Available online at www.sciencedirect.com
ScienceDirect
Speech Communication xxx (2010) xxx–xxx
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners

Yan-Chen Lu [a,*], Martin Cooke [b,c]

[a] Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK
[b] Ikerbasque (Basque Foundation for Science), 48011 Bilbao, Spain
[c] Language and Speech Laboratory, Faculty of Letters, University of the Basque Country, Spain

Abstract

Localisation in azimuth and distance of sound sources such as speech is an important ability for both human and artificial listeners. While progress has been made, particularly for azimuth estimation, most work has been directed at the special case of static listeners and static sound sources. Although dynamic sound sources create their own localisation challenges such as motion blur, moving listeners have the potential to exploit additional cues not available in the static situation. An example is motion parallax, based on a sequence of azimuth estimates, which can be used to triangulate sound source location. The current study examines what types of listener (or sensor) motion are beneficial for localisation. Is any kind of motion useful, or do certain motion trajectories deliver robust estimates rapidly? Eight listener motion strategies and a no motion baseline were tested, including simple approaches such as random walks and motion limited to head rotations only, as well as more sophisticated strategies designed to maximise the amount of new information available at each time step or to minimise the overall estimate uncertainty. Sequential integration of estimates was achieved using a particle filtering framework. Evaluations, performed in a simulated acoustic environment with single sources under both anechoic and reverberant conditions, demonstrated that two strategies were particularly effective for localisation. The first was simply to move towards the most likely source location, which is beneficial in increasing signal-to-noise ratio, particularly in reverberant conditions. The other high performing approach was based on moving in the direction which led to the largest reduction in the uncertainty of the location estimate. Both strategies achieved estimation errors nearly an order of magnitude less than those obtainable with a static approach, demonstrating the power of motion-based cues to sound source localisation.

Keywords: Active hearing; Sound source localisation; Interaural time difference; Motion parallax; Particle filtering

1. Introduction

Listeners routinely solve problems of navigation, object identification and avoidance in challenging environments with an efficiency which disguises the true complexity of the task. Localisation of sounds in space is particularly critical. Anyone who has ridden a bicycle on a busy road will appreciate that the ability to rapidly and robustly locate sounds source of interest is of great everyday importance. In general, knowing *where* to listen can improve speech intelligibility in the presence of other sound sources (Kidd et al., 2005).

Most work on understanding and modelling human sound localisation has taken place in a setting where the sensors and sources of interest are assumed to be static. While there are a limited number of situations, such as recording studios or human–machine spoken interaction with headset microphones and headphones, where this assumption is approximately true, there are also many contexts where listeners and/or sound sources are mobile. Further, certain types of application, such as hearing aid processors or other forms of wearable audio (Sawhney

* Corresponding author.
E-mail addresses: y.c.lu@dcs.shef.ac.uk, luyanchen@gmail.com (Y.-C. Lu).

0167-6393/$ - see front matter © 2010 Published by Elsevier B.V.
doi:10.1016/j.specom.2010.06.001

and Schmandt, 2000; Lukowicz et al., 2004) presuppose mobile sensors. While these active scenarios require more sophisticated processing to handle issues such as auditory motion blur and the need to track dynamic sources, they also contain opportunities to exploit cues not available to static listeners. It is well-known that head movements can be used to resolve front-back ambiguities in localisation (Wallach, 1940; Thurlow et al., 1967; Mackensen, 2004) and other studies (e.g. Speigle and Loomis, 1993; Ashmead et al., 1995) suggest that body motion helps in distance estimation. There are many other ways in which motion might help in audition (Cooke et al., 2008). For example, moving towards a source can improve the ratio of direct to reverberant energy. Head rotation can locate the target source in the frontal plane where spatial resolution is at its finest. Movement away from hard surfaces can reduce the effect of reverberation. The head and body can attenuate intense competing sources by a significant amount. Further, a target sound object is easier to segregate from omni-directional reverberation by exploiting its relative response to body motion (Martinson and Schultz, 2006).

Engineering solutions to the localisation problem typically exploit microphone arrays containing $N > 2$ sensors. However, it is of interest to explore binaural/stereo approaches, not only because there is a wealth of existing behavioural data and models of binaural processing to draw upon, but also to test predictions about the value of particular motion strategies emerging from the model in listeners. For some applications involving wearable audio such as personal audio diaries, retaining the link to listener behaviour may be important in making sense of data collected. In any case, one useful way to think about mobile binaural audio is that it provides N binaural *asynchronous* sensors. Even for mobile sources, the additional data provided by sampling at distinct spatial locations can be exploited to improve localisation, as we show in this article.

One intriguing question is: what kind of motion is beneficial for sound source localisation? Does arbitrary motion help in triangulating sources, or are certain trajectories optimal? Answering this question will help in the design of motion plans for mobile robots equipped with audio sensors and may lead to predictions about human movement strategies in adverse conditions. The primary purpose of the work described in this paper is to evaluate the performance of distinct motion strategies, ranging from simple approaches limited solely to head movement, to more sophisticated motions based on information-theoretic criteria such as moving in the direction which maximises estimate entropy. Of interest are those strategies which lead to robust estimates of source location, for both static and moving sources, and which also converge rapidly to a good solution.

Full source localisation involves estimating source azimuth, elevation and distance relative to the listener. Of these, azimuth and distance are of most practical relevance to human listeners. Location in azimuth has been thoroughly investigated (Jeffress, 1948; Blauert, 1997) and a number of computational models exist which produce levels of performance similar to listeners (Lindemann, 1986; Bodden, 1993; Gaik, 1993). Distance has received far less attention (Zahorik et al., 2005). Here, we introduce procedures for estimating both azimuth and distance.

The approach taken in the current study is depicted in Fig. 1. Left and right ear signals are derived from room simulations (anechoic and with mild and moderate reverberation). These signals are processed by an auditory filterbank and successive cross-correlations performed to allow estimation of sound source location in both azimuth and distance through triangulation. Microphone motion enables triangulation for sound distance by sequentially integrating azimuth observations obtained from binaural input (Lu et al., 2007) or array input (Sasaki et al., 2006). A range of location estimates are maintained with a sequential Bayesian particle filtering framework, and a given motion strategy is applied. A specific motion is chosen and updated binaural signals are computed based on the new location. The process cycles in this manner during the "walk".

Section 2 describes the source tracking framework and the extraction of cues used as the basis for sound source localisation. The particle filtering architecture is introduced in Section 3 together with an extension for moving listeners. Section 4 details the different motion strategies, or *strategic walks*, evaluated in the current study. A subset of walks is based on *motion entropy*, which is proposed as a measure of the uncertainty associated with the sound source location in response to listener movement. The outcome of an extensive evaluation of different motion strategies is presented in Section 5.

2. Source localisation in distance and azimuth

2.1. Source tracking framework

The current study addresses localisation in azimuth and distance of a single source, which may be in motion, based on binaural inputs received by an artificial listener, also potentially in motion. Source location x_t at time t is defined in terms of distance and azimuth components, $x_{r,t}$ and $x_{\phi,t}$ specified in a listener-centred spherical coordinate system whose 2D transverse plane passes through the ears:

$$x_t = [x_{r,t}, x_{\phi,t}]. \quad (1)$$

An output variable of the state x_t is denoted y_t and is assumed to be described by the output equation:

$$y_t = g(x_t, v_t), \quad (2)$$

where v_t represents the combined effect of distortions due to reverberation and noise. The unknown and possibly non-linear function g links the state to the noisy measurement. The observation of the underlying state x_t, i.e. the current source's location in distance and azimuth, is derived from the left and right binaural signals at time t,

Please cite this article in press as: Lu, Y.-C., Cooke, M., Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners, Speech Comm. (2010), doi:10.1016/j.specom.2010.06.001

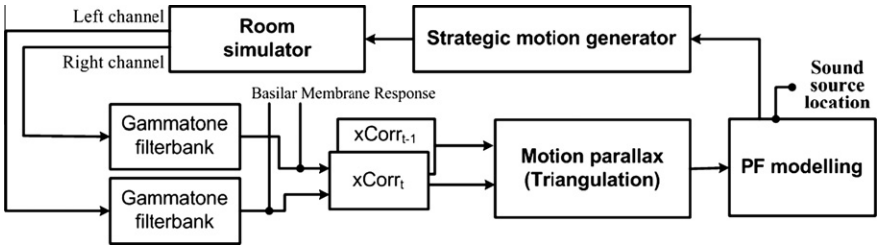

Fig. 1. Computational model for evaluating simulated listener walk strategies in sound source localisation.

$Y_{l,t}$ and $Y_{r,t}$ as described in the following sections. The state-space approach attempts to estimate and track source location using the sequence of noisy measurements, based on the assumption that sound source dynamics can be modelled as a Markov process:

$$x_t = f(x_{t-1}, u_t), \tag{3}$$

where f is the dynamic model describing the relation between current and previous states and incorporating a noise term u_t. Both f and g can be constructed in probabilistic forms to build up a state-space model where the states are represented in terms of random variables.

State estimates are updated with each new observation. The process operates like a recursive filter which avoids the memory requirements of storing complete data sets. Each observation is processed in two stages: prediction and update. The prediction stage uses a model of source dynamics to predict how states evolve in time. It alters the prior probability density functions (PDFs) of state variables depending on an analytical transfer function learned off-line. Noise parameters (e.g. u_t) approximate the effect of unknown disturbances. The update stage integrates instantaneous estimates of source location into the predicted state variable PDFs. These operations follow the principle of Bayesian filtering (West and Harrison, 1997), under which prior information about the target state is updated in accordance with information from new observations.

2.2. Computation of azimuth observations from interaural time differences

The most common approach to azimuth estimation uses stereo inputs to compute interaural time differences (ITD). Inspired by Jeffress' coincidence detection model (Jeffress, 1948), ITDs are normally computed from the cross-correlation of the two channels (Knapp and Carter, 1976), with the ITD determined by the location of the maximum cross-correlation along the delay line. However, the performance of ITD estimation decreases with increasing amounts of room reverberation due to spurious peaks (Champagne et al., 1996), in spite of attempts to compensate for reverberation using, for example, the phase transform (Knapp and Carter, 1976). Recent techniques to tackle the effect of reverberation filter coincident peaks based on their temporal consistency (Brandstein, 1997; Brandstein and Silverman, 1997; Faller and Merimaa, 2004; Jan and Flanagan, 1996; Rui and Florencio, 2004; Ward et al., 2003). A recent review is provided in (Chen et al., 2006).

The cross-correlation of left and right channel signals $Y_{l,t}$ and $Y_{r,t}$ is defined by

$$CC(m) = \sum_{t=1}^{T} Y_{l,t} Y_{r,t+m}, \quad m = \left\lfloor \frac{-M+1}{2} \right\rfloor, \ldots, \left\lfloor \frac{M-1}{2} \right\rfloor, \tag{4}$$

where M is the length of the delay line and T defines the frame size in samples over which the cross-correlation is computed. The lag time τ in Eq. (5) that maximises the cross-correlation is used to estimate the interaural time difference

$$\tau = \arg\max_m CC(m). \tag{5}$$

An important modification of the cross-correlation method was proposed by (Knapp and Carter, 1976), and termed "generalized cross-correlation (GCC)". A weighting function was employed to process the cross-power spectrum to improve the time delay estimation by incorporating knowledge of the environment or source. Among a variety of weighting algorithms, the phase transform (PHAT) method is well-known for its effective resistance to reverberation contamination (Wang and Chu, 1997). PHAT operates by pre-processing the filtered binaural signal through whitening the energy distribution across frequency channels, which has the effect of accentuating components of the spectrum with high direct-to-reverberant energy ratios. When a constant frequency weighting is adopted, the GCC becomes a frequency-domain implementation of the cross-correlation method.

Here, ITD estimates were extracted through cross-correlation of equally-weighted outputs of auditory filters modelled using a bank of $N = 32$ gammatone filters (GTF) with centre frequencies equally spaced on an ERB-rate scale between 50 and 8000 Hz (Patterson et al., 1988). In the GTF-GCC, pairs of cochlear filters are applied to the binaural input to generate filtered binaural signals, $Y_{l,t}(f)$ and $Y_{r,t}(f)$, rather than through the use of the discrete Fourier transformation, to better approximate auditory frequency resolution. The cross-correlation function is modified to include summation across frequency channels:

$$CC(m) = \sum_{f=1}^{32} \sum_{t=1}^{T} Y_{l,t}(f) Y_{r,t+m}(f). \tag{6}$$

In certain circumstances, such as robot motor control, source lateralisation information is more easily handled in terms of azimuth angle rather than ITD. In our case, where the source location is encoded in polar coordinates, an azimuth–ITD transformation and its inverse are required. Given the relationship observed between ITD and azimuth from head-related transfer function (HRTF) data, it is possible to generate an azimuth–ITD transformation based on table lookup (Viste and Evangelista, 2004). However, for the current study, it was necessary to check whether such a function is sensitive to source distance. Binaural stimuli associated with sources distributed uniformly in distance and azimuth were synthesized using a room simulator (see Section 5.1) in a variety of reverberation configurations. ITDs were estimated for these stimuli, and based on the known azimuths, an average ITD curve as a function of azimuth was computed. No significant effect of source distance was found, so a distance-independent azimuth–ITD transformation was considered adequate for the work described here.

