (see fig. 1, left). This model is referred to as MP in table 2. As mentioned earlier, acoustic τ requires a distance estimate from another source, so if the instantaneous azimuth-based distance estimate is employed, we arrive at a second instantaneous model whose distance estimate is simply the mean of the motional parallax and acoustic τ at current time. This is the MP+AT model in table 2.

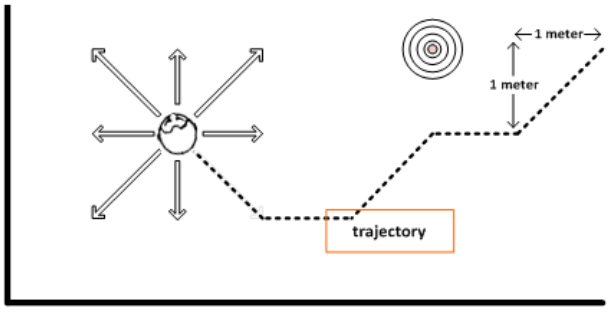

Figure 5: An example of a simulated listener trajectory.

3. Evaluation

Evaluations employed a simulated acoustic environment. In order to add reverberation and spatial location to the original monaural stimuli, impulse responses were created using the Roomsim simulator [14] with a room of size 18m x 18m x 2.75m. The simulated sound source was fixed at the centre of the room, 1.2m above the ground. All surfaces of the room were assumed to have identical reverberation characteristics. Two reverberation surfaces, "acoustic plaster" and "platform floor wooden", were used, with mean estimated T60 reverberation times of 0.34s and 0.51s respectively. The simulated listener can start from any place of the room. At each time step, the simulated listener moved either directly or diagonally forward (i.e. ±45°), avoiding the room boundary and the sound source. For collision avoidance, head orientation was changed until a forward ±45° or 0° movement was possible. For each new position and orientation of the listener, a new pair of room responses was generated with Roomsim. These were convolved with the point sound source, which was a pink noise, to produce the binaural signal.

A distance estimate was generated for each forward movement. The root mean square error (RMSE) of distance for the latter part of the trajectory was taken as an indication of algorithm performance. Blind initialization of the particle hypotheses can lead to poor performance during first few iterations. To avoid adding noise to the RMSE due to this initial "transient", a measure called frame convergence was monitored. Frame convergence occurs when the distance estimate error is smaller

than the standard deviation of the entire particle set estimates [11], and can be used as a performance indicator for different PF algorithms. Since frame convergence requires prior knowledge of the true distance, it cannot be used in live estimation, but convergence times for different PF approaches can be computed offline and the mean time to convergence used in live evaluations. Using a development set of stimuli, we measured an average time to convergence of 15.7 movements for SIR and 9.2 for ASIR. Consequently, the RMSEs reported here were based on that part of the trajectory from the 15^{th} and 9^{th} time steps to the end of the trajectory for SIR and ASIR respectively.

In addition to determination of the value of the different cues in distance estimation, a goal of the evaluation was to compare non-sequential and sequential methods. To ensure a fair comparison, the parameters of each PF algorithm were independently tuned using a reference audio sample to achieve the best performance. This process was done empirically by running each algorithm a number of times with varying parameters until a satisfactory performance was achieved. Table 1 presents the parameter settings chosen for each PF algorithm.

Table 1. *Parameter setting for PF algorithms.*

	Anechoic	RT-0.3s	RT-0.5s
"a" in eqn. 1	0.08	0.134	0.217
"b" in eqn. 2	0	0.002	0.019
	SIR		ASIR
Number of particles	180		225
Robust window size	36		23

*Both PF algorithms used the robust mean estimation method

The fitted parameters from data produced in simulations using Roomsim for intensity-based cues in different acoustic environments are also listed in Table 1. Note that increasing *a* in eqn. 1 leads to better PF performance as more reverberation is present. Strong ceiling and floor echo might account for this value because their contributions in reverberation are also degraded as the source-listener distance increases.

