# Binaural Estimation of Sound Source Distance via the Direct-to-Reverberant Energy Ratio for Static and Moving Sources

Yan-Chen Lu and Martin Cooke

***Abstract*—One of the principal cues believed to be used by listeners to estimate the distance to a sound source is the ratio of energies along the direct and indirect paths to the receiver. In essence, this "direct-to-reverberant" energy ratio reveals the absolute distance component of the direct energy by normalising by what is assumed to be distance-independent reverberant energy. Earlier approaches to direct-to-reverberant energy ratio calculation made use of the estimated room impulse response, but these techniques are computationally expensive and inaccurate in practice. This paper proposes and evaluates an alternative approach which uses binaural signals to segregate energy arriving from the estimated direction of the direct source from that arriving from other directions, employing a novel binaural equalization-cancellation technique. The system is integrated with a probabilistic inference framework, particle filtering, to handle the nonstationarity of energy-based measurements. The algorithm is capable of using reverberation to estimate source distance in large rooms with errors of less than 1 m for static sources and 1.5-3.5 m for sources with varying degrees of motion complexity. Model performance can be accounted for largely in terms of a competition between auditory horizon and source energy fluctuation effects.**

***Index Terms*—Acoustic distance measurement, direct-to-reverberant energy ratio, particle filtering, binaural sound source localization.**

## I. Introduction

JUDGEMENT of ego-centric distance to nearby objects is an important human sensory capability and is at times wholly - and critically - dependent on auditory input. A cyclist uses engine noise from motor vehicles to estimate distance in order to pick up speed in time and prepare for evasive action. A runner listens to the footsteps and breathing of his or her opponents to evaluate whether their lead in the race is adequate enough. The approach of a mosquito person at night can be detected through the sound made by its vibrating wings.

A preliminary report of this work appeared at the International Workshop on Acoustic Echo and Noise Control (Seattle, July 2008). This work was funded in part by EU Cognitive Systems STREP Project POP (Perception On Purpose), FP6-IST-2004-027268.

Yan-Chen Lu is with the Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, U.K. (e-mail: yanchen@dcs.shef.ac.uk)

Martin Cooke is with the Language and Speech Laboratory, University of the Basque Country, Spain and IKERBASQUE (Basque Foundation for Science) (e-mail: m.cooke@ikerbasque.org).

Absolute sound energy at the receiver is a function of intrinsic source energy and source distance, both of which may be time-varying, precluding the use of energy alone as a cue to source-listener distance. However, the combination of energy received along the direct source-listener path with energy arriving following reflections has potential as a means of estimating source distance. The "direct-to-reverberant" energy ratio (DRR) has been suggested as part of the mechanism for source distance judgements in listeners [3-6]. Distance judgments are more accurate in a reverberant space than in an anechoic space, with small inter-test variation in judgements in the same environment [3]. Listeners may use reverberation as an absolute distance cue given that accurate distance judgments were obtained at first stimulus presentation [4]. Zahorik [7] suggested that the principal role of the DRR cue was to provide absolute distance information rather than support fine distance discriminations and was poor as a relative cue. Zahorik also suggested that DRR was perceptually more salient than an intensity cue, especially in a situation where prior knowledge of natural speech level could not be used due to other more variable and complex acoustic information in the surrounding environment [8].

Attempts have been made to exploit the DRR cue both monaurally and binaurally. Bronkhorst and Houtgast [9] proposed a computational model to predict human distance judgement in a controlled condition where the DRR cue is dominant. Their model demonstrated accurate prediction of subjects' distance responses based on prior knowledge of certain acoustical properties of the environment (room volume, reverberation time and source directivity) using monaural data. Relying on blind identification of the room impulse response from the monaural signal, Larsen and colleagues [10] developed a technique to compute DRR based on certain assumed room acoustics parameters, such as the duration of direct sound. They found that source distance could also be determined as an intermediate output and arrived at underestimated judgments for sources at moderate and large distances (further away than 2 m), similar to the pattern found with human listeners. Other models based on binaural signals utilised either prior knowledge of the environment (e.g. room impulse responses [11]) or extensive spectral training data [12] to formulate source distance inference using the DRR concept. While these studies attempted to demonstrate that distance inference can be further improved with binaural input, neither study, surprisingly, emphasised the role of directional information.

One basic step in computing DRR involves segregating the direct and reverberant signals from the acoustic mixture. A common approach uses the difference in arrival time of the two components [9], usually applied by specifying an integration window for the room impulse response (e.g. treating the leading 4 ms portion of the signal as direct) to determine the direct sound energy. However, it is difficult to extract a precise long room impulse response by de-convolving the raw signal in a reasonable run time [13].

It is not clear whether the auditory system also extracts DRR via a similar temporal scheme. The hypothesis that the separation of direct and reverberant components correlated highly with the detection of sharp onsets (or offsets) was not supported following a study which showed that distance perception is unaffected by the shape of the envelope of the sound [14]. Other experiments by Bronkhorst [11] demonstrated that a reduction of interaural correlation led to a strong decrease of apparent distance. This suggests that human listeners might also use binaural information to determine sound source distance.

Here, we explore the possibility of performing direct/reverberant energy segregation based on estimated source direction. By removing the energy of a target signal which occupies a particular azimuthal region, the reverberant signal can be identified by its diffuse (i.e non-directional) characteristic. An adaptive sub-band scheme proposed by Liu et al. [15] to address a different problem, that of separating multiple sources, motivated the approach developed in this study. Their two-microphone system exploited sound location information to steer independent nulls that suppressed the strongest interference in each time-frequency region, using a dual delay-line structure. We adapt this technique to extract signal energy for each angular position as a means of separating the direct signal from the reverberant signal. The result is used to generate the DRR from which a likelihood function can be derived.

The approach taken in the current study is depicted in Fig. 1. Left and right ear signals are first processed to allow estimation of sound source location in both azimuth and distance based on cross-correlation and DRR features. These cues are then combined to create either an instantaneous (non-sequential) estimate of listener-source distance, or integrated through time using a sequential method. In the latter case, location priors are updated with new observations in a sequential particle filtering (PF) framework in an iterative manner for inferring sound location.

Section II addresses the extraction of cross-correlation and DRR cues to location, while section III evaluates the relationship between DRR cues and distance for both real and synthetic environments. The DRR-distance relationship is modelled using Gaussian mixtures as outlined in section IV. Section V describes the use of particle filtering to integrate distance estimates through time. An evaluation of the DRR cue for judging the distance to both static and moving sound sources in reverberant conditions is presented in section VI.

## II. Equalization-Cancellation Method for Estimation of the Direct-to-Reverberant Energy Ratio

A system for DRR estimation is introduced capitalizing on target source directional information. It is fundamentally an equalization-cancellation (EC) operation applied on a reverberant binaural signal. The EC concept was proposed to explain the masking suppression process in the presence of a single noise source [16]. Equalization renders the magnitudes of noise components to be identical between channels, while cancellation subtracts the noise component in one channel from that of the other channel. In our application of the EC principle, the direct signal, which is identified by its angular position, is the "noise" component.

The EC-based DRR estimation system is outlined in Fig. 2. First, successive windowed frames of a binaural signal are processed by a pair of N-channel gammatone filterbanks [17]. Next, individual filter outputs from left and right ear models feed two binaural interaction processes, namely cross-correlation (CC) and equalization-cancellation, operating on an M-element delay line. A cross-frequency integration stage enables robust localisation of the direct sound source and estimation of the source power distribution as a function of interaural delay. Finally, a single DRR value is generated for each frame of data input. The direct energy is estimated via azimuthal information from the source localizer which is used to select the direct source power at the corresponding delay-line index, denoted $j_{source}$. The DRR is estimated as the ratio of direct energy to reverberant energy, the latter computed as the

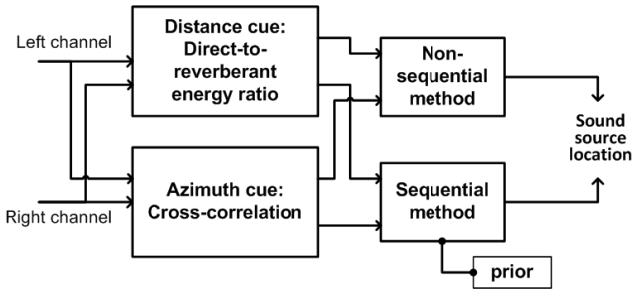


Fig. 1. Computational mode for evaluating reverberation cues in sound source localisation.