2.3. Distance estimation from motion parallax

A listener's body translation produces changing azimuthal measurements with respect to a static source, inducing a potential cue to source distance based on *motion parallax*. Speigle and Loomis (1993) tested the roles of dynamic auditory cues on listener distance judgement when motion parallax was available. The results in a quiet outdoor space demonstrated a small advantage from listener motion. It is possible that larger benefits from motion are to be had in more complex environments.

Assuming that ego-motion S is known or can be estimated, triangulation enables source distance to be quantified, as illustrated in Fig. 2. The sound source undergoes a change in azimuth (from $x_{\phi,t-1}$ to $x_{\phi,t}$) as the listener translates through distance S. For static sources it is possible to use $x_{\phi,t-1}$, $x_{\phi,t}$ and S to triangulate the source distance $x_{r,t}$. For moving sources, some error in distance estimation (e.g. Δ_e in Fig. 2) will occur. Consequently, an approach which maintains a set of location estimations is employed, as described in the next section.

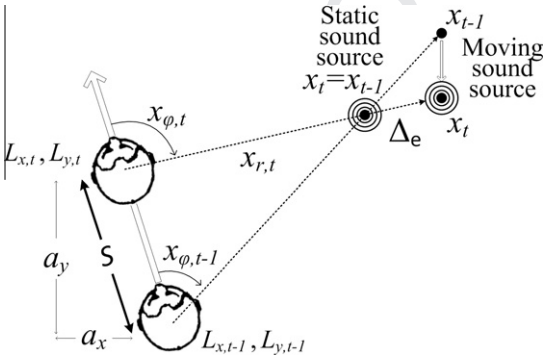

Fig. 2. Motion parallax cue to source distance. A moving sound source sharing the same set of $x_{\phi,t-1}$ and $x_{\phi,t}$ with a static source may introduce an error Δ_e in the distance estimate $x_{r,t}$. The ego-motion S in Section 4 is described in terms of a_x and a_y in Cartesian coordinates for ease of representation.

3. Particle filtering framework for localisation

3.1. Particle filtering

To address the tracking problem in the state-space approach, the target evolution is represented as the time-varying state sequence x_t at discrete points in time, $t = 1, 2, \ldots$, as in Eqs. (1) and (3). A new observation y_t is received at each time t in the form of a noisy measurement of the state x_t as in Eq. (2). The *Bayesian filtering approach* (West and Harrison, 1997) combines all the observations $y_{1:t}$ up to time t to arrive at a target estimate $\hat{x}_t$. Particle filtering (PF) provides a powerful sampling-based approximation to the Bayesian approach. Particles are state samples representing hypotheses about the target space. A set of randomly sampled particles with corresponding importance weights approximate the posterior PDF of the tracking problem. The target (here, the source location) is represented as a set of discrete particles and their weights.

A family of PFs has been developed (Arulampalam et al., 2002). The sampling importance resampling (SIR) filter is the simplest variant among all PF methods (Gordon et al., 1993; Del Moral, 1997) and is reviewed here to illustrate the operation of a particle filtering system. Fig. 3 depicts a SIR PF and associated operations.

Each iteration of the particle filtering operation starts with a set of state samples x^n_{t-1}, and associated importance weights ω^n_{t-1}, $n = 1, \ldots, N$, forming an approximation to the desired posterior PDF $p(x_{t-1}|y_{1:t-1})$ at time $t-1$ using

$$p(x_{t-1}|y_{1:t-1}) \approx \sum_{n=1}^{N} \omega^n_{t-1} \delta(x_{t-1} - x^n_{t-1}), \quad (7)$$

where δ is the Dirac delta function. It can be shown that as $N \to \infty$, Eq. (7) approaches the true posterior PDF $p(x_{t-1}|y_{1:t-1})$ (Doucet et al., 2000). The weights are normalised to unity:

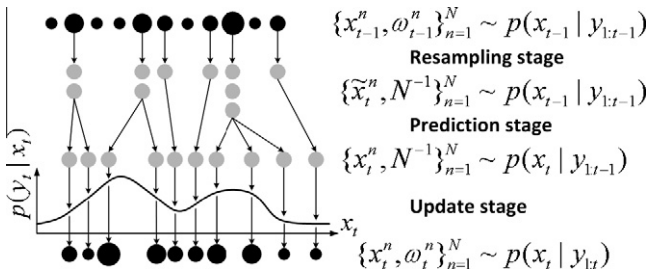

Fig. 3. Representation of a PF iteration with the resampling mechanism. The particles are represented as circles, the size of which denotes the magnitude of corresponding importance weights. The horizontal axis corresponds to the target space. After each iteration, the set of particles and weights is an approximate representation of a specific PDF (noted on the right-hand side). Figure modified from de Freitas et al., 1998.

Please cite this article in press as: Lu, Y.-C., Cooke, M., Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners, Speech Comm. (2010), doi:10.1016/j.specom.2010.06.001

$$\sum_n \omega_t^n = 1. \quad (8)$$

At the prediction stage, each particle is relocated within the target space according to the system dynamics model. In particular, particles are drawn independently according to Eq. (9), moved away from the previous states x_{t-1} conditioned on the transition PDF $p(x_t|x_{t-1})$

$$x_t^n \sim p(x_t|x_{t-1}^n). \quad (9)$$

The prior density $p(x_t|y_{1:t-1})$, Eq. (10), which can be regarded as the predicted version of Eq. (7), is approximated by this new set of particles

$$p(x_t|y_{1:t-1}) \approx \sum_{n=1}^{N} \omega_{t-1}^n \delta(x_t - x_t^n). \quad (10)$$

The update stage renews the posterior PDF $p(x_t|y_{1:t})$ approximated with an updated set of particles and weights as

$$p(x_t|y_{1:t}) \approx \sum_{n=1}^{N} \omega_t^n \delta(x_t - x_t^n), \quad (11)$$

where ω_t^n, the particle importance weight, is modified based on the likelihood function $p(y_t|x_t)$ as defined in Eq. (12), whose parameters are adjusted upon the receipt of the new observation y_t

$$\omega_t^n \propto \omega_{t-1}^n p(y_t|x_t^n). \quad (12)$$

An estimate of the current state $\hat{x}_t$ can then be constructed based on the information derived from the posterior PDF approximation (e.g. mean, median).

An additional resampling stage whose goal is to fine tune the posterior PDF approximation can be applied prior to the prediction stage, as shown in Fig. 3. Particles with low weights (shown as small circles) are eliminated and replaced by the resampling mechanism which usually maintains a proper sample-based representation of the true PDF with all particles being generated with a uniform importance weight, $1/N$. This step attempts to control the variance of the particle importance weights to a sufficiently lower level without compromising sample diversity. The variance may increase over time due to an inaccurate dynamic model, leading to degeneracy, and without the resampling stage a considerable amount of computational power is spent on processing particles with low importance weights and whose contributions to the approximated posterior PDF are of low significance (Doucet et al., 2000).

The resampling stage is better for being activated at the iteration when the effective sample size N_{eff} (given by Eq. (13)) goes beneath a particular threshold value. This selection mechanism prevents a situation known as "sample impoverishment" (Gordon et al., 1993) from developing, in which all particles are compacted to a very limited region

$$N_{eff} = 1/\sum_{n=1}^{N} (\omega_t^n)^2. \quad (13)$$

A range of resampling schemes exist to counteract degeneracy (Douc and Cappe, 2005; Kitagawa, 1996; Liu and Chen, 1998). The systematic resampling scheme (Kitagawa, 1996) is adopted here due to its simplicity. Note that system state estimation should be computed before the resampling stage due to the introduction of irrelevant random variation to particles.

3.2. Auxiliary sampling importance resampling filter

Auxiliary sampling importance resampling (ASIR) PF is a variant of SIR PF which aims to improve the prediction of target dynamics. At the prediction stage, particles of a SIR PF are relocated based only on the previous states x_{t-1} as defined in Eq. (9). This can lead to the problem of bias in Monte Carlo simulations if a poor dynamic model is used, since certain parts of the target space with high posterior likelihood are omitted by poor particle placement. The ASIR PF was developed in an attempt to circumvent this problem with the aid of the observation generated immediately afterwards (Pitt and Shephard, 1999). In this way, particle redistribution during resampling is dependent on both the current observation (through processing over x_{t-1}) and the observation available at the next time frame. The aim is to produce a particle distribution prior to the update stage which is better sampling the true state.

Fig. 4 illustrates the differences between the SIR and ASIR PFs. The figure also shows how the resampling stage operates for processing observations corresponding to two successive frames. One additional pair of prediction and update stages (in filled boxes) conditioned on later observations is employed in the ASIR PF. The last stage of the SIR PF, resampling, becomes the first stage of the ASIR PF iteration, where the iteration boundary is defined to synchronise with the receipt of observations. The resampling stage of a SIR PF removes particles with low importance weights ω_{t-1}. In the ASIR PF, weak particles are defined as those with low auxiliary weights ξ_t. The auxiliary weight ξ is updated based on observation y_t as Eq. (15) with respect to particular particle hypothesis estimates μ. Rewriting Eq. (9) as Eq. (14), μ is specified as the predicted

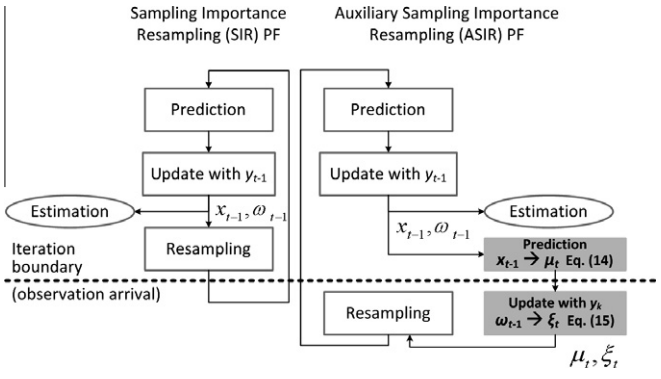

Fig. 4. Comparison of a SIR PF (left) and an ASIR PF (right).

particle distribution based on the dynamic model used at the earlier prediction stages

$$\mu_t^i \sim p(x_t|x_{t-1}^i), \quad i = 1, \ldots, N, \tag{14}$$

$$\xi_t^i \propto \omega_{t-1}^i p(y_t|\mu_t^i). \tag{15}$$

Fig. 5 illustrates synchronisation of the ASIR PF for prediction, update and resampling stages among iterations. The auxiliary weight generation block replaces the pair of prediction and update stages shown in filled boxes in Fig. 4. Unlike the SIR PF, the resampling stage applies at every iteration, despite the possibility of suffering from the aforementioned "sample impoverishment". Resampled particles conditioned on the auxiliary weights are input to the regular prediction stage to encourage less variant particle importance weights than for the SIR PF.

The auxiliary density $p(y_t|\mu_t^i)$ in Eq. (15) for generating auxiliary weights in the ASIR PF also involves calculating the ASIR importance weights as Eq. (16), used primarily for target state estimation. Eq. (16) is modified from the SIR importance weight update, Eq. (12), by substituting the product with ω_{t-1}^n into the division over the auxiliary density, which is an approximation to the true PDF (Pitt and Shephard, 1999)

$$\omega_t^n \propto p(y_t|x_t^n)/p(y_t|\mu_t^{i^n}), \quad n = 1, \ldots, N, \tag{16}$$

where i^n is the index of the parent of particle n during the resampling process. This indexing value retains the pairing information of particles before and after resampling to help locate the correct auxiliary density in Eq. (15).

3.3. Model of source dynamics

Experiments have demonstrated that PF is efficient in tracking sound sources in a realistic reverberant environment (Asoh et al., 2004; Ward et al., 2003) for static sensors. PF is also considered useful for providing continuous tracking of sources with short periods of silence, in contrast to techniques which use the local running average of spatial cues (Aarabi, 2002). For the current study, motion of both the target source and the listener contribute to the observed dynamics. Source location hypotheses of all particles are relocated according to the inverse of listener movement in the PF prediction stage. Noise terms are additionally applied to particles to effect an additional dispersion in both azimuth and distance. Specifically, the azimuth coordinate $x_{\phi,t}^n$ of the state variable (i.e. hypothesised source location) corresponding to particle n at time t is determined by

$$x_{\phi,t}^n = x_{\phi,t-1}^n + u_{\phi,t}^n - l_{\phi,t}^n, \tag{17}$$

where the development of source location $x_{\phi,t}^n$ is formulated as an increment to the previous location $x_{\phi,t-1}^n$ with noise term $u_{\phi,t}^n$ and inverse listener motion in azimuth coordinate $-l_{\phi,t}^n$. As defined in Eq. (18), $u_{\phi,t}^n$ is a zero mean Gaussian variable with a time-variant standard deviation $\sigma_{\phi,t}^n$, which is a function of the particle importance weight $\omega_{\phi,t-1}^n$ and the constant parameter o_ϕ

$$u_{\phi,t}^n \sim N(0, {\sigma_{\phi,t}^n}^2), \quad \sigma_{\phi,t}^n = (1 - \omega_{t-1}^n) \cdot o_\phi. \tag{18}$$

Since $\omega_{\phi,t-1}^n$ ranges between 0 and 1, o_ϕ specifies the maximum value of the standard deviation. The effect is that a weaker particle is more likely to possess a larger noise term, and encourage particle evolution (see Table 3 for the value used). In the same vein, $x_{r,t}^n$ is determined by inverse listener motion in the distance coordinate $-l_{r,t}^n$ and a noise term $u_{r,t}^n$ which is a function of ω_{t-1}^n and o_r:

$$x_{r,t}^n = x_{r,t-1}^n + u_{r,t}^n - l_{r,t}^n. \tag{19}$$

Only the altered hypothesis x_ϕ^n is subsequently evaluated with the observation model (see below) to update the particle weight. After integration with listener motion l_t (i.e. S in Fig. 2) at time t, a highly weighted $x_{\phi,t-1}^n$ may be evaluated as less important due to its inaccurate $x_{r,t-1}^n$, the paired source distance hypothesis, in the context of motion parallax.