4. Results

Table 2 presents the estimation results in three simulated environments. Results were obtained by averaging over 100 various trajectories, each of which had 50 time steps. In each condition, the benefit of employing intensity-based cues and the sequential PF framework can be observed, although in the most severe reverberant condition the use of acoustic τ and power led to a drop in performance compared to the use of motion parallax only. This may stem from the independence assumption used to combine cues whose estimation is in fact coupled.

Table 2. *Performance comparison of distance estimate error (in meter) for 3 types of simulated environments. Results are given for two instantaneous models (MP and MP+AT) and ASIR PF models (MP+PF and MP+AT'+PF). Numbers in brackets are the SIR PF results.* ***MP****: motion parallax;* ***AT****: acoustic τ;* ***AT'****: acoustic τ + acoustic power cue (see sec. 2.3.1);* ***PF****: particle filtering.*

Model type	Anechoic	RT-0.3s	RT-0.5s
MP	8.3	7.2	7.4
MP+AT	5.6	5.2	6.2
MP+PF	1.9 (2.7)	2.0 (3.3)	3.4 (4.4)
MP+AT´+PF	1.4 (1.8)	1.8 (2.5)	4.4 (5.3)

Particle filtering outperforms the instantaneous methods, and the ASIR algorithm typically works better than the generic SIR approach. Reverberation leads to some degradation in distance estimation accuracy. The ASIR PF approach results in RMSEs below 3.4m in the worst conditions. Although no equivalent data for listeners is available, an approximation can be obtained using the power function proposed by Zahorik et al. [1], leading to an estimate of around 3.5m error (resulting from distance underestimation) at the average true distance value of 7.3m used here. However, this approximation is based on the average of many studies which tested both static and dynamic cues, so the comparison should not be regarded as definitive.

5. Conclusions

A computational model was developed to estimate source-listener distance using dynamic acoustic cues for non near-field sources based on a model of binaural hearing. The method assumed no detailed knowledge of the acoustic environment. Compared to baseline non-sequential models, two sequential particle filtering algorithms operating in a simulated acoustic environment decreased the estimation error to levels commensurate with some estimates of human performance. Future work will compare the performance of listeners and automatic distance estimation in real acoustic environments.

Acknowledgements. This work has funded by the EU Cognitive Systems STReP project POP (Perception On Purpose), FP6-IST-2004-027268. The authors thank Ning Ma for supplying optimized gammetone filterbank C-code implementation.

References

[1] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory Distance Perception in Humans: A Summary of Past and Present Research," *Acta Acustica united with Acustica* 91, 409-420, May/June 2005.

[2] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," *Annu Rev Psychol* 42, 135-159, 1991.

[3] D. S. Brungart, "Preliminary model of auditory distance perception for nearby sources," in *Computational models of auditory function*, G. S. and M. Slaney, Eds.: IOS Press, 2001, 83-96.

[4] H. R. Hirsch, "Perception of the range of a sound source of unknown strength," *J Acoust Soc Am* 43, 373-374, Feb 1968.

[5] J. Molino, "Perceiving the range of a sound source when the direction is known," *J Acoust Soc Am* 53, 1301-1304, May 1973.

[6] A. W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature* 397, 517-520, Feb 1999.

[7] J. M. Speigle and J. M. Loomis, "Auditory distance perception by translating observers," in *Proc. IEEE Symposium on research frontiers in virtual reality* San Jose, CA, 1993, 92-99.

[8] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Processing* 50, 174-188, Feb 2002.

[9] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," Applied Psychology Unit Cambridge, UK, TR 2341, 1988.

[10] W. G. Gardner, "3D Audio and Acoustic Environment Modeling," Wave Arts Inc., Arlington, MA. USA, TR, March 1999.

[11] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech Audio Process* 11, 826-836, Nov 2003.

[12] M. K. Pitt and N. Shephard, "Filtering via simulation: auxiliary particle filters," Journal of the American Statistical Association 94, 590-599, June 1999.

[13] R. S. Woodworth and H. Schlosberg, Experimental Psychology. New York: Holt, Rinehart and Winston, 1962.

[14] D. R. Campbell, K. J. Palomäki, and G. Brown, "A MATLAB simulation of "shoebox" room acoustics for use in research and teaching," Computing and Information Systems J. 9, 2005.