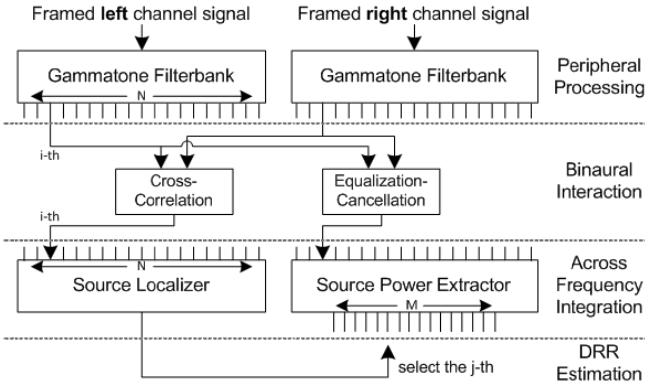


Fig. 2. Schematic diagram of EC-DRR system.

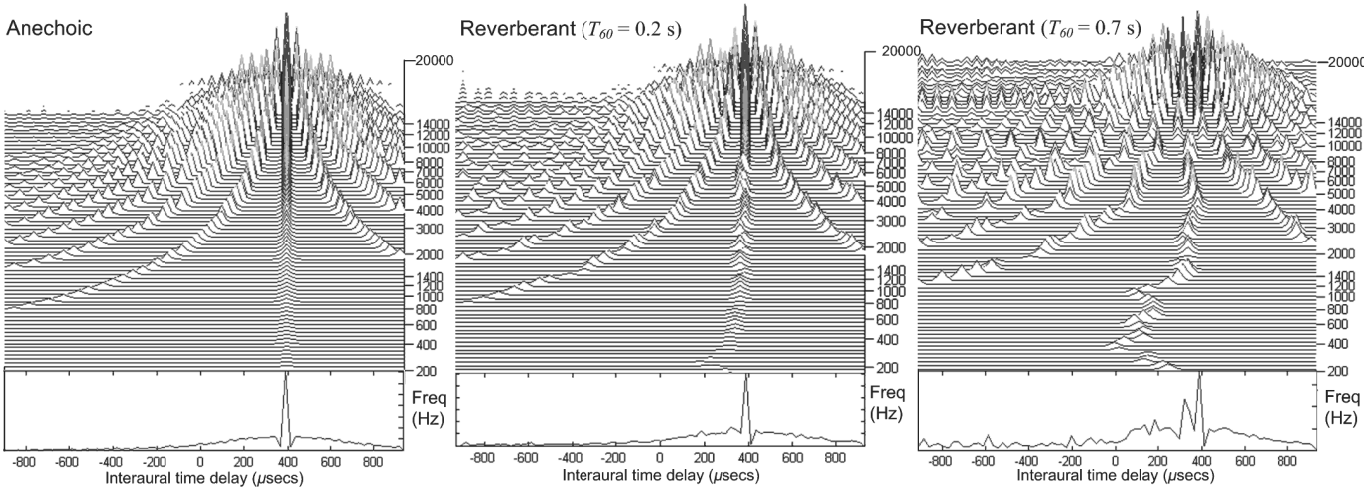


Fig. 3. ITD extraction for a pink noise source located 45° to the right. The lower plots in each panel are cross-correlogram summaries across frequency (from 200 Hz to 20 kHz). The location corresponding to the source is identified by detecting the peak in the lower plot at which the summarised cross-correlogram is maximized. Spurious peaks occur due to reverberation (right). By contrast, a single salient peak is easily observed in the anechoic case (left) and in the less reverberant condition (middle).

residual of total signal energy $S$ after subtraction of the direct energy component as

$$DRR = Dj_{source}/(S - Dj_{source}) \tag{1}$$

The individual steps are now described in more detail.

### A. Cross-correlation

ITD estimates were computed from the summary cross-correlation (2) of the left and right "ear" outputs of auditory filters, $X_L(t,f)$ and $X_R(t,f)$, modelled using a bank of $N$=32 gammatone filters [17] with centre frequencies equally spaced on an ERB-rate scale between 50 and 8000 Hz with $k$ and $T$ the start of the current frame and the number of samples per frame respectively:

$$CC(m) = \sum_{f=1}^{N} \sum_{t=k}^{T+k-1} X_L(t,f) X_R(t+m,f) \tag{2}$$

ITD is estimated by identifying the maximum value of summary cross-correlation:

$$\tau = \underset{m}{\operatorname{argmax}}\, CC(m) \tag{3}$$

Examples of cross-correlations and their average in both anechoic and reverberant ($T_{60}$= 0.2 s and 0.7 s) spaces are shown in Fig. 3. As can be discerned in the right panel of the figure ($T_{60}$= 0.7 s), fluctuations resulting from reverberation contribute to spurious peaks in addition to the peak at the desired ITD (395 $\mu$s), while the middle panel with a moderate reverberation ($T_{60}$= 0.2 s) shows a smaller degree of peak shifting in the individual frequency bands which has little effect on the average.

It is more convenient to work with azimuth angle rather than directly with ITDs. Given the relationship observed between ITD and azimuth from head-related transfer function data, it is possible to generate an azimuth-ITD transformation based on table lookup [18]. In previous work [19] we demonstrated using a room simulator for a range of reverberation settings that this transformation was not significantly affected by source distance.

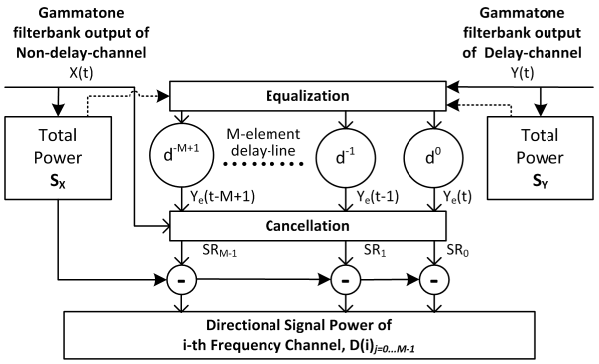


Fig. 4. EC module at $i$-th frequency channel in EC-DRR system.

### B. Equalization-cancellation

A block diagram of the delay-line EC module is shown in Fig. 4. The in-phase position of a target source in one channel with respect to the other channel is determined by the azimuthal information derived above. The in-phase signal components in both channels are assumed to be identical after equalization and can be cancelled by subtracting one from the other. One of the two channels is selected as the delay-channel which is compensated by equalization and delayed prior to cancellation. Power in the non-delayed channel is computed as

$$S_X(f) = \sum_{t=k}^{T+k-1} |X(t,f)|^2, f = 1 \ldots N \tag{4}$$

The compensation factor $E(f)$, (5), used by the equalization block is updated every frame to generate $Y_e$, the compensated delay-channel signal as (6).

$$E(f) = (S_X(f) / S_Y(f))^{0.5} \quad (5)$$

$$Y_e(t, f) = Y(t, f)E(f) \quad (6)$$

The delayed channel is equalized with respect to the other channel to compensate for differences in energy captured through the two microphones. The cancellation block subtracts the compensated delayed signal from the non-delayed channel and accumulates the residual energy for each delay as

$$SR_j(f) = \frac{T}{T-j} \sum_{t=k}^{T+k-1} \left| X(t,f) - Y_e(t-j,f) \right|^2, j = 0 \ldots M-1 \quad (7)$$

The estimated direct energy $D_j(f)$ represented by the cancelled component is integrated with those from other frequency channels in the source power extractor as

$$D_j = \sum_{j=1}^{N} D_j(f) = \sum_{j=1}^{N} [S_X(f) - SR_j(f)] \quad (8)$$

## III. Relationship between EC-DRR and Distance

Ideally, DRR is a quantity that varies only with source distance and is independent of source power. The effectiveness of the proposed EC-DRR system was judged according to how well DRR reflected actual source distance. A pink noise source with constant power was used to generate simulated audio sequences in an 18 m by 18 m by 10 m rectangular space using a room simulator (see Section VI.A for more details). The distance between the simulated listener (KEMAR head model [20]) and the noise source was increased from 2 m to 11 m with $T_{60}$= 0.2 s, where $T_{60}$ indicates the time required for sound level to drop by 60 dB following sound onset. The resulting binaural signals were processed by the EC-DRR system, resulting in DRR estimates.