3.4. Observation model

An observation model based on ITD determination is used to update x_t. Particle importance weights are altered from the correlation between x_t and y_t, realised in two steps. First, the learned azimuth to ITD function (described in Section 2.2) converts the predicted azimuth observation $x_{\phi,t}^n$, Eq. (17), to an ITD. Second, a GTF-GCC scheme assigns this measure corresponding to particle n an importance weight. Cross-correlation in Eq. (6) can be used directly as the ITD likelihood function for source azimuth, as in the pseudo-likelihood approach of (Ward et al., 2003). The length of the delay line is $M = 65$ samples corresponding to 0.75 ms at the sampling frequency of 44.1 kHz used here.

3.5. Source location estimation

The posterior PDF of x_t is calculated in terms of a collection of N particles x_t^n, ω_t^n as specified in Eq. (16). Based on the posterior PDF, a number of approaches can be used to generate a single estimate $\hat{x}_t$ of the source location. Apart from calculating the estimate to minimise the mean square error (MMSE) as Eq. (20), $\hat{x}_t$ can be directly

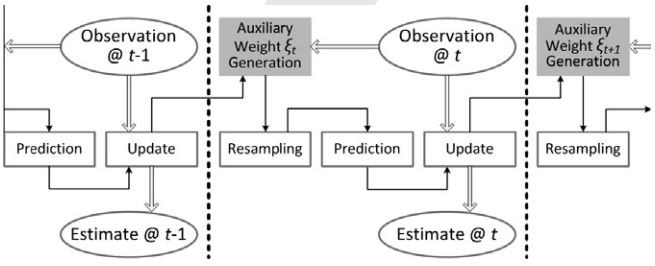

Fig. 5. ASIR PF iterations and the synchronisation between incoming observations and PF stages.

assigned to the "best" particle j characterized by the highest importance weight as the maximum a posteriori (MAP) estimate, (21). Alternatively, $\hat{x}_t$ can be the weighted mean of particles within a robust "window", whose size is specified by $w \in (0,1)$. The robust window includes the particles with importance weights larger than a certain threshold which is the product of w and the weight of the best particle j, Eq. (22). This third approach will be referred as the "robust mean" (Rekleitis, 2003)

$$\hat{x}_t = E\{x_t|y_{1:t}\} = \sum_{n=1}^{N} \omega_t^n x_t^n, \tag{20}$$

$$\hat{x}_t = x_t^{j_t}, j_t = \arg\max_j \omega_t^j, \tag{21}$$

$$\hat{x}_t = \sum_m \omega_t^m x_t^m, \quad m \in \{\omega_t^m \geqslant w\omega_t^{j_t}\}. \tag{22}$$

The best particle hypothesis can introduce a discretization bias due to the coarse sampling over the target space with a limited number of particles. The MMSE estimate spreads the noise of weak particles and is often sensitive to disrupt observations. As the outlier particles are filtered with a robust window, the robust mean is a preferable approach when process measurements are noisy and is adopted in our simulations introduced in Section 5. However, it is also the most computationally expensive on account of the determination of "robust" particles in the robust mean method.

4. Strategic walk design

The previous section demonstrated how source location estimates in azimuth and distance can be sequentially-integrated within the particle filtering framework. The goal of strategic walk design is to determine which listener motion strategies achieve fast and robust localisation. Fig. 1 illustrates how listener motion plans are integrated into the system architecture. A "strategic motion generator" block provides instant-by-instant listener location for the room simulator, generating one position coordinate per frame as part of a single PF iteration.

Six types of systematic motion plans were studied. For ease of description, we revert to (x,y) Cartesian coordinates to characterise motion strategies (see Fig. 2). In addition, listener head orientation h varied from $\pi/2$ and $-\pi/2$. The current listener location is denoted L_t which can be expressed in terms of the previous listener location L_{t-1} and current action a_t:

$$L_{x,t} = L_{x,t-1} + a_{x,t}, \tag{23}$$

$$L_{y,t} = L_{y,t-1} + a_{y,t}, \tag{24}$$

$$L_{h,t} = L_{h,t-1} + a_{h,t} (\mathrm{mod} 2\pi). \tag{25}$$

In the following subsections, strategic motions are introduced in terms of the evolution of a_t and classified into the following categories: (1) no motion, (2) head rotation only, (3) random walk, (4) smooth random walk, (5) walk towards estimated source location and (6) strategic entropy walk. The last category contains four variations, leading to a total of nine strategic motions. While a standard uniform distribution random generator is employed for generating the strategic motions of the second to fourth category, the instantaneous PF state parameters determine those of the last two categories. In other words, only the listener walks of non-random motion strategies are affected by the PF estimate for sound source location. Examples of walk trajectories are illustrated in Section 5.

4.1. No motion (NM)

In this baseline condition, the listener remains static with zero $a(t)$ by facing towards zero azimuth. No dynamic localisation cue is available.

4.2. Head rotation only (HRO)

This strategy involves no listener displacement and therefore the absence of motion parallax cues[1] for distance estimation. The direction (clockwise or anti-clockwise) and size (up to $\pi/2$) of head rotation was determined randomly according to

$$a_{h,t} = \pi/2 - R_h \cdot \pi, \quad R_h \sim U(0,1), \tag{26}$$

where $U(0,1)$ indicates a uniformly-distributed random variable taking values between 0 and 1. Azimuth localisation can benefit from this kind of motion by resolving front-back ambiguity (Wallach, 1940). However, localisation in distance is nearly equivalent to random guessing since both static and dynamic cues to distance are lacking.

4.3. Random walk (RW)

This strategy allows both random head rotation and body movement and tests whether motion-induced cues are beneficial regardless of direction or smoothness of trajectory. A random head rotation produces the next orientation, Eq. (26). Decoupled movements in the (x,y) coordinates are determined by Eqs. (27) and (28) respectively

$$a_{x,t} = (1 - 2 \cdot R_x) \cdot v \cdot \delta t, \quad R_x \sim U(0,1), \tag{27}$$

$$a_{y,t} = (1 - 2 \cdot R_y) \cdot v \cdot \delta t, \quad R_y \sim U(0,1). \tag{28}$$

The maximum number of displacements in (x,y) coordinates are independently specified by model parameters for velocity v and frame duration δt.

4.4. Smooth random walk (SRW)

A smoothed version of random walk is proposed to create a more natural walking trajectory. The strategy is

[1] The head and ears have the same centre of rotation in our simulation, although real head movements may induce small motion parallaxs due to the centre of rotation being offset from the interaural axis.

implemented simply by limiting the maximum allowed head rotation change between time steps, reduced from $\pi/2$ to $\pi/9$:

$$a_h(t) = \begin{cases} -\pi/9, & R < 1/3, \\ \pi/9, & R \geqslant 2/3,\ R \sim U(0,1), \\ 0, & \text{otherwise.} \end{cases} \quad (29)$$

After head orientation is determined, a forwarding action is subsequently taken as defined by

$$a_{x,t} = \sin[a_{h,t}] \cdot v \cdot \delta t, \quad (30)$$

$$a_{y,t} = \cos[a_{h,t}] \cdot v \cdot \delta t \quad (31)$$

generating (x, y) displacements whose sizes are specified by the product of velocity and frame duration. The resulting motion is more like real human movement and leads to a greater exploration of the room space compared to an unsmoothed random walk (cf. Fig. 9).

To avoid non-physical motion across the room boundary, head rotation is re-randomised[2] until listener motion stays inside the room boundary.[3]

4.5. Walk towards estimated source location (W2S)

Here, the current source location estimate $\hat{\phi}_{t-1}$ from particle filtering determines the next head displacement

$$a_h = \hat{\phi}_{t-1}. \quad (32)$$

Walking towards the source is of potential benefit for two reasons. First, placing the source in front of the listener enhances azimuth perception due to a finer resolution of cross-correlation in the frontal plane. Additionally, the subsequent forward motion decreases listener–source distance, leading to a possible improvement in SNR. However, a listener might approach the wrong direction due to a biased or incorrect estimate of the time-varying source location. No prediction of source motion is applied here.

4.6. Strategic entropy walk

A number of listener motion strategies were derived based on the information obtainable with a single "look-ahead" step. While such strategies may not be used in practice by human listeners (since they would require a rather unnatural "stop-and-search" behaviour), they are of interest since quite different criteria applied to the resulting distribution of motion-induced state-spaces. Specifically, using a quantity we call "motion entropy" (introduced below and defined in more detail in Appendix A), it is possible to contrast motion which leads to the greatest increase in information with motion which increases certainty.

Different "next" motions lead to distinct sets of hypothesised source locations represented in the particle filtering framework, as illustrated in Fig. 6. Informally, motion entropy is a measure of the uncertainty of sound source location associated with a given listener movement, a quantity which can be calculated from the particle set (see Appendix A). In general, one listener motion leads deterministically to a single motion entropy. For example, the resulting distribution of particle importance weights from a forward move (zero head rotation) normally differs from that of a reverse move (rotate $+/-\pi$). By evaluating motion entropy for a number of next steps, the direction which results in greatest uncertainty (maximum motion entropy) or least uncertainty (minimum motion entropy) can be chosen. While the latter may be beneficial for environments where the particle set provides an accurate estimate of source location, it is conceivable that moving in the direction which provides most information gain might be an effective strategy when the source estimate is unclear.

A number of rotation choices defined in Eq. (33) were evaluated in terms of their induced motion entropies

$$a_{h,k,t} = \frac{2\pi}{9} \cdot S \cdot (k - \lfloor (K+1)/2 \rfloor), \quad k = 1, \ldots, K, \quad (33)$$

where K is the number of choices and S determines the variation in magnitude between them. A smaller S leads to a smoother head motion. The subsequent forward motion is determined by

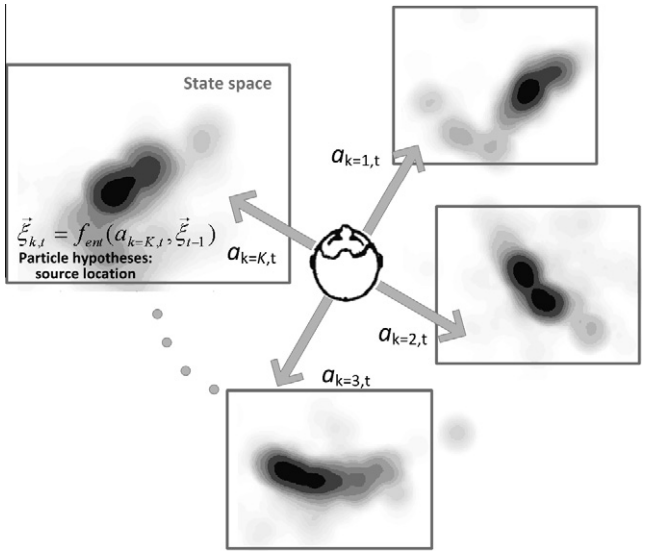

Fig. 6. Source location hypotheses induced by K different listener motions. These K distinguished sets of hypotheses originate from the same prior particle distribution, $\vec{\xi}_{t-1}$. The resultant variation is due to different prospective listener motions $a_{k,t}$. Greyscale clouds represents the distribution of particles as a function of their location in room space, with darker regions indicating denser particle populations.

[2] It is possible that in the restricted head movement case of the smooth random walk, no choice of rotation will lead to a physical motion, in which case Eq. (26) is used instead of Eq. (29).

[3] Room boundary information for strategic motion generator is assumed available. In practice, it could be obtained from another modality (e.g. vision).

$$a_{x,k,t} = \sin[a_{h,k,t}] \cdot v \cdot \delta t, \quad (34)$$

$$a_{y,k,t} = \cos[a_{h,k,t}] \cdot v \cdot \delta t. \quad (35)$$

The selection of either the maximum or the minimum among K motion entropies is associated with two different S values (1 and 0.25). A strategic entropy walk family with four members is accordingly developed as listed in Table 1. $e_{k,t}$ is the motion entropy of kth motion choice at frame t.

5. Evaluation of motion strategies

The eight motion strategies introduced in the previous section were evaluated with respect to their ability to estimate source distance for static and moving sources in simulated anechoic and reverberant conditions.

5.1. Stimuli

Virtual binaural stimuli were synthesized using the "Roomsim" room simulator, with KEMAR head modelling (Campbell et al., 2005). Non-individual HRTFs recorded with a KEMAR dummy head (Gardner and Martin, 1996) were used to generate binaural room impulse responses (BRIRs) based on the image method (Allen and Berkley, 1979) along with parameters such as the listener–source geometry and the absorption properties of reflective surfaces. Training and test stimuli were generated by convolving a speech source with the BRIR. Source and listener motion was created by altering their geometric parameters in the room simulator in each time frame. Subjectively, some motion discontinuity is audible due to the limitations of the image-based method in synthesizing dynamic objects, especially for a moving noise source (Otani and Hirahara, 2007).

An 18 m by 18 m by 10 m gym space was approximated in free-field and two reverberant conditions characterized by T_{60} of 0.2 and 0.7 s. The octave band reverberation times of these two reverberant spaces are shown in Fig. 7 according to the Norris–Eyring reverberation formula (Eyring, 1930)

$$T_{60} = \frac{0.161 \cdot V}{4 \cdot m \cdot V - S \cdot \ln(1-a)}, \quad (36)$$

where V is the room volume, m is the air absorption coefficient representing the relative humidity of air, S is the total room surface area and a is the averaged absorption coefficient over the six reflecting surfaces of the room interior. T_{60} is an average across octave bands.

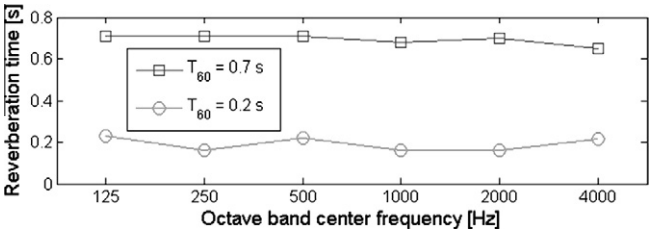

Fig. 7. Octave band reverberation times of the synthetic 18 m by 18 m by 10 m gym space for two reverberant conditions.

Three motion patterns of increasing complexity were generated: static, linear motion and zigzag motion. Example trajectories for each are shown column-wise in Fig. 9.

Using 60 time steps of 0.75 s each, sources with linear motion moved towards the south-east corner of the room in a straight line at a constant velocity (0.14 ms^{-1}). Those sources with zigzag motion moved towards the north-east corner with a variable velocity (ranging from 0 to 0.42 ms^{-1}). Here, the listener–source distances of all stimuli are greater than 2 m, so no near-field distance cues are available (Zahorik et al., 2005). Table 2 lists the Roomsim parameters common to all simulations.

5.2. Localisation accuracy

The Euclidean distance between the true and estimated source location, as well as the unsigned estimation errors in azimuth and distance, were used to quantify localisation performance. Since blind initialization of particle hypotheses may lead to poor performance during first few iterations in PF based algorithms, a "frame convergence" measure was monitored to avoid adding noise to the distance estimate based on the initial transient. Convergence was defined to occur in the frame for which the Euclidean distance was smaller than the standard deviation of the estimation errors across the entire particle set. The time to converge was different for each stimulus. We used the mean time to convergence in evaluations, computed offline. Using a development set of stimuli, an average time to convergence of 15 frames was measured. The performance figures reported below were thus based on the partial trajectory from the 15th frame to the end. The averaged Euclidean distance (AED) over the converged frames was used to assess the algorithm.

Table 1
Variants of entropy walks in terms of different argument combination of $e_{k,t}$ and S.

MinEnt	$a_{k,t} = \arg\min_k e_{k,t},\ S = 1$
MaxEnt	$a_{k,t} = \arg\max_k e_{k,t},\ S = 1$
SMinEnt	$a_{k,t} = \arg\min_k e_{k,t},\ S = 0.25$
SMaxEnt	$a_{k,t} = \arg\max_k e_{k,t},\ S = 0.25$

Table 2
Roomsim parameters.

Parameter	Value
Temperature	20 °C
Sample-rate	44,100 Hz
Sound speed	342.7 mps
Dummy head inter-ear distance	0.152 m
Relative humidity of air	40%

The PF algorithm is based on a stochastic simulation, so different localisation results, using "robust mean" method as described in Eq. (22), are produced for each run even for the same input data. Evaluation was therefore based on sufficient runs to allow statistical distribution of a particular assessment parameter, e.g. geometric mean of AEDs, to be estimated.

5.3. Simulation settings

Each simulation run involved stimuli lasting 45 s and 60 runs were conducted for each combination of walk strategy, reverberation level and source motion. The resulting 60 AEDs differed from each other due to the randomness inherent in Monte Carlo simulation and listener walk development. Nine walk strategies, three reverberation levels and three source motions led to 72 combinations and a total of 4860 simulations. For the three strategic motions synthesized independent of particle filtering states, namely NM, HRO, RW and SRW, 10 different listener trajectories were generated off-line, and each stimulus was processed six times for a total of 60 simulations in each of nine conditions (three reverberation levels by three source motions). The simulations ran at 10 times real time on a 2 GHz CPU with 2 GB main memory. Approximately 16 days were required to complete all 4860 simulations.

Table 3 lists parameter values used in PF modelling with simulated listener walks (see Appendix A for the details of the last two parameters associated with motion entropy). Values were determined empirically based on pilot simulations that achieved a reasonable run-time and performance.

5.4. Examples of listener walks

Typical examples of walk trajectories for all strategies, except NM, are presented here for the more reverberant condition ($T_{60} = 0.7$ s). Source motion affects the listener walks of non-random strategies via continuous particle evaluation. In contrast, for the other strategies (HRO, RW and SRW) the simulated listener walks show no temporal coherence with source location.

Fig. 8 shows an example of the head rotation only (HRO) strategy. The upper part of the figure displays an 18 m × 18 m × 10 m simulated room space viewed from above. In this example, the source moves towards the south-east corner in a straight line at a constant velocity (0.14 mps). Location accuracy, shown in the lower half of the figure, shows the typical poor performance of this strategy. Here, the mean estimation error is 6.81 m.

The top row of Fig. 9 depicts the typical movements of the random walk (RW) strategy for static and moving sources. The listener walk exhibits no clear pattern and only explores the region near the initial point of departure. Note that for all the simulations discussed in this paper, the listener started from the same point, (3, 1) in (x, y) coordinates, in gym space, allowing for variation of source distance up to 19 m.

The smooth random walk (SRW) strategy is shown in the second row of Fig. 9. Here, limiting the allowed head movement (and hence the change in forward angle) produces a smooth trajectory which is quite distinct from the RW case. SRW exhibits a more rapidly decreasing estima-

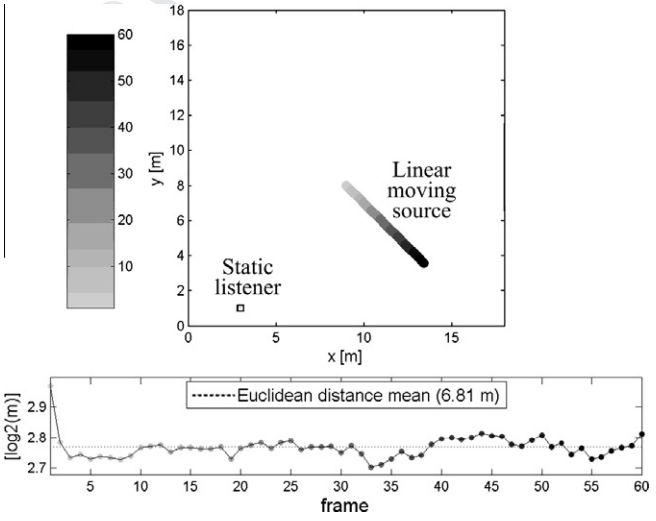

Fig. 8. An example of the head rotation only strategy with a linearly-moving source (plan view). The listener is static and represented by the square in the lower left corner. Grey level is used to indicate time index, ranging from light ($t = 0$ steps) to dark ($t = 60$ steps). The lower plot shows the accuracy of the location estimate as $\log_2$ of the Euclidean distance between the true and estimated positions. The dotted line is the arithmetic average of 60 Euclidean distance values.

Table 3
Parameter settings for strategic listener walk simulations.

Parameter	Value	Description
N	225	Number of particles
M	65	Delay line length (samples) of cross-correlation function
o_ϕ	15	Source dynamic model parameter (°) for azimuth component in Eq. (17)
o_r	2	Source dynamic model parameter (m) for distance component
v	1.0	Simulated source velocity (m/s)
δt	0.75	Frame length (s) in PF simulation
K	7	Number of candidates for strategic entropy walk in Eq. (33)
σ_r	8 ($T_{60} = 0.2$ s)	Standard deviation for the predicted Gaussian function
	5 ($T_{60} = 0.7$ s)	Eq. (39)
σ_s	8	Standard deviation for the weight function, Eq. (43)

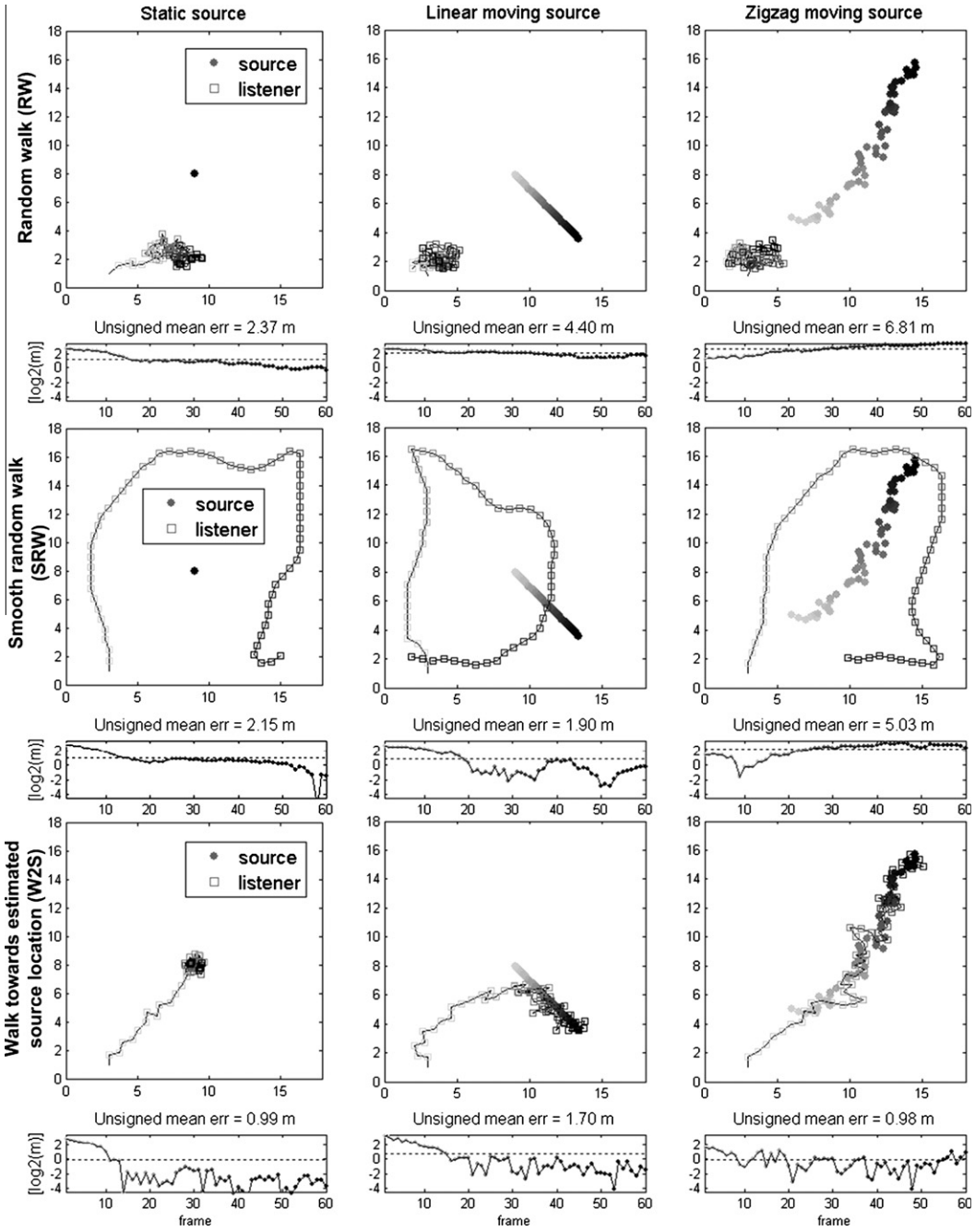

Fig. 9. Examples of motion strategies, random walk (RW), smooth random walk (SRW) and walk towards estimated source location (W2S) for three source dynamics.

tion error through time than RW, suggesting that source localisation benefits from the greater motion parallax in consecutive frames. However, as seen in the third column of first two rows, the error increases with time, suggesting that tracking a zigzag moving source with limited listener movement is challenging.

The walk towards estimated source (W2S) strategy is illustrated in the bottom row of Fig. 9. Clear evidence of listener–source following behaviour is visible, with the listener intercepting the source when the intervening distance becomes small. Estimation error is typically much smaller than in the static and random walk strategies.