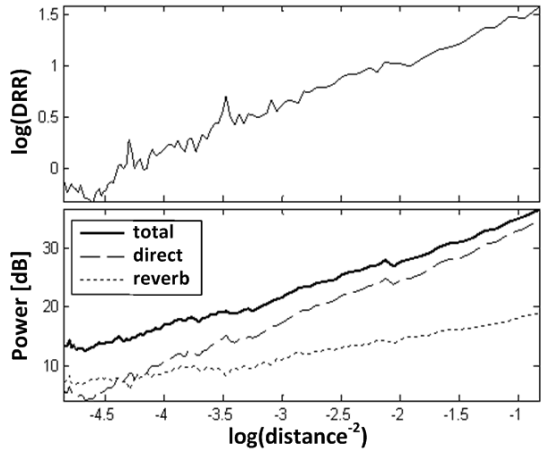


Fig. 5. DRR calculated by EC-DRR system (upper) and its segregated components (lower).

Estimated DRR increases with decreasing distance in this simulated case as shown in the upper part of Fig. 5, suggesting that the EC-DRR approach does generate a distance related feature. The relationship between log(DRR) and the inverse squared distance is approximately linear. The lower part of Fig. 5 displays total energy, estimated direct energy and corresponding reverberant energy as a function of source distance. The total and direct energy increased as the distance decreased. However, reverberant energy was not constant, as would be expected in a truly diffuse sound field. Instead, the estimated reverberant component also increased as the distance decreased, albeit at a slower rate than the estimated direct energy. This outcome may be due to non-ideal direct signal extraction in the delay-line structure as well as the limited number of reflecting surfaces employed in the room simulator.

### A. Evaluation using real stimuli

Real test sequences collected from a 9 m by 6 m by 4 m classroom were processed to compare with the simulated case. Five sides of the room were hard concrete walls while the upper part of remaining wall was glazed. No acoustically hard objects were present in the room during the measurements. Two different static sources, pink noise and speech, were used. Sources were placed in front of a pair of Bruel & Kjaer (B & K) type 4190 1/2-in microphones, placed 10.6 cm apart at two sides of a manikin head. The signal was preamplified by a B & K Nexus model 2690 conditioning amplifier prior to digitization at 44.1 kHz by a M-Audio MobilePre A-D processor. Eighteen different distances were used from 0.75 m to 5 m at 0.25 m intervals. Recorded sequences for each distance were 10 s long and processed in 200 ms frames to obtain 50 DRR values per distance, and 900 in total.

The DRR estimates are depicted in Fig. 6 with respect to their distance to the noise or speech source along with their mean and standard deviation. Both speech and noise show a clear relationship between source distance and *log*(1/DRR) for distances up to around 2.5 m. Thereafter, DRRs show less dependence on distance. Noise sources have a narrower DRR distribution than those of speech sources, and the width of the distribution narrows with distance. The somewhat wider DRR distribution for the speech source shows that the output of the EC-DRR system is not perfectly independent of source power.

### B. Evaluation with synthetic stimuli

To verify that reverberation contributed to the variation of DRR with distance, the same set of spatial configurations as

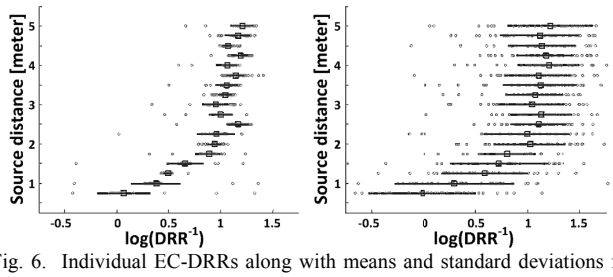


Fig. 6. Individual EC-DRRs along with means and standard deviations for real data; noise source (left); speech source (right).

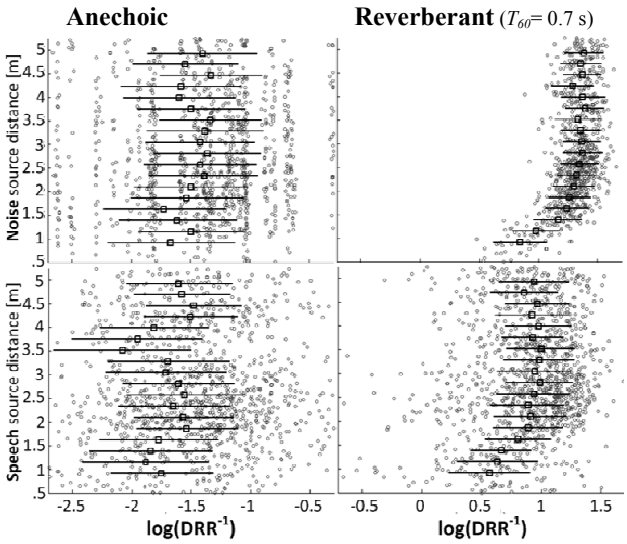


Fig.7. EC-DRRs for simulated pink noise (top) and speech (bottom) data in anechoic (left) and reverberant (right) spaces.

used in the real room recordings was simulated using a room simulator based on the image source algorithm (see Section VI.A) to evaluate the effect of reverberant ($T_{60}$= 0.7 s) versus anechoic conditions. As before, pink noise and speech sources were used, with responses collected at the ears of a simulated KEMAR head model. 1000 DRR values were collected with distances ranging from 0.75 m to 5 m, discretised into 18 states. Although a diffuse sound field cannot be perfectly approximated due to the computational complexity of simulating a very large number of imaging sources, a similar pattern was found to be the real world case. Comparing the top and bottom rows in Fig. 7, a narrower DRR distribution of the noise source was found than for the speech source. Fig. 7 also demonstrates the presence of a systematic DRR effect in the reverberant space but not in the anechoic space. The observation of an effect up to 2.5 m in both real and synthetic cases suggests that the room volume might impose a constraint upon the effective DRR operating range. DRRs for sources exceeding 2.5 m show a smaller change with respect to the increase of distance, and beyond 3 m there was no relationship with distance.

Bronkhorst and Houtgast [9] reported a similar effect with a 2 m upper boundary in a 65 m$^3$ room space ($T_{60}$= 0.5 s) and recognized it as the "auditory horizon" effect identified by Mershon and colleagues [3, 4]. They accounted for this effect via a direct energy calculation with a fixed-length integration window after signal onset. Beyond the auditory horizon, the calculated direct energy stopped decreasing further as a function of distance since the real direct energy was considered small compared to that of the fraction of reverberant energy included in the integration window.

## IV. EC-DRR Likelihood Function

Due to the absence of a simple analytic relationship between DRR and distance, a trainable probabilistic modelling approach was adopted. Given a DRR observation measured from the current frame, it is possible to obtain a likelihood function as shown in schematic form in the right part of Fig. 8, which estimates listener-source distance based on the previously collected training data in the same environment. Training data is represented as a Gaussian mixture (GM) with each component mapped to a discretized distance value with mean and variance describing the distribution of log(1/DRR) measurements around this distance range. In the schematic example of Fig. 8, the distance space is discretised into 8 segments and forms an eight-component GM used to derive the distance likelihood function.

With reference to the EC-DRR GM parameters, Gaussian means $\mu$ and variances $\sigma^2$ essentially differ with the reverberation properties, e.g. reverberation time. Given training stimuli of a particular room space, GM parameters can be learned through the EM algorithm [21]. There is no guarantee that EM will converge to the global maximization unless appropriate initial conditions are used. Means $\mu$ and variances $\sigma^2$ of the EC-DRR GM can be initialized based on the statistics ($\gamma_{range/max/min}$) of $L$ training stimuli for a $K$-element GM as shown in equations (9) to (13). The distance space is uniformly discretized into $K$ states. The smallest sampled distance state is mapped to the first Gaussian element.

$$\gamma_{\min} = \min_{\forall l} \log(DRR(l)^{-1}), l = 1 \ldots L \tag{9}$$

$$\gamma_{\max} = \max_{\forall l} \log(DRR(l)^{-1}) \tag{10}$$

$$\gamma_{range} = \gamma_{\max} - \gamma_{\min} \tag{11}$$

$$\mu(i) = \gamma_{\max} - \gamma_{range} \cdot (e^{(K-i/K)} - 1)/(e-1), i = 1 \ldots K \tag{12}$$

$$\sigma^2(i) = \gamma_{range} \cdot \frac{K - i + 1}{K^2} \tag{13}$$

The effect is to initialize components at smaller distances with higher variances, while component means increase logarithmically with distance. An example is illustrated in the left panel of Fig. 8.

Values of *log*(1/DRR) for 12241 virtual sources with various listener-source distances up to 18 m in a 18 m by 18 m by 10 m reverberant indoor space ($T_{60}$= 0.7 s) are plotted with their

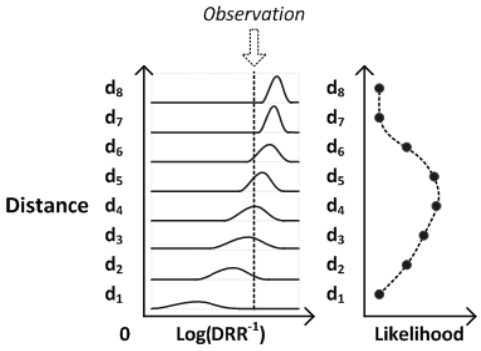


Fig. 8. EC-DRR likelihood function for source distance Bayesian filter. The left panel shows a Gaussian mixture (GM) of which each component is mapped to a discrete distance. The right panel shows the probability distribution over source distances derived from the GM's correspondence to the DRR observation noted in the dotted vertical line.

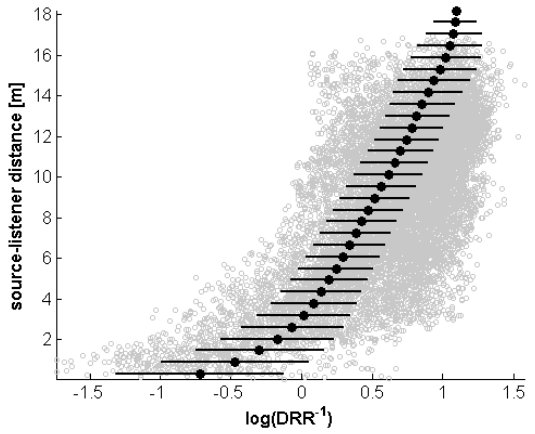


Fig. 9. DRR estimates derived from stimuli synthesised for a 18 m by 18 m by 10 m space with $T_{60}$= 0.7 s. Grey circles are plotted for each of the 12241 DRRs, along with the derived GM parameters.

means and standard deviations in Fig. 9. Again, as observed with the smaller room described in Section III, room volume seems to impose a constraint upon the effective DRR operating range, in this case up to 16 m. The same data set in Fig. 9 is redrawn in Fig. 10 separated into three sets of azimuth angles (near frontal, intermediate and lateral), using a greyscale encoding to further represent azimuth information within each range. It can be seen that, even at identical distances, stronger direct energy is yielded for sources in front of the listener (bottom panel of Fig. 10) relative to lateral sources (top panel). This figure makes it clear that EC-DRR varies as a function of both distance and azimuth, with less correspondence between *log*(1/DRR) and distance for lateral sources. This effect is considered as a natural limitation of EC-DRR owing to the deteriorating azimuthal resolution of the delay-line from frontal to lateral position [22].

The finding that EC-DRR does not correlate solely with distance suggests that the accuracy of distance estimation can be improved by introducing an additional variable, the source azimuth, in GM modelling. A DRR observation associated with a particular source distance is often misjudged due to overlapping Gaussian distributions. One way to minimise the overlap is to reduce Gaussian variance through the use of multiple Gaussian distributions to describe equally-distant sources with varying azimuth. Conditioned on the same training stimuli, multiple GMs, instead of one, are learned through the EM algorithm associated with different source azimuths. At least three GMs (each formulated as illustrated in the left panel of Fig. 8) were found to sufficiently enhance the distance inference performance in a pilot study. Upon receipt of new observations, ITD information helps to determine the appropriate GM for generating the likelihood function. No ITD-azimuth mapping is needed since only the delay in samples from the centre is required.

## V. Sequential Modelling for Temporal Integration of Distance Estimates

Previous sections presented an algorithm which processes binaural signals to produce instantaneous estimates of source

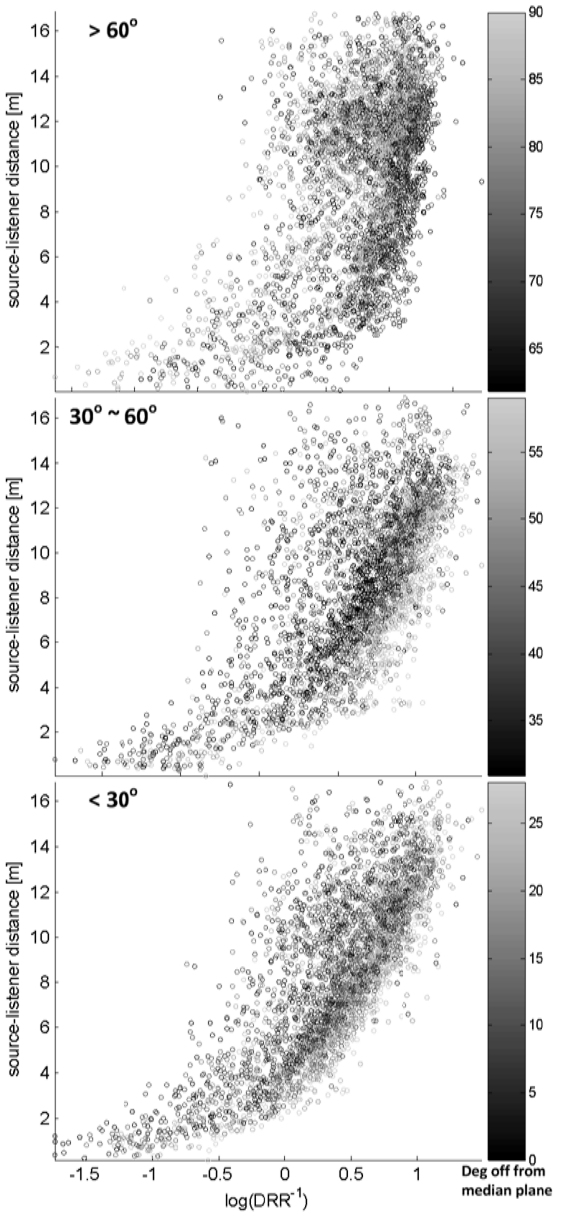


Fig. 10. Re-plotting of Fig. 9 with the associated azimuth information displayed. The lateral sources (at least 60° offset from the medial plane) are drawn at the top panel with a grey color representation of azimuth (60°:black; 90°:light grey). The middle and bottom panel display sources in the ranges (60°~30°) and (30°~0°) from the medial plane respectively.

distance based on the direct-to-reverberant energy ratio. In practice, instantaneous estimates can be affected by factors such as noisy observations, and fail to take account of any prior information in the case of moving sources. This section explores the localisation of a single static or moving source by a static virtual listener with a rotating head, using particle filtering. Particle filtering is a sampling-based approach to approximate the continuous distribution of belief about the system state (here, sound source location) via Monte Carlo simulation [23]. It allows the modelling of non-linear state variation and the continuous tracking under temporary dominance of observation noise. The use of particle filtering for

sound localisation is described in [24, 25]. Details of the PF architecture used in the current study can be found in [19].

In essence, PF operates by iterating 3 steps: prediction, update and estimation. Each iteration alters the particles (i.e. independent hypotheses over the system state) and associated particle likelihood weights based on a predictive model of the sound source dynamics with the likelihood transformed by updated observations. Estimated source location is then calculated as the weighted mean across all particles. The implementation of these steps in the current study is described next.