Strategies involving motion entropy give rise to the four distinct walking patterns displayed in Fig. 10. It is interesting that MinEnt exhibits a similar walking pattern as W2S in the top panel of Fig. 10. This suggests that a good policy to reduce uncertainty in estimating source location is to approach the source itself. In contrast, a MaxEnt listener aiming to augment information gain tends to keep a constant distance from the source as in the second row of

Please cite this article in press as: Lu, Y.-C., Cooke, M., Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners, Speech Comm. (2010), doi:10.1016/j.specom.2010.06.001

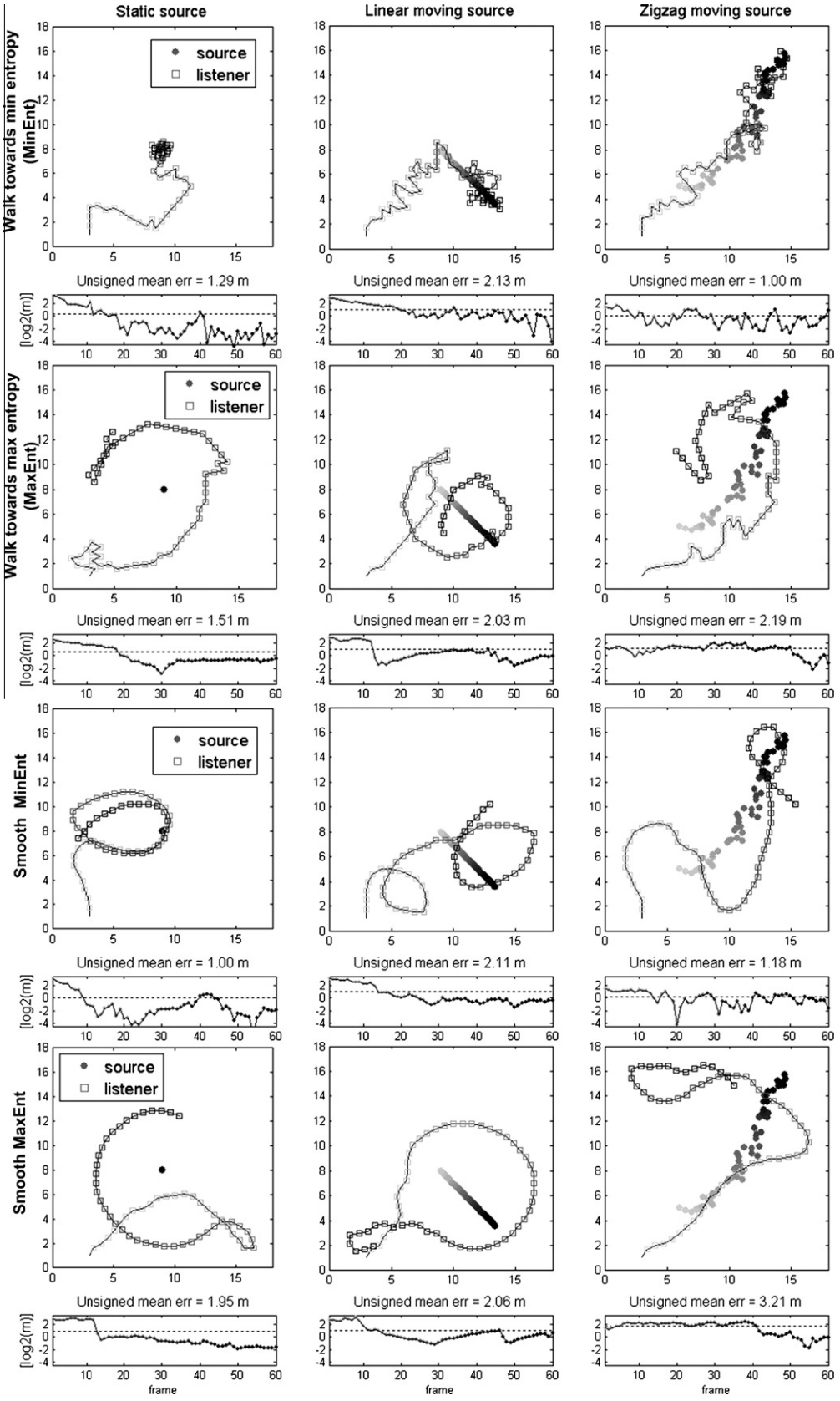

Fig. 10. Motion entropy walk strategies for three source dynamics.

Fig. 10. Reducing the extent of head rotation leads to the smooth versions of MinEnt and MaxEnt, SMinEnt and SMaxEnt. Here, smoothing does not appear beneficial, but instead limits the rapidity with which these "listeners"

can perform adaptations to track moving sources. As a result, SMinEnt transverses the source instead of intercepting it. SMaxEnt revolves around the source while seeking a constant changing rate of azimuth. MaxEnt and SMaxEnt both exhibit "source-orbiting" behaviours, with the former showing rapid directional changes at certain points, particularly for the zigzag source motion.

In these examples, W2S and MinEnt delivered relatively fast convergence of PF estimation as well as greater precision. The gain from reducing the effect of reverberation relative to the direct signal is probably responsible for the better performance by improving azimuth cues since MaxEnt appears to have at least the same, if not greater, opportunity to acquire useful motion parallax information as W2S and MinEnt.

Estimation errors of MinEnt and SMinEnt varied more abruptly than MaxEnt and SMaxEnt. The errors in the azimuth component while intercepting the source may lead to rapid changes in estimation error. This *source-intercepting* effect will be illustrated in more details below.

Another key factor degrading localisation performance is the distortion in cues due to reverberation, leading to a biased likelihood function. This effect is particularly found for motion strategies such as MaxEnt and SMaxEnt which result in a low ratio of direct to reverberant energy by maintaining a certain listener–source distance. Particles which are evaluated with relatively higher importance weights by the biased likelihood function are temporarily favoured by the "robust mean" method (originally designed for outlier filtering) in the PF estimation stage and lead to an increase in estimation error.

Additionally, an incorrect model of source dynamics may lead to biased particle weighting and erroneous estimation. For example, when tracking zigzagging sources, particles cannot always be redistributed to the right place (the peak of the likelihood function) via the prediction function based on the noise term only. An incorrect posterior PDF for source location is thereby obtained by the loss of particles in the high likelihood region, which amounts to a failure to represent the importance density. If biased particle weighting, either from distorted cues or an improper dynamic model, occurs along with the degeneracy phenomenon (all but one particle will have negligible weight), a PF resampling operation may trigger a long term performance decline. As a result, most particles are duplicated (reborn) around the region with low importance after resampling and herding particles again towards the high likelihood region may require a significant amount of time. A possible solution to degeneracy is to offset the likelihood function from zero so as to reduce particle weight diversity. Nevertheless, the PF convergence rate may suffer as a consequence of emphasising least important particles.

5.5. Results

Table 4 lists the simulation results in terms of average error in Euclidean distance (AED), as means across the three reverberant conditions, for the three motion types. The no motion (NM) and head rotation only (HRO) approaches resulted in very poor performance, while random motion (i.e. RW) almost halved the estimation error. This demonstrates that motion in itself is beneficial for source localisation. Smoothing the random walk produced a further halving of error, suggesting that motion through a larger portion of space can help. However, the best strategy overall was to walk towards the source (W2S), which led to more than an eightfold decrease in estimation error relative to the static strategy (i.e. NM and HRO). Of the entropy-based approaches, MinEnt outperformed MaxEnt, and smoothing was not beneficial. The overall difference between W2S and MinEnt, though small, was statistically significant ($p < 0.01$). However, the advantage was largely due to the superior performance of W2S for sources with the more complex zigzag motion.

Table 4
Average Euclidean distance errors (in m) for each combination of source motion and listener walk. Results are averages across the three reverberation conditions ($T_{60} = 0.0$ s, 0.2 s and 0.7 s). The overall average is calculated across source motions and reverberation conditions.

	Static	Linear motion	Zigzag motion	Overall
NM	8.20	7.47	12.52	9.40
HRO	6.27	6.54	10.58	7.80
RW	1.51	2.97	5.87	3.45
SRW	0.96	1.20	3.62	1.93
W2S	0.79	0.89	1.08	0.92
MinEnt	0.83	0.96	1.22	1.00
MaxEnt	0.78	1.57	3.00	1.78
SMinEnt	0.94	1.35	2.33	1.54
SMaxEnt	1.07	2.39	2.92	2.13

Fig. 11 presents averaged Euclidean distance errors for both reverberation and source motion factors. A three-way ANOVA with factors of source motion, reverberation and strategy confirmed main effects of strategy [$F(7,4248) = 5566$, $p < 0.001$], motion [$F(2,4248) = 1835$, $p < 0.001$] and reverberation [$F(2,4248) = 356$, $p < 0.001$] together with significant pairwise and three-way interactions. Inspection of Fig. 11 suggests that increased motion complexity led to a reduction in performance for most strategies apart from NM and HRO, which showed no or little degradation in going from a static to a linearly-moving source. The presence and strength of reverberation was detrimental for the localisation performance of most strategies, although little effect of the $T_{60} = 0.2$ s condition over the anechoic case was seen for several types of motion.

The effect of a better direct-to-reverberant energy ratio achieved by travelling behind the source may explain the low estimation errors delivered by W2S and MinEnt, particularly for the more challenging conditions (high reverberation and complex source motion). However, it is noteworthy that other strategies delivered better estimates in the less challenging conditions. One possibility is that motion towards a source (exhibited both directly by W2S and indirectly by MinEnt) leads to less robust triangulation

Please cite this article in press as: Lu, Y.-C., Cooke, M., Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners, Speech Comm. (2010), doi:10.1016/j.specom.2010.06.001

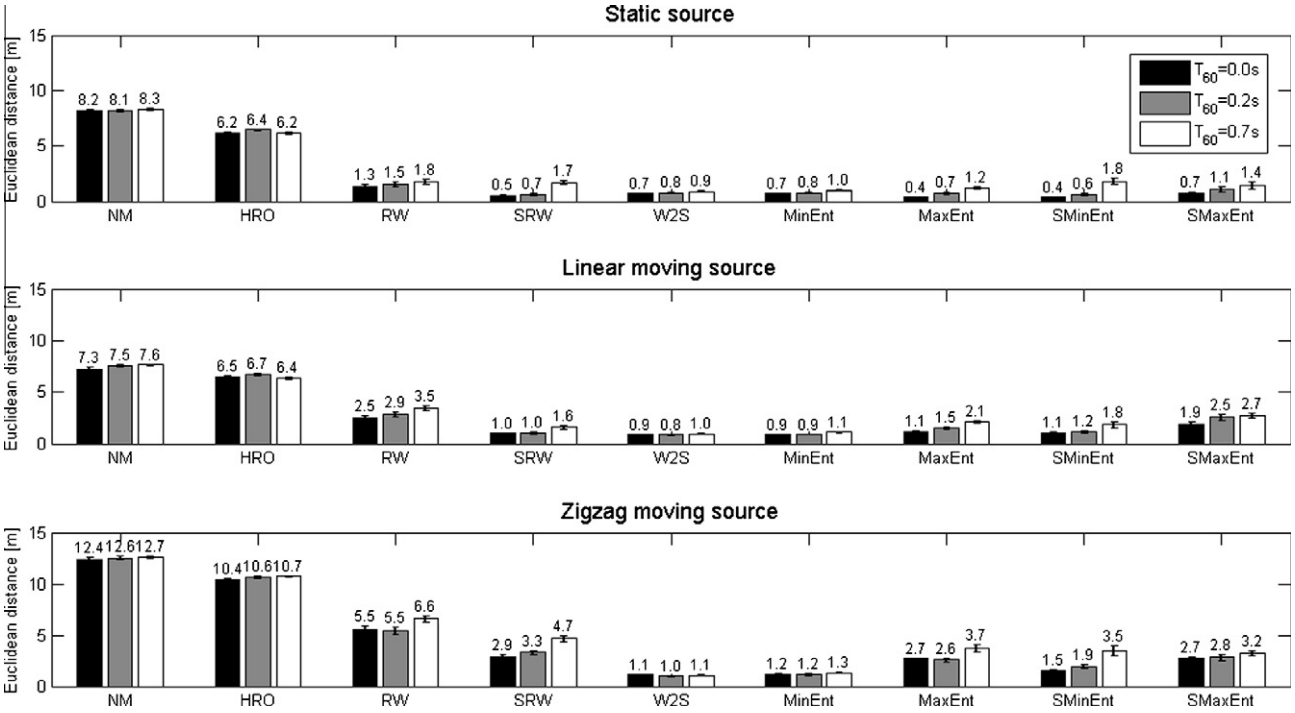

Fig. 11. Localisation performance in 81 conditions (nine walk strategies × three reverberation levels × three source motions). Values above bars show mean *AED* over 60 simulations. Here and elsewhere error bars indicate 95% confidence intervals.

via motion parallax cues since the change in azimuth in successive steps approaches zero. Motion towards the estimated source location may involve a tradeoff between better suppression of reverberant components but reduced efficacy of motion parallax cues. The superior performance of other strategies in anechoic conditions (e.g. SRW, MaxEnt and SMinEnt outperformed W2S for static sources shown in Fig. 12) where no benefit is obtained by reverberation suppression supports the existence of such a tradeoff, and suggests that a hybrid strategy of walking towards the source but maintaining sufficient lateral excursions to produce useful a motion parallax may outperform a pure W2S approach. As a control, a condition where the walk was towards the exact source location produced distance estimation errors an order of magnitude larger than walking towards the estimated source. This outcome is due to an inability to benefit from motion parallax cues and demonstrates their critical role in these simulations.