### A. PF prediction stage

If available, a model of source dynamics can be used to modify particle hypotheses at the PF prediction stage. Since no prior knowledge of source dynamics is assumed here, source motion is approximated by a zero-mean Gaussian noise term which is characterized by an adaptive standard deviation, itself a function of the corresponding particle likelihood weight. The observed dynamics in the current study is due to motion of the target source. Noise terms are applied to relocate source location hypotheses of all particles in both azimuth and distance in the PF prediction stage. Specifically, the state variable (i.e. hypothesised source location in distance $x_{r,t}^n$ and azimuth $x_{\phi,t}^n$) corresponding to particle *n* at time *t* is determined by

$$x_t^n = [x_{r,t}^n, x_{\phi,t}^n] \tag{14}$$

where the evolution of source azimuth location $x_t^n$ is formulated as an increment to the previous location $x_{t-1}^n$ with noise term $\mu_t^n$ As defined in (15), $\mu_t^n$ is a zero mean Gaussian variable with a time-variant standard deviation $\sigma_t^n$, which is a function of the particle likelihood weight $\omega_{t-1}^n$ and the constant $\sigma_{\max}$.

$$\mu_t^n = N[0,(\sigma_t^n)^2],\ \sigma_t^n = (2\cdot\omega_{t-1}^n - 1)\cdot\sigma_{\max} \tag{15}$$

Since $\omega_{t-1}^n$ ranges between 0 and 1, $\sigma_{\max}$ specifies the maximum value of the standard deviation. The effect is that a weaker particle is more likely to possess a larger noise term, encouraging a wider range of exploration of the state space. The altered location hypotheses are subsequently evaluated with the observation model introduced next to update particle weights.

A fixed maximum standard deviation, empirically determined as 15° ($\sigma_{\max,\phi}$) and 2 m ($\sigma_{\max,r}$) in azimuth and distance coordinates respectively, was applied across all particles. Note that the source azimuth is measured relative to the medial plane, as a function of head orientation. Particle hypotheses for source azimuth therefore need to be adjusted in accordance with head rotation.

### B. PF update stage

At the PF update stage, cues were fused by multiplying the likelihood functions associated with the two different localisation cues, namely ITD and EC-DRR, for updating each particle's likelihood weight. The derivation of the EC-DRR likelihood function according to the current observation and the pre-collected training data was demonstrated in Section IV. The cross-correlation function, (2), computed from the input binaural signal can be directly used as the ITD likelihood function for source azimuth as in the pseudo-likelihood approach of Ward *et al.* [25]. 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. As illustrated in the lower plots of Fig. 3, the cross-correlation function acts as a weight for each time delays, which are further mapped into source azimuths through the transfer function described in Section II.A.

In parallel to PF, the peaks of the likelihood function produce instantaneous estimates of source azimuth and distance. Instantaneous azimuth information is also used to calculate EC-DRR and select the correct GM for generating the EC-DRR likelihood function for a given azimuth range, as outlined in Section IV.

### C. PF estimation stage

The mean of the 20% of particles with the highest likelihood weight was used to generate source location estimates. This approach minimises the negative effect of weak particles and results in a degree of resistance to noisy observations. Since the instantaneous azimuth estimate is made without exploiting the head rotation cue, substantial localisation degradation is expected resulting from front-back reversals. This estimation error, which may be up to $\pi$, is principally on account of the use of the ITD cue rather than the detrimental effects of noisy environments. Therefore, the judgements suffering from front-back reversals occurring in the instantaneous estimation approach were rectified to the hemispheres used to generate the input stimuli. In this way, the advantage of sequential integration provided by PF can be properly assessed.

## VI. Evaluation

### A. Stimuli

The "Roomsim" room simulator [26] was used to generate binaural stimuli for various listener-source configurations and reverberation conditions. Non-individual HRTFs recorded with a KEMAR dummy head [20] with inter-ear distance of 0.152 m were used to generate binaural room impulse responses based on the image method [27]. Stimuli were generated at a sampling rate of 44.1 kHz. Temperature and relative humidity were set at 20° C and 40% respectively during the simulations. The virtual room was an 18 m by 18 m by 10 m gym space. This was generated in two reverberant conditions with $T_{60}$= 0.2 s and 0.7 s.

Training and test stimuli were generated by convolving a speech or pink noise source with the binaural room impulse responses. The simulated listener was static and placed in the south-west corner of the room in order to permit the use of quite large listener-source distances, all of which greater than 2 m, at no near-field distance cues are available [2], allowing a focus on the range where DRR cues are expected to dominate distance perception.

Variation of listener-source geometry at each time frame simulated motion of the sound source. The sound source was either static, moved linearly, or exhibited more complex motion which we refer to here as "zigzag motion". Each simulation run

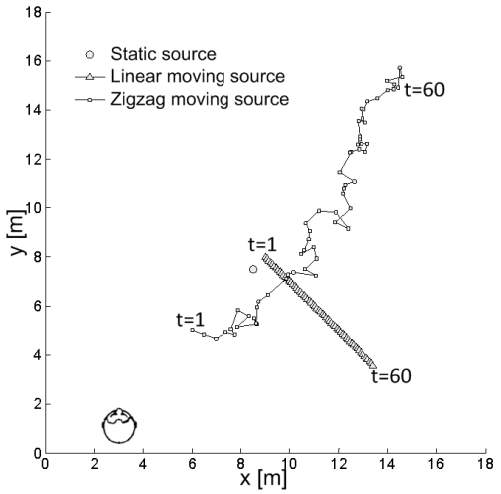


Fig. 11. Synthetic sound source dynamics. Virtual binaural stimuli were generated by a room simulator given the spatial attributes of the listener and the source. The position of the virtual listener as well as examples of linear and zigzag motion are shown.

consisted of 60 time steps of 0.75 s each. The linear source moved towards the SE corner of the room at a constant velocity of 0.14 $ms^{-1}$. The zigzag source moved in a northeasterly direction but varied in velocity from stationary to 0.42 $ms^{-1}$. Fig. 11 shows the source trajectory of each of the three cases.

### *B. Localisation Performance*

Localisation performance was measured using (i) the average Euclidean distance (AED) between the true and estimated source location; (ii) the unsigned estimation error in azimuth; and (iii) the unsigned estimation error in distance. The unsigned error refers to the absolute value of the difference between the estimate and the ground truth. To avoid undue influence from initial conditions, since particle hypotheses were set randomly, performance measures are the averages over frames 15-60 of the simulation. Frame 15 was chosen based on an offline analysis of the convergence of particle hypotheses in a development set using a frame convergence criterion described in [19]. Evaluation in the sequential case (i.e. using particle filtering) was based on arithmetic means over 100 simulations. By contrast, the AED of the non-sequential algorithm is constant between simulations.

Static auditory cues, DRR and ITD, were used for estimating source distance and azimuth respectively across six conditions consisting of two reverberant spaces and three source motions. Fig. 12 presents results for the non-sequential (DRR+ITD) and the sequential algorithms (PF+DRR+ITD). Particle filtering led to a large improvement in localisation performance in virtually all conditions. For static and linear moving sources, the estimation error increased as the level of reverberation increased, whereas motion complexity resulted in poorer distance estimation for zigzag moving sources with mild reverberation.

#### *1) Comparison to human listener performance*

Human listener localisation performance in azimuth and distance was tested in [1] using a similar experimental configuration as in the current study. Fig. 13 shows listener performance in localising a static sound source when the listener is placed in the center of a 18 m by 18 m by 10 m rectangular virtual space with $T_{60}$= 0.7 s, alongside estimates from the model for the same stimuli.

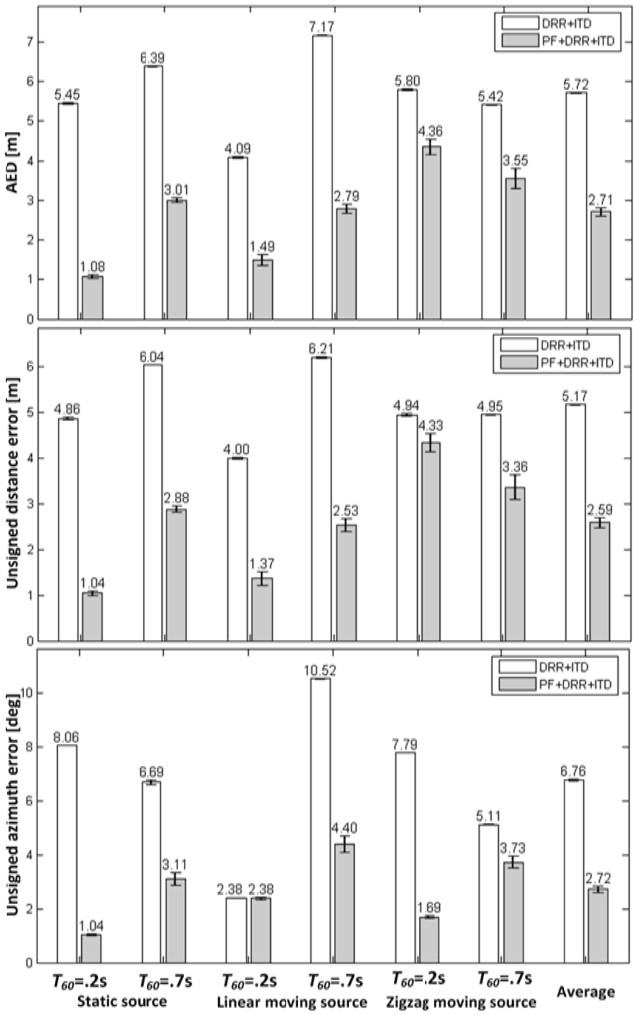


Fig. 12. Localisation performance comparison of sequential (PF+DRR+ITD) and non-sequential (DRR+ITD) algorithms as a function of different source motions and reverberation levels. Values above bars are mean AEDs (top panel), averaged unsigned estimation error of distance (middle panel) and azimuth (bottom panel) over 100 simulations of 45 s each. The rightmost bars show averages over the 6 conditions. Error bars denote 95% confidence intervals.

While listeners and the model resulted in very similar overall average Euclidean distance estimates, human listeners were less able to estimate azimuth but outperformed the model in distance.

An alternative measure of human distance estimation is provided by Zahorik [8], who proposed a compressive power function relationship, $D_e = k \cdot D_s^{\,a}$, between the perceived distance $D_e$ and the true distance $D_s$. Based on 84 data sets, Zahorik et al. [2] found a best fit at $a$=0.54 and $k$=1.32. This fit is shown in Fig. 14, although it should not be taken too literally since it is an average across a large range of conditions, only some of which are compatible with the configuration used in the current study.

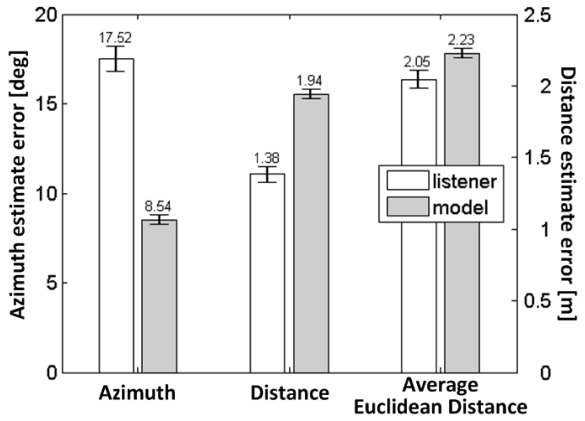


Fig. 13. Performance comparison between human listeners (from [1]) and the PF-based model of source localization in azimuth and distance. The left y-axis corresponds to azimuth errors while the right y-axis serves for both distance and average Euclidean distance.

Given the distance estimates obtained in our PF method, $a$ and $k$ were fitted 0.65 and 2.02 (the dotted line in Fig. 14). Note that EC-DRR can only function effectively in far-field (source distance beyond 2 m), while Zahorik's function is derived from multiple studies which used both near- and far-field stimuli. From the figure, distance underestimation in our model is not as large as predicted by Zahorik's curve, probably because the auditory horizon effect of EC-DRR is present at a rather distant region (see Fig. 9).

*2) Effect of reverberation and target motion on EC-DRR*

A longer reverberation tail is considered helpful in distance perception by enhancing later-arriving reflections which characterise the diffuse sound field. The room size controls the delays of early strong reflections, while the absorptive properties of reflected surfaces determine their magnitudes. For a fixed room size, a longer reverberation tail contributes to the existence of a diffuse sound field, offering better conditions for DRR cue utilization. Mershon *et al.* [5] reported that more accurate distance judgments were found for a more reverberant space, given time to reach sufficient familiarity with the room acoustics. A similar effect was also found in Bronkhorst and Houtgast [9], who discovered a closer "auditory horizon" for a space with less reverberation (in their case, with $T_{60}$= 0.1 s vs. 0.5 s). As a result, distance underestimation was less frequent when the room reverberation was higher compared to the less reverberant case.

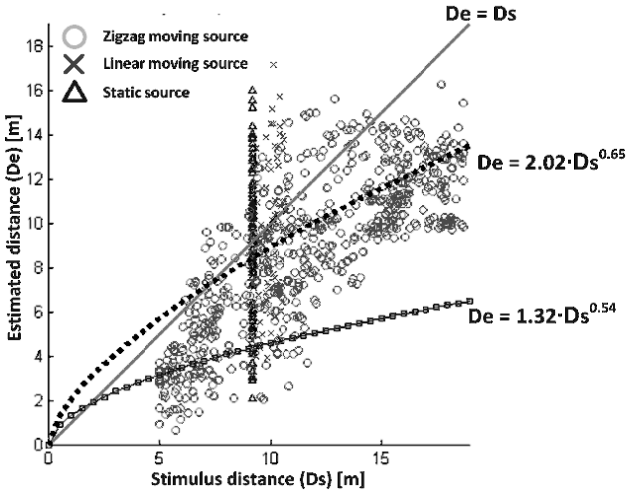


Fig. 14. Estimated distance as a compressive power function of stimulus distance. The stimulus-estimate relationship of outputs from PF runs for the three different target motions are presented. The black solid line ($D_e$=$D_s$) is the ideal relationship. Fitted power function curves are generated for data in [2] ($D_e$=1.32·$D_s^{0.54}$, squares) and in this paper ($D_e$=2.02·$D_s^{0.65}$, diamonds).

The distance estimation results of the current study were analysed using a two-way repeated-measures ANOVA. Reverberation time and source motion were treated as between-group factors so as to study the effect of reverberation level and auditory horizon on estimation performance. The interaction of the two between-group factors, source motion × $T_{60}$, was significant [$F$(2,594) = 231.4, $p$ < 0.001]. *Post hoc* analysis, using Tukey's HSD, revealed that judgements for zigzag moving sources were significantly less accurate than those for the other two ($p$ < 0.001). The performance for $T_{60}$= 0.2 s was better than $T_{60}$= 0.7 s for static and linear moving sources ($p$ < 0.001), while the reverse was true for the zigzag moving sources ($p$ < 0.001). We return to this conflicting finding below after examining the differences between source motions.

The simulated sources with zigzag motion have a wider range of stimulus distance variation than the other source motions, as shown in Fig. 15. The listener-source distance was a constant (9.22 m) for the static sources. Stimulus distance variations of up to 1.55 and 16.9 m were possible for the linear moving sources and the zigzag moving sources respectively. The possible EC-DRR distribution of zigzag sources is indicated by the region of the larger box in Fig. 16, whereas the smaller box contains the other two motions. Thus, a more limited range of EC-DRR is seen in the stimuli of static and linear sources (0~1.25 in the dimensionless horizontal axis) when compared to that of zigzag sources (-0.25~1.5). The range of the training stimuli for the $T_{60}$= 0.7 s condition is -1.25 to 1.5. The heights of the boxes in Fig. 16 are determined by corresponding curves in Fig. 15.

Fig 16 displays the GM priors for the EC-DRR likelihood function in the low and high reverberation conditions. These

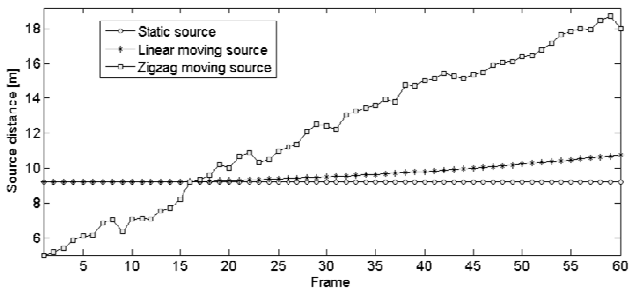


Fig. 15. Listener-source distance variations for three simulated source motions. The static sources were 9.22 m away from the listener, while varying distances from 9.19 to 10.74 m and from 5 to 18.69 m were specified for the linear moving sources and the zigzag moving sources respectively.

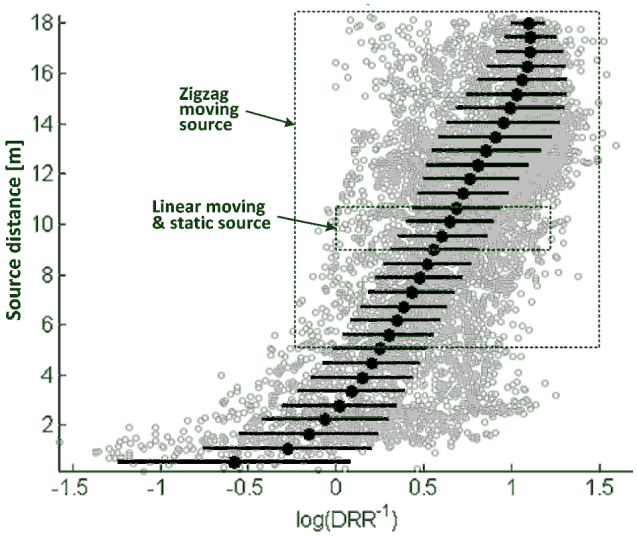


Fig. 16. The operating regions of EC-DRR likelihood function ($T_{60}$= 0.7 s) for stimuli corresponding to the three simulated source motions.

priors were learned from the training stimuli synthesised for the same gym space. As shown, the greater reverberant energy in the $T_{60}$= 0.7 s condition leads to larger values of log-inverse EC-DRR. Continuously curves are formed by connecting the Gaussian means. For $T_{60}$= 0.7 s, log-inverse EC-DRR increases as the source distance increases. However, for $T_{60}$= 0.2 s, only a slight increase occurs for distances beyond 12 m along while the variance continues to increase. This may result from the "auditory horizon" phenomenon. The auditory horizon of the low reverberation condition is indicated under the box in Fig. 17, indicating the operating range applied to zigzag sources. Consequently, it may explain the resulting less accurate distance estimations in low reverberation for the increasing number of sound sources falling beyond the auditory horizon. By contrast, by operating in the listener-source distance region away from the auditory horizon effect, distance estimates for the static and linear sources became more accurate in low reverberation. This suggests that energy fluctuations from

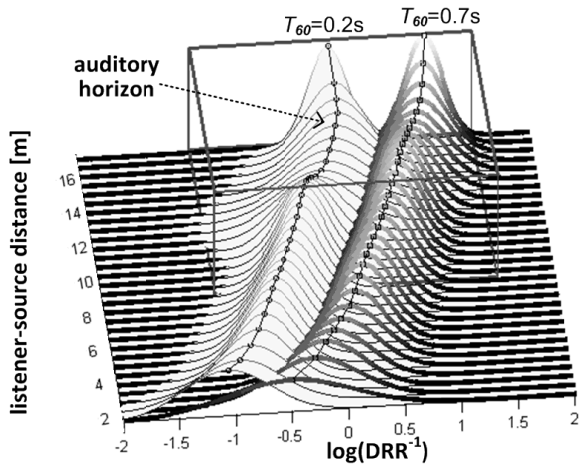


Fig. 17. Two Gaussian mixtures for generating EC-DRR likelihood function associated with $T_{60}$= 0.2 s (grey) and 0.7 s (black).

reverberation are more influential when the auditory horizon effect is absent. The measurements of EC-DRR were disrupted more severely in high reverberation and led to less accurate location estimates.

*3) Benefit of sequential modelling*

As shown in the top panel of Fig. 12, the localisation error of PF can be reduced to about half of that for approaches which do not involve PF. The effect of sequential modelling on AED was analysed using a one-way repeated-measures ANOVA with two within-group factors (source motion × $T_{60}$). The sequential (PF) algorithm was significantly superior to the non-sequential (non-PF) algorithm [$F(1,198) = 956.2$, $p < 0.001$]. Highly correlated performance distributions are found for the mean AED (the top panel of Fig. 12) and the distance error (the middle panel) except for the condition of non-PF under zigzag motion.

Simulation results for both PF and non-PF algorithms at $T_{60}$= 0.2 s for static sources are drawn as a function of time frame in Fig. 18. Static sources only are presented here so as to examine the effects of sequential modelling independently from effects of source motion. From the upper panel, the distance errors of the non-PF approach show little temporal correlation and range roughly from 1 to 9 m for 9.22 m distant stimuli. On the contrary, PF demonstrates a smoothly decreasing distance error. This decrease stops when it reaches the best EC-DRR can offer, i.e. 1 m error from the figures of non-PF.

Considering the azimuth component, spurious peaks are present in the ITD likelihood function which may disrupt non-PF azimuth localisation. By contrast, the PF algorithm is relatively insensitive to disturbance caused by reverberation.

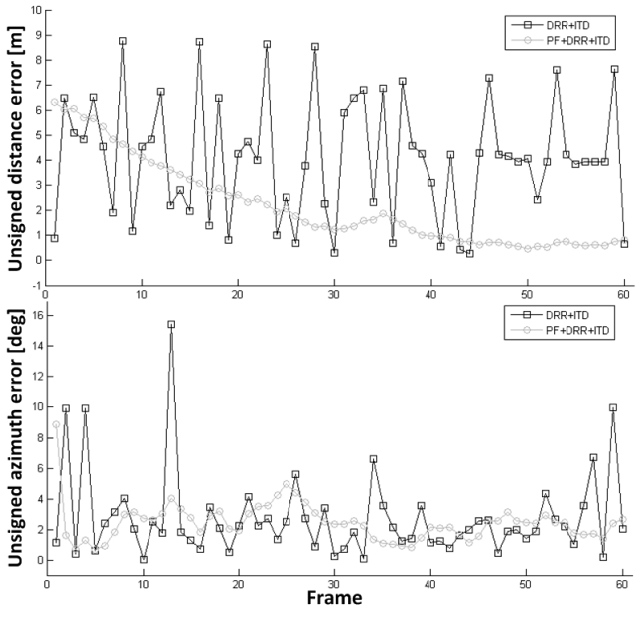


Fig. 18. The running unsigned error of distance (upper) and azimuth (lower) for the static sources with $T_{60}$= 0.2 s. Each symbol represents an average of over 100 simulations for the PF methods. The non-PF methods show more variation in estimates from frame to frame than the PF methods.

PF particles close to the true source position are constantly supported by that part of the likelihood function, which may produce a lower weighing due to the existence of spurious peaks. This effect is depicted in the lower panel of Fig. 18. A rather smoother azimuth error is found for PF than non-PF despite the fact that the latter has been improved by correcting for front-back reversals. It seems likely that the sequential approach embodied in the PF framework provides some robustness to occasional observation errors, leading to more stable performance as a result of temporal cue integration.

It is generally found across conditions that the PF-based estimate for the azimuth component exhibits a faster convergence (sometimes showing convergence as early as the second frame, as exemplified in Fig. 18) than that for distance. This suggests that EC-DRR may be too sensitive to instantaneous energy fluctuations to be able to provide as precise a localisation estimate as ITD.

## VII. Discussion

The current study describes a system to extract useful information for determining sound source distance from reverberant binaural input. The direct-to-reverberant energy ratio was measured by assuming direct and reverberant signals can be separated with their differential incoming directions. A binaural equalization-cancellation (EC) technique was proposed to separate energy belonging to each lateral direction corresponding to a resolution of one delayed sample in a delay-line structure. The EC-based direct-to-reverberant energy ratio (EC-DRR) was calculated by dividing the energy aligned with the target location over the energy sum of the remaining directions. A cross-correlation based azimuth estimator was used to determine the target location.

During the evaluation of its potential as a distance cue for both simulated and real data, estimated EC-DRRs with substantially varying values even from a static sound source were commonly observed due to energy fluctuations of both target and background. Their correlation with source distance was only be observed in the longer term, i.e. under sequential examination over a few seconds.

An EC-DRR based likelihood function was developed in combination with interaural time differences to update estimates of sound source location in both distance and azimuth. The likelihood function was derived based on a given reverberation-related prior as a function of source azimuth, in the form of a set of Gaussian mixtures, uniquely determined for each room. Only the information of azimuth offset from the median plane is needed when integrating the reverberation-related prior. The latter can be derived as part of the process of learning the acoustics of the room. Our approach is hence insensitive to front-back reversals and consistent with Simpson and Stanton's finding [28] on the irrelevancy of head movements to distance perception.

The framework of particle filtering (PF) integrates evidence over time to reduce the distorting effects of errors associated with instantaneous observations. The use of PF nearly halves the error in localisation compared to instantaneous estimates. Further, PF smoothes out the variation in reverberation effects caused by the changing geometry between room surfaces and target source (or listener). PF also allows the learning of room acoustics (here, in terms of GM prior) independent from the positions of sound source and listener, and may offer a less complex alternative to the distance learning problem than presented in [12].

The auditory horizon effect was observed in both real and synthetic stimuli as a function of room reflection properties and probably led to the under-estimation of distance found over the upper part of the distance range (e.g. see Fig. 14). The auditory horizon is effectively an upper limit on perceived distance and may result from the mechanism that human listeners use to separate direct and reverberant energy components [9]. In the scheme used in the current study to separate direct energy from reverberation via a delay-line structure, an auditory horizon may be established if the direct signal is too weak (being sufficiently far away from the listener) to dominate the energy arriving from the target source azimuth.

The performance drop compared to real listeners suggests that the EC-DRR approach to distance estimation does not fully exploit the DRR cue. One weakness of the evaluated distance estimator is that EC-DRR was designed to operate principally in the non-near-field range i.e. for listener-source distances of greater than 2 m. It is known that other cues, such as interaural differences, are implicated in near-field distance estimation [2].

The evaluation used synthetic stimuli which reproduced the diffuse field with limited fidelity, perhaps resulting in a substantial amount of distance-dependent but non-direct energy. Further, the acoustic properties of synthetic stimuli were not completely unknown to the computational models during simulation: reverberation time served as a prior for configuring associated likelihood function parameters. Further work using both real stimuli and methods to learn reverberation time is required to overcome these limitations. The analytic function describing the relationship between DRR and listener-source distance can also be improved to reduce GM parameters by simplifying the learning of room reverberation.

A further development of the equalization-cancellation technique for identifying multiple source energies in the delay-line structure may be necessary to improve EC-DRR accuracy by removing strong early reverberation or interfering sources.

## VIII. Conclusions

The direct-to-reverberant energy ratio has long been considered as an absolute auditory cue in human distance perception. Traditional methods for extracting this energy ratio are based on the post-processing of room impulse response, whose estimation is computationally expensive and inaccurate in practice. An alternative, which was employed here, is to estimate the energy arriving from the direction of the direct source, under the assumption that reverberant components result in a spatially-diffuse sound field. We proposed a binaural equalization-cancellation technique to calculate this energy ratio (EC-DRR) by locating the direct energy in a delay-line structure.

Simulations using synthetic stimuli indicated that the degrading effect of reverberation could be effectively mitigated

by sequential cue integration for both distance and azimuth estimations. The localisation error of the PF approach is in this way reduced to 47% of that for instantaneous estimation. The advantage of particle filtering was found to be mainly attributable to better localisation in distance, suggesting the importance of sequential integration for auditory cues based on energy measurements.

The performance of source distance estimation using DRR can be accounted for largely in terms of a competition between auditory horizon and energy fluctuation effects. As found for the sources with non-linear motions, better distance localisation performance for PF was obtained in the more reverberant space with a farther auditory horizon. Although energy fluctuations increased with the level of reverberation, their detrimental effect was smoothed by PF cue integration over time.

## REFERENCES


[1] Y.-C. Lu, "Active hearing strategies for binaural sound localisation in azimuth and distance by mobile listeners," Ph.D. thesis, Computer Science department, Sheffield University, Sheffield, UK, 2009.

[2] 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,* vol. 91, no. 3, pp. 409-420, May/Jun., 2005.

[3] D. H. Mershon, and E. King, "Intensity and reverberation as factors in the auditory perception of egocentric distance," *Percept. Psychophys.,* vol. 18, pp. 409-415, 1975.

[4] D. H. Mershon, and J. N. Bowers, "Absolute and relative cues for the auditory perception of egocentric distance," *Perception,* vol. 8, no. 3, pp. 311-22, 1979.

[5] D. H. Mershon, W. L. Ballenger, A. D. Little *et al.*, "Effects of room reflectance and background noise on perceived auditory distance," *Perception,* vol. 18, no. 3, pp. 403-16, 1989.

[6] C. W. Sheeline, "An investigation of the effects of direct and reverberant signal interaction on auditory distance perception," Ph.D. thesis, Hearing & Speech Sciences department, Stanford University, 1984.

[7] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *J Acoust Soc Am,* vol. 112, no. 5, pp. 2110-2117, Nov., 2002.

[8] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *J Acoust Soc Am,* vol. 111, no. 4, pp. 1832-46, Apr, 2002.

[9] A. W. Bronkhorst, and T. Houtgast, "Auditory distance perception in rooms," *Nature,* vol. 397, no. 6719, pp. 517-520, Feb., 1999.

[10] E. Larsen, C. D. Schmitz, C. R. Lansing *et al.*, "Acoustic scene analysis using estimated impulse responses," in Proc. the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2003, pp. 725-729.

[11] A. W. Bronkhorst, "Modeling auditory distance perception in rooms," in Proc. EAA Forum Acusticum Sevilla, Sevilla, Spain, 2002.

[12] S. Vesa, "Binaural sound source distance learning in rooms," *IEEE Trans Audio Speech Language Process,* vol. 17, no. 8, pp. 1498-1507, Nov., 2009.

[13] K. Furuya, and Y. Kaneda, "Two-channel blind deconvolution for non-minimum phase impulse responses," in Proc. ICASSP, 1997, pp. 1315-1318.

[14] A. W. Bronkhorst, "Effect of stimulus properties on auditory distance perception in rooms," in Proc. 12th International Symposium on Hearing (ISH): Physiological and Psychological Bases of Auditory Function, Shaker, Mastricht, Holland, 2001, pp. 184-191.

[15] C. Liu, B. C. Wheeler, W. D. O'Brien, Jr. *et al.*, "A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers," *J Acoust Soc Am,* vol. 110, no. 6, pp. 3218-31, Dec., 2001.

[16] N. I. Durlach, "Note on the equalization and cancellation theory of binaural masking level differences," *J Acoust Soc Am,* vol. 32, pp. 1075-1076, 1960.

[17] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth *et al.*, *An efficient auditory filterbank based on the gammatone function,* TR 2341, Applied Psychology Unit Cambridge, UK, 1988.

[18] H. Viste, and G. Evangelista, "Binaural source localization," in Proc. 7th International Conference on Digital Audio Effects (DAFx), Naples, Italy, 2004, pp. 145-150.

[19] Y.-C. Lu, and M. Cooke, "Motion strategies for binaural localisation of speech of speech sources in azimuth and distance by artificial listeners," *Speech Commun.*, 2009 (submitted).

[20] W. G. Gardner, and K. D. Martin, "HRTF measurements of a KEMAR dummy-head microphone," *Standards in Computer Generated Music*, G. Haus and I. Pighi, eds.: IEEE CS Tech. Com. on Computer Generated Music, 1996.

[21] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society,* vol. B, no. 39, pp. 1-38, 1977.

[22] A. W. Mills, "On the minimum audible angle," *J Acoust Soc Am,* vol. 30, pp. 237-246, 1958.

[23] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing,* vol. 10, pp. 197-208, 2000.

[24] H. Asoh, F. Asano, T. Yoshimura *et al.*, "An Application of a Particle Filter to Bayesian Multiple Sound Source Tracking with Audio and Video Information Fusion," in Proc. Fusion, 2004.

[25] 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,* vol. 11, no. 6, pp. 826-836, Nov, 2003.

[26] 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.,* vol. 9, no. 3, pp. 48-51, 2005.

[27] J. B. Allen, and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J Acoust Soc Am,* vol. 65, no. 4, pp. 943-50, 1979.

[28] W. E. Simpson, and L. D. Stanton, "Head movement does not facilitate perception of the distance of a source of sound," *The American Journal of Psychology,* vol. 86, no. 1, pp. 151-159, March, 1973.



**Yan-Chen Lu** received the B.Sc. and M.Sc. degrees in electronics engineering from National Chiao-Tung University, Hsinchu, Taiwan, in 1997 and 1999. He is currently pursuing the Ph.D. degree in the computer science department, University of Sheffield, UK. His research interests lie in spatial audio perception and computational audition.

**Martin Cooke** received a B.Sc. in Computer Science and Mathematics from the University of Manchester in 1982 and a Ph.D. in Computer Science from the University of Sheffield in 1991. He has worked at the UK National Physical Laboratory, the University of Sheffield and is currently Ikerbasque