It is instructive to examine the contribution of errors in azimuth and distance separately to the overall Euclidean distance error (Figs. 12 and 13). From these plots, it is abundantly clear that the AED measure is dominated by the distance component. A counter-intuitive picture emerges for the azimuth component, where the smallest errors are associated with HRO and two random walk strategies, and the poorest estimates come from NM and the two strategies which are best overall. The limited effect of azimuth errors on overall Euclidean distance is due to the fact that the contribution of azimuth estimation is bounded by twice the error in distance, and distance errors are typically small after convergence. Since the upper bound on Euclidean distance error is a function of distance, errors in distance estimation have a great effect.

The contrastingly poor azimuth estimates may be due to the source-interception pattern of behaviour shown by the strategies involving body translations. This effect is illustrated in Fig. 14, which shows what happens to the azimuth estimate when the source is about to be intercepted by the SMinEnt listener at around the 35th frame. At this point there is a sudden increase in azimuth estimation error with little effect on distance estimation, as shown in the left part of Fig. 14. The azimuth error deceases rapidly as the simulated listener moves away from the source. The frequency of source-intercepting also increases with an increase in source motion complexity (this behaviour pattern can be seen for zigzag motion in Figs. 9 and 10). At the PF prediction stage, particles are shifted as a function of current listener motion. Unless the true source location has been probed with a sufficient number of particles after the shift, a certain estimation error may come about due to the loss of diversity among the particles. When the listener intercepts the source, it may develop quite a considerable error in azimuth in some circumstances as illustrated in the right part of Fig. 14. From the figure, one possibility to account for an error increase is the scarcity of particles in the shaded area HL which is the highly likely part of the target space at the 35th frame. On the other hand, particles falling in the area FB are evaluated with higher weights by the likelihood function, and the obtained estimate is slightly right-shifted from that of the 34th frame. The area FB mir-

Please cite this article in press as: Lu, Y.-C., Cooke, M., Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners, Speech Comm. (2010), doi:10.1016/j.specom.2010.06.001

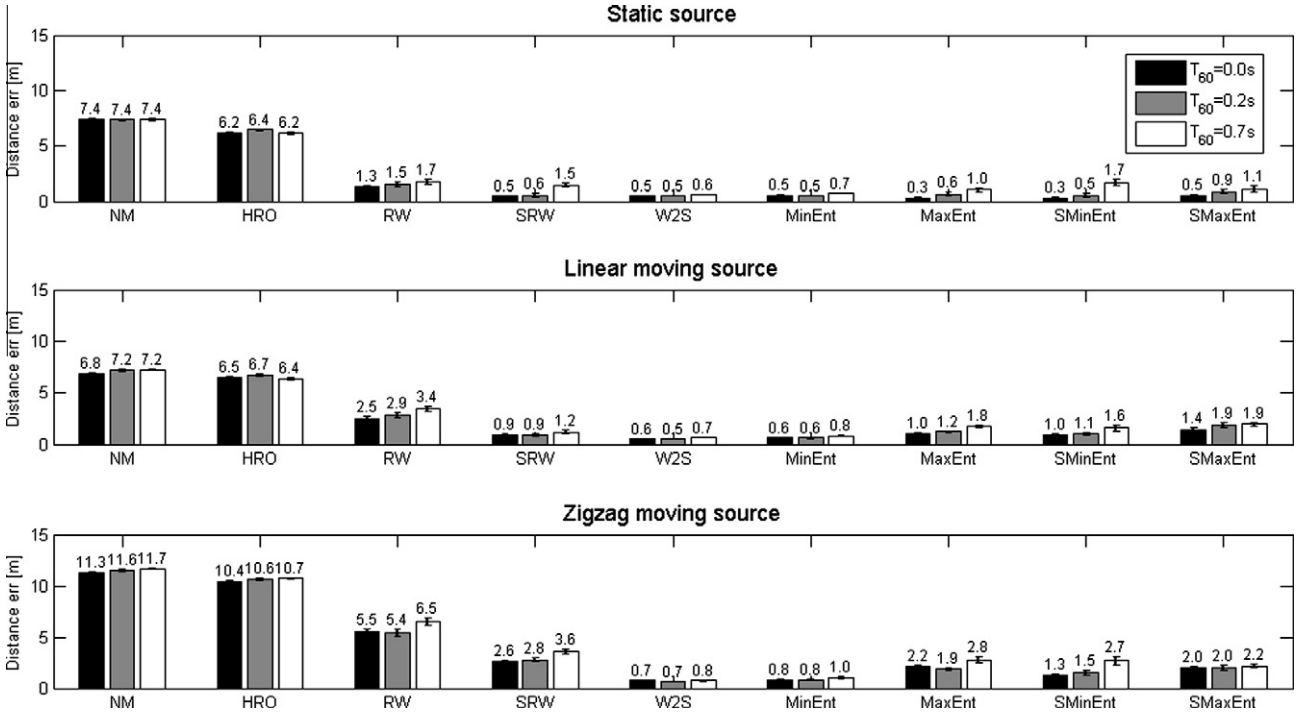

Fig. 12. Distance estimation error.

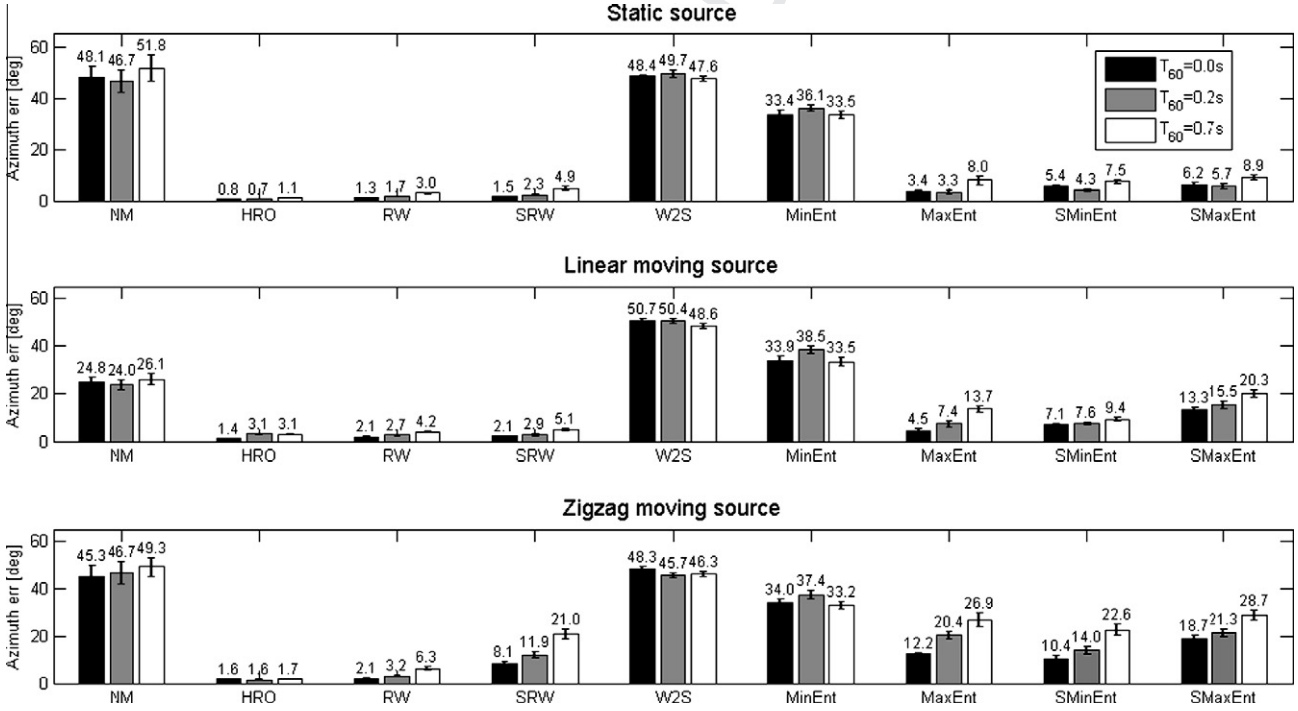

Fig. 13. Azimuth estimation error.

rors area HL with respective to the interaural axis and commonly leads to front–back reversals in estimation due to the similar density of its likelihood function as in the HL region. A few iterations (e.g. 35th to 38th frame in Fig. 14) may be needed to generate more particles in the shaded area via the resampling algorithm to regain good estimation accuracy.

Lacking the head rotation cue, NM suffered from front–back ambiguities, but the azimuth error was halved for the linear motion case, probably due to its approaching the

Please cite this article in press as: Lu, Y.-C., Cooke, M., Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners, Speech Comm. (2010), doi:10.1016/j.specom.2010.06.001

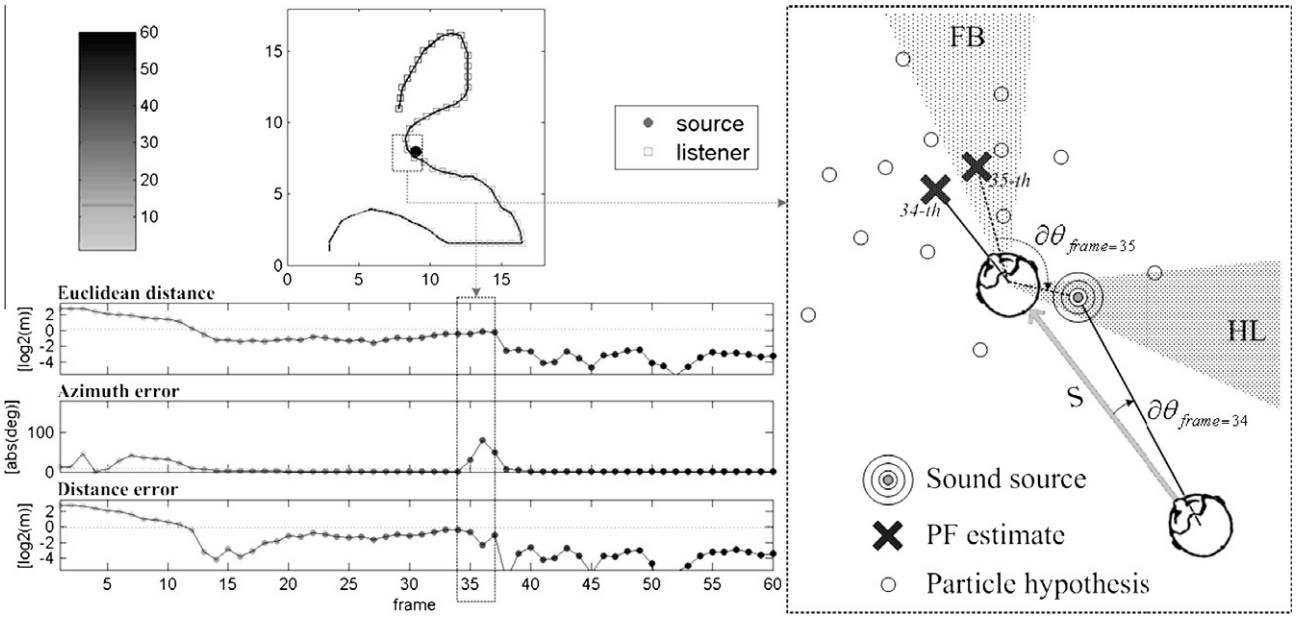

Fig. 14. Effect on localisation accuracy as the simulated listener intercepts a sound source. An example SMinEnt walk is shown in the left with errors in Euclidean distance, unsigned azimuth and distance over time. The listener intercepts the source at the 35th frame and the PF estimation around this frame is illustrated in the right of figure. Since the particles are sufficiently close to the sound source (i.e. evaluated with high importance weights), the noise terms applied as the dynamic model for particle shifting in response to motion S are very small. At the 35th frame, the peak region of likelihood function is mapped to the areas HL and its front-back mirror, FB. The particles in the area FB are incorrectly weighted and result in an increasing $\partial\theta$ from the 34th to 35th frame shown in the schematic and in the running azimuth error panel on the left, while only slight changes are found in Euclidean distance and distance error.

interaural axis of the listener. HRO produced a smaller distance error than NM, suggesting a beneficial effect from head rotation on distance localisation.

To evaluate the effect of Monte Carlo simulation on motion strategy evaluation, additional simulations were conducted with source motion assumed as known a priori at the PF prediction stage. Smaller distance estimation errors for moving sources were produced compared to adopting a source dynamic model, since this may lead to highly variant particle weights. Although there was no change in the ranking of strategies, the performance difference between strategies was reduced. However, the front–back confusion resulting from source-intercepting behaviour still occurred as the listener–source distance was less wide than the particle distribution.

6. Discussion

The simulation results presented here demonstrate a clear advantage of listener motion in localising a speech source in azimuth and distance. Any motion helps relative to the no motion baseline, and arbitrary motion which explores more ground is even more beneficial. However, strategies which minimise source–listener distance produce the lowest errors in reverberant conditions, presumably by reducing the influence of non-direct sound energy.

Similar motion patterns emerged from the smooth random walk (SRW) and the MaxEnt strategy, with comparable performance in low reverberant conditions. The key difference between the two walk styles is that MaxEnt exhibits a search largely centred on the target source while SRW is independent of source location. This finding suggested that as long as exploration is wide-ranging, random motion is no worse than source-directed investigation, again supporting the idea that motion *per se* is of some value in localisation even if it does not reduce source–listener distance.

Two further strategies which exhibited a highly similar style of walk were W2S and MinEnt. Listener motion attempting to place the source in the frontal region ("auditory fovea") yielded a lower motion entropy value than that in the lateral region due to the non-uniform resolution along the ITD delay line. One sample delay in the frontal region corresponds to a span of 2° for the ears separated by 16.2 cm (KEMAR dummy) sampled at 44.1 kHz. On the other hand, a less fine angular discrimination is found for the lateral region where one sample delay is equivalent to around 20° span (Viste and Evangelista, 2004). Due to the non-uniform resolution along the delay line, the set of particles for a frontal source have a less concentrated distribution than an identical geometric configuration for a lateral source. Consequently, the importance is that weights updated with more widely-spread particles possess a lower entropy value since variations among weights are smaller. In our simulations, this effect was especially obvious when particle hypotheses had converged to the actual source location.

By contrast, when particles are more randomly distributed in the target space (normally during the early part of a run), no listener motion, whether placing the source

in front or to the side, is capable of a substantial lowering of entropy. It follows that a MinEnt listener might choose a forwarding direction which is independent of previous motions, since motion entropies at this stage show little directional dependence. This type of "aimless" search behaviour continues until the set of particles starts to converge, reducing uncertainty.

To illustrate distinct types of behaviour during localisation, Fig. 15 shows head orientations (bottom panel) along with the localisation performance (middle panel) during the tracking of a linear moving source by a MinEnt listener. Head orientations were measured with respect to the source direction counter-clockwise from $-\pi$ to π, so a zero head orientation indicates walking towards the source. Three different behaviours are evident. In the first stage (up to the 27th frame), no single strong belief in source location dominates and the MinEnt listener exhibits exploratory searching characterized by unregulated head orientations. For the next few frames (27–30), the head orientations lie roughly in the direction of the source, leading to below-average estimation errors. In the third stage (frame 31 onwards), the listener has started to intercept the source. Large head rotations are generated when passing by the source (as depicted in Fig. 14) and attempting to re-steer towards the source.

Although the current study focused on virtual environments and simulated listeners, some of the emergent behaviours shared characteristics with those observed for human listeners homing in on virtual sound sources. Loomis et al. (1990) used a head tracker to measure head location and orientation in order to evaluate the effectiveness of a virtual auditory display system in conveying sound source location. During the localisation experiments, subjects were instructed to walk to the perceived source location. They varied in the extent to which they used head rotations to localize sources during motion. As the subject–source distance reduced, they ceased rotating the head and attempted to sense solely through head translation (i.e. body movement). Fig. 15 depicts somewhat similar behaviour (cf. Loomis et al., 1990, Fig. 3) when a MinEnt listener attempts to locate and track a linearly-moving source in the highest reverberation condition.

This comparison highlights one weakness of the simulations employed here. While real listeners are capable of independent (if limited) head rotations and body translation, our artificial listeners were constrained to move in the direction of the head rotation, restricting the range of possible emergent behaviours. Further, the simulations used fixed length translational motion, while real listeners are capable of a range of "step" sizes as well as more complex head motions such as tilting (Thurlow et al., 1967).

While the lookahead strategy employed in the motion entropy-based strategies are non-physical, they are presumably approximately realisable as a searching behaviour over short time periods, under the assumption that auditory "memory" is available to be able to compare information acquired at various stages of the search process.

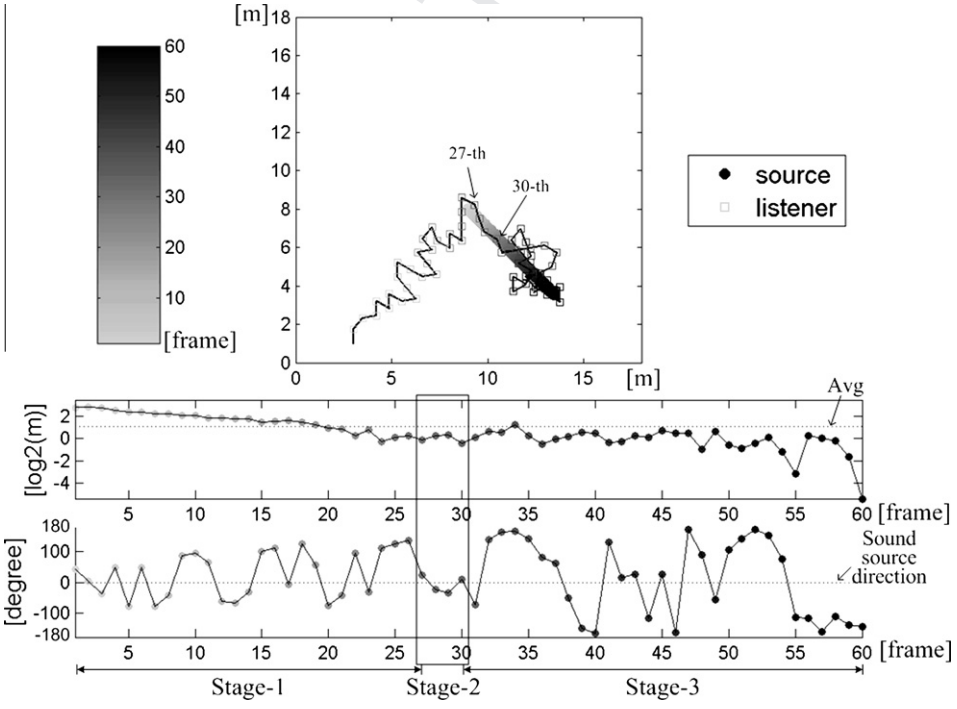

Fig. 15. MinEnt listener head orientations with respect to a linearly-moving sound source during localisation in simulated reverberation ($T_{60} = 0.7$ s). Listener and source trajectories are shown in the top panel. The Euclidean distance between the estimated and true source location, and the simulated listener's head orientation are displayed for each time frame in the middle and bottom panels respectively. Head orientation is encoded counter-clockwise and measured with respect to the source direction in order to verify the existence of steering towards the source behaviour. According to the observed head rotation, the epoch is divided into three stages: exploration, approaching-source and intercepting-source.

Please cite this article in press as: Lu, Y.-C., Cooke, M., Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners, Speech Comm. (2010), doi:10.1016/j.specom.2010.06.001

There is a need to evaluate listener motion strategies in real rather than simulated spaces where factors such as reverberation are more variable. Listener motion was assumed to be known, which may be a valid assumption in some artificial listener scenarios (such as mobile robotic platforms) but will be only approximately true for real listeners without additional sensors.

7. Conclusions

This study investigated whether motion is useful for localising sound sources by simulated listeners in a dynamic acoustic environment. Eight motion strategies were evaluated in anechoic and reverberant conditions, for static and moving sources, relative to a no motion baseline. The principal findings are summarised here.

1. Relative to no motion, all motion strategies resulted in better localisation.
2. Limiting motion to head rotation only (no translation) resulted in a small improvement over the no motion baseline.
3. Translational motion was always beneficial relative to the static case, with maximum estimation errors (i.e. those produced by random motion) at a fourth of those for the static listener-static source configuration, and around half that of the static listener for moving sources.
4. A smooth random walk strategy, produced by limiting the degree of head rotation, was as an effective strategy as purposeful movement for static and linearly-moving sources. However, purposeful motion produced significantly lower errors for more complex source motion.
5. Of the purposeful motion strategies, those which resulted in a reduction of distance to the estimated source location led to the best overall performance, probably due to an improvement in the ratio of direct to reverberant energy. Movement towards the estimated source was obtained either explicitly or as a result of a strategy which minimised motion entropy.
6. Different types of simulated listener searching behaviour were evident at different stages during trajectory evolution. Some similarities with real listener motion were apparent, although more behavioural studies are needed to fully explore the question of model–listener comparisons.
7. Overall estimation errors were dominated by the distance component. Large azimuth errors were evident for the best strategies due to source-searching behaviour in the initial stages and source tracking following interception.

Appendix A. Definition and evaluation of motion entropy

A sequence of motion entropies e associated with K choices of the next step can be calculated by accessing the state parameters of the particle filter using

$$e_{k,t} = M_{ent}(a_{k,t}, \hat{\phi}_{t-1}, \vec{\omega}_{t-1}, \vec{\xi}_{t-1}), \quad k = 1, \ldots, K, \tag{37}$$

where M_{ent} is the motion entropy function, which requires the current potential listener motions $a_{k,t}$, Eqs. (33)–(35), the previous azimuth estimate $\hat{\phi}_{t-1}$, particle weights $\vec{\omega}_{t-1}$ and location hypotheses $\vec{\xi}_{t-1}$. Entropy values are calculated based on the particle weights evaluated by a predicted likelihood function for time t, which is estimated according to preset simulation parameter for the level of reverberation. These motion entropies are used to determine strategic motion among K choices as introduced in Section 4.6. The process requires motion candidate enumeration, particle weight update through prediction and motion entropy calculation. These three stages are detailed below.

A.1. Motion candidate enumeration

K potential listener motions $a_{k,t}$ are generated based on Eq. (33) in order to relocate particle hypotheses (i.e. source location in distance and azimuth):

$$\vec{\xi}_{k,t} = f_{ent}[a_{k,t}, \vec{\xi}_{t-1}], \tag{38}$$

where $\vec{\xi}_{t-1}$ is a N-element vector and N is the number of particles. Function f_{ent} shifts these N source location hypotheses in polar coordinates according to K different listener motions into a two dimensional vector $\vec{\xi}_{k,t}$ as illustrated in Fig. 16. f_{ent} operates like the PF state transition function, Eq. (17), without the noise term. Each shifted hypothesis set generates an associated motion entropy quantity.

A.2. Particle weight update through prediction

The derived source location priors $\vec{\xi}_t$ corresponding to the K motion candidates are evaluated with an identical set of Gaussian-based likelihood functions η_m. Gaussians are centred on M distinct positions, denoting M possible different observations for source location, along the delay line used for ITD calculation:

$$\eta_m \approx \mathcal{N}\left(m - \frac{M+1}{2}, \sigma_r^2\right), \quad m = 1, \ldots, M, \tag{39}$$

where $\mathcal{N}$ denotes the Gaussian function and M indicates the length of the delay line, which divides the frontal hemisphere into M non-uniformly separated regions. All M Gaussian-based likelihood functions η_m are specified with a constant standard deviation σ_r, as a function of the level of room reverberation. A higher level of reverberation, which increases the ambiguity in discerning the source azimuth, leads to a large σ_r. The function specifying the relationship between the reverberation level and σ_r was derived empirically using a small data set (see Table 3 for the used value). This step can be considered an "environment-related" adaptation.

Particle weights are renewed by the predicted likelihood functions as specified by Eq. (40), which is a modification of the SIR PF update stage, Eq. (12)

Please cite this article in press as: Lu, Y.-C., Cooke, M., Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners, Speech Comm. (2010), doi:10.1016/j.specom.2010.06.001

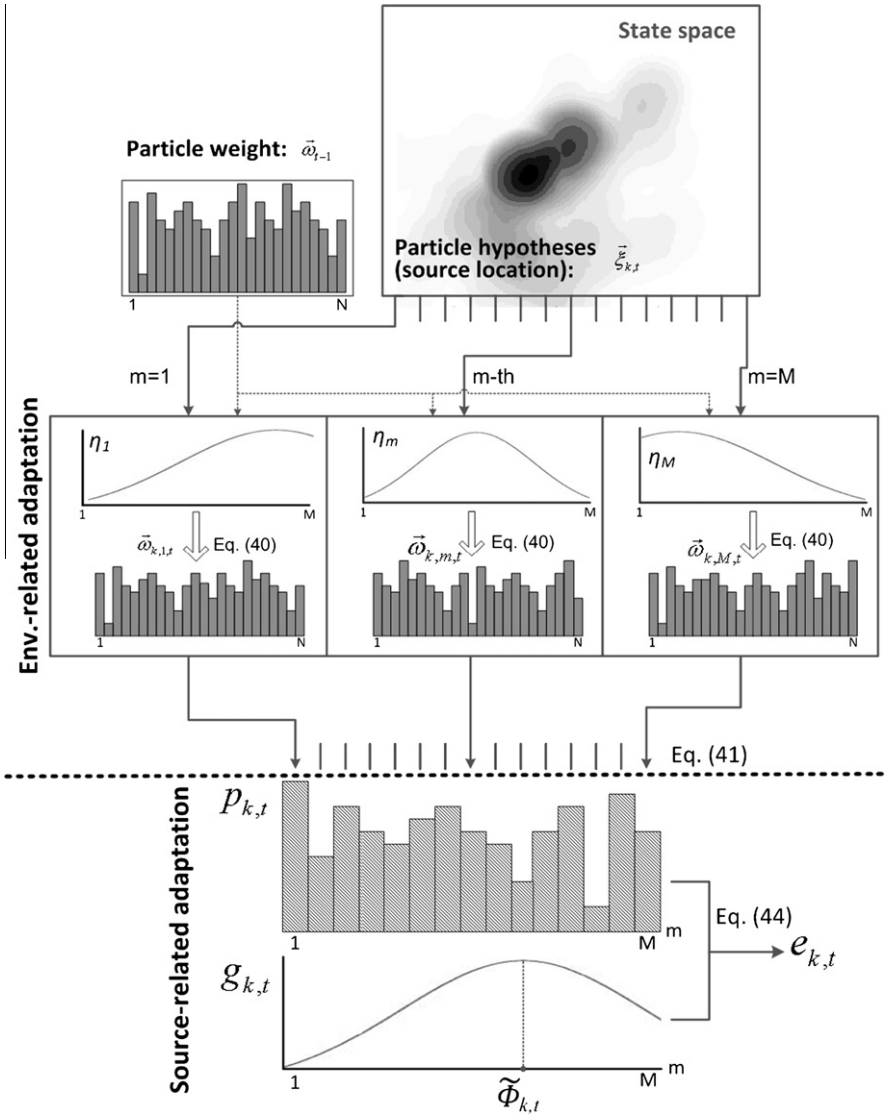

Fig. 16. Motion entropy $e_{k,t}$ calculation for a particular motion candidate. The processing depicted in the upper half operates as an "environment-related" adaptation to generate M entropy values $p_{k,t}$ based on predictive importance weights $\vec{\omega}_{k,m,t}$. The lower half can be considered a "source-related" adaptation which further processes the entropy vector $p_{k,t}(m)$ into motion entropy $e_{k,t}$ via a weighting function $g_{k,t}$ which employs the source dynamic factor into the strategy development.

$$\vec{\omega}_{k,m,t} \propto \vec{\omega}_{t-1} \cdot \eta_m(\vec{\xi}_{k,t}), \tag{40}$$

M sets of importance weights $\vec{\omega}_{k,m,t}$ are updated based on η_m and the particle hypotheses $\vec{\xi}_{k,t}$ associated with particular listener motions k. They are subsequently calculated through the standard definition of entropy, leading to the M-element entropy $p_{k,t}(m)$

$$p_{k,t}(m) = -\sum_{n=1}^{N} \omega_{k,m,t}(n) \log(\omega_{k,m,t}(n)), \tag{41}$$

where N is the number of particles and the summation of N importance weights is normalised to unity prior to the calculation of entropy. The upper half of Fig. 16 depicts how the M-element $p_{k,t}(m)$ is derived through M different predicted likelihood functions η_m corresponding to the kth motion candidate. The motion entropy is derived from these M $p_{k,t}(m)$ according to the likelihood of observing their associated η_m. For example, equally-occurring η_m lead to a uniform probability of producing each element of $p_{k,t}(m)$ as motion entropy.

A.3. Motion entropy calculation

Motion entropy $e_{k,t}$ is calculated based on the $p_{k,t}(m)$ after the relative importance of its elements has been estimated. At the position along the delay line where the peak of the predicted likelihood function coincides with the sound source, the corresponding $p_{k,t}(m)$ score is considered to be high. For a static source, listener motion is the only source accounting for azimuth alteration from the previous frame. Consequently, the most probable azimuth observation $\tilde{\phi}_{k,t}$ can be obtained by summing the previous azimuth estimate $\hat{\phi}_{t-1}$ and the kth "look ahead" head rotation $a_{h,k,t}$

$$\tilde{\phi}_{k,t} = \hat{\phi}_{t-1} + a_{h,k,t}. \quad (42)$$

However, estimation bias and sound source motion may produce a poor prediction of $\tilde{\phi}_{k,t}$. Thus, a Gaussian centred at $\tilde{\phi}_{k,t}$ in Eq. (43) is used to regularise the likelihood of $p_{k,t}(m)$

$$g_k(t,m) \approx \mathcal{N}(m; \tilde{\phi}_{k,t}, \sigma_s^2), \quad (43)$$

where m is a dummy variable on the abscissa (delay line) axis and σ_s indicates the degree of influence from estimation bias and possible source motion. A larger bias or motion gives rise to a higher σ_s (see Table 3 for the used value).

Finally, motion entropy $e_{k,t}$ is obtained by picking the centroid of M weighted $p_{k,t}(m)$ values

$$e_{k,t} = \sum_{m=1}^{M} p_{k,t}(m) g_{k,t}(m). \quad (44)$$

The lower half of Fig. 16 depicts this weighting process, which can be considered as a "source-related" adaptation. The output of the entropy calculation is K motion entropy values which lead to the development of strategic entropy walks as specified in Table 1.

References

Aarabi, P., 2002. Self-localizing dynamic microphone arrays. IEEE Trans. Systems, Man Cybernet. 32 (4), 474–484.

Allen, J.B., Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Amer. 65 (4), 943–950.

Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T., 2002. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. Signal Process. 50 (2), 174–188.

Ashmead, D.H., Davis, D.L., Northington, A., 1995. Contribution of listeners' approaching motion to auditory distance perception. J. Exp. Psychol. – Human Percept. Perform. 21 (2), 239–256.

Asoh, H., Asano, F., Yoshimura, T., Yamamoto, K., Motomura, Y., Ichimura, N., Hara, I., Ogata, J., 2004. An application of a particle filter to bayesian multiple sound source tracking with audio and video information fusion. In: Proc. Fusion, pp. 805–812.

Blauert, J., 1997. Spatial Hearing – The Psychophsics of Human Sound Localization. The MIT Press, London, England.

Bodden, M., 1993. Modeling human sound-source localization and the cocktail-party-effect. Acta Acust. 1 (1), 43–55.

Brandstein, M.S., 1997. A pitch-based approach to time-delay estimation of reverberant speech. In: Proc. WASPAA.

Brandstein, M.S., Silverman, H.F., 1997. A robust method for speech signal time-delay estimation in reverberant rooms. In: Proc. ICASSP, pp. 375–378.

Campbell, D.R., Palomaki, K.J., Brown, G., 2005. A Matlab simulation of "shoebox" room acoustics for use in research and teaching. Comput. Inform. Systems J. 9 (3), 48–51.

Champagne, B., Bedard, S., Stephenne, A., 1996. Performance of time-delay estimation in the presence of room reverberation. IEEE Trans. Audio Speech Process. 4 (2), 148–152.

Chen, J., Benesty, J., Huang, Y.A., 2006. Time delay estimation in room acoustic environments: An overview. EURASIP J. Appl. Signal Process. 2006, 1–19.

Cooke, M., Lu, Y.-C., Lu, Y., Horaud, R., 2008. Active hearing, active speaking. In: Dau, T., Buchholz, J.M., Harte, J.M., Christiansen, T.U. (Eds.), Auditory Signal Processing in Hearing-Impaired Listeners. Centertryk.

de Freitas, N., Niranjan, M., Gee, A., Doucet, A., 1998. Sequential Monte Carlo Methods for Optimisation of Neural Network Models. Technical Report CUED/F-INFENG/TR 328, Cambridge University Department of Engineering.

Del Moral, P., 1997. Non-linear filtering: Interacting particle resolution. Markov Process. Related Fields 2 (4), 555–580.

Douc, R., Cappe, O., 2005. Comparison of resampling schemes for particle filtering. In: Proc. ISPA, pp. 64–69.

Doucet, A., Godsill, S., Andrieu, C., 2000. On sequential Monte Carlo sampling methods for Bayesian filtering. Statist. Comput. 10, 197–208.

Eyring, C.F., 1930. Reverberation time in "dead rooms". J. Acoust. Soc. Amer. 1 (2A), 168.

Faller, C., Merimaa, J., 2004. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. J. Acoust. Soc. Amer. 116 (5), 3075–3089.

Gaik, W., 1993. Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling. J. Acoust. Soc. Amer. 94 (1), 98–110.

Gardner, W.G., Martin, K.D., 1996. HRTF measurements of a KEMAR dummy-head microphone. In: Haus, G., Pighi, I. (Eds.), Standards in Computer Generated Music. IEEE CS Tech. Com. on Computer Generated Music.

Gordon, N.J., Salmond, D.J., Smith, A.F.M., 1993. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEE Proc. Radar Signal Process. 140 (2), 107–113.

Jan, E.-E., Flanagan, J., 1996. Sound source localization in reverberant environments using an outlier elimination algorithm. In: Proc. ICSLP, pp. 1321–1324.

Jeffress, L.A., 1948. A place theory of sound localization. Comp. Physiol. Psychol. 41, 35–39.

Kidd Jr., G., Arbogast, T.L., Mason, C.R., Gallun, F.J., 2005. The advantage of knowing where to listen. J. Acoust. Soc. Amer. 118 (6), 3804–3815.

Kitagawa, G., 1996. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. J. Comput. Graphical Statist. 5 (4), 1–25.

Knapp, C.H., Carter, G.C., 1976. The generalized correlation method for estimation of time delay. IEEE Trans. Speech Audio Process. 24 (4), 320–327.

Lindemann, W., 1986. Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals. J. Acoust. Soc. Amer. 80 (6), 1608–1622.

Liu, J.S., Chen, R., 1998. Sequential Monte Carlo methods for dynamic systems. J. Amer. Statist. Assoc. 93 (443), 1032–1044.

Loomis, J.M., Hebert, C., Cicinelli, J.G., 1990. Active localization of virtual sounds. J. Acoust. Soc. Amer. 88 (4), 1757–1764.

Lu, Y.-C., Cooke, M., Christensen, H., 2007. Active binaural distance estimation for dynamic sources. In: Proc. Interspeech, pp. 574–577.

Lukowicz, P., Ward, J.A., Junker, H., Stager, M., Troster, G., Atrash, A., Starner, T., 2004. Recognizing workshop activity using body worn microphones and accelerometers. Lecture Notes Comput. Sci. 3001, 18–32.

Mackensen, P., 2004. Auditive Localization. Head Movements, an Additional Cue in Localization. Ph.D. Thesis, Dept. of Physics and Communication Science, Technical University Berlin, Berlin.

Martinson, E., Schultz, A., 2006. Auditory evidence grids. In: Proc. IROS, pp. 1139–1144.

Otani, M., Hirahara, T., 2007. A dynamic virtual auditory display: Its design, performance, and problems in HRTF switching. In: Proc. the Japan–China Joint Conf. of Acoustics 2007.

Patterson, R.D., Nimmo-Smith, I., Holdsworth, J., Rice, P., 1988. APU report 2341: An efficient Auditory Filterbank based on The Gammatone Function. Applied Psychology Unit Cambridge, UK.

Pitt, M.K., Shephard, N., 1999. Filtering via simulation: Auxiliary particle filters. J. Amer. Statist. Assoc. 94 (446), 590–599.

Rekleitis, I.M., 2003. Cooperative Localization and Multi-robot Exploration. Ph.D. thesis, School of Computer Science, McGill University, Montreal, Quebec, Canada.

Rui, Y., Florencio, D., 2004. Time delay estimation in the presence of correlated noise and reverberation. In: Proc. ICASSP, pp. 133–136.

Please cite this article in press as: Lu, Y.-C., Cooke, M., Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners, Speech Comm. (2010), doi:10.1016/j.specom.2010.06.001

Sasaki, Y., Kagami, S., Mizoguchi, H., 2006. Multiple sound source mapping for a mobile robot by self-motion triangulation. In: Proc. IROS, pp. 380–385.

Sawhney, N., Schmandt, C., 2000. Nomadic radio: Speech and audio interaction for contextual messaging in nomadic environments. ACM Trans. Comput. Human Interact. (COCHI) 7 (3), 353–383.

Speigle, J.M., Loomis, J.M., 1993. Auditory distance perception by translating observers. In: Proc. IEEE Symposium on Research Frontiers in Virtual Reality, pp. 92–99.

Thurlow, W.R., Mangels, J.W., Runge, P.S., 1967. Head movements during sound localization. J. Acoust. Soc. Amer. 42 (2), 489–493.

Viste, H., Evangelista, G., 2004. Binaural source localization. In: Proc. 7th Internat. Conf. on Digital Audio Effects, pp. 145–150.

Wallach, H., 1940. The role of head movements and vestibular and visual cues on sound localization. J . Exp. Psychol. 27, 339–368.

Wang, H., Chu, P., 1997. Voice source localization for automatic camera pointing system in videoconferencing. In: Proc. ICASSP, pp. 187–190.

Ward, D.B., Lehmann, E.A., Williamson, R.C., 2003. Particle filtering algorithms for tracking an acoustic source in a reverberant environment. IEEE Trans. Speech Audio Process. 11 (6), 826–836.

West, M., Harrison, J., 1997. Bayesian Forecasting and Dynamic Models. Springer-Verlag, New York.

Zahorik, P., Brungart, D.S., Bronkhorst, A.W., 2005. Auditory distance perception in humans: A summary of past and present research. Acta Acust. United Acust. 91 (3), 409–420.

Please cite this article in press as: Lu, Y.-C., Cooke, M., Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners, Speech Comm. (2010), doi:10.1016/j.specom.2010.06.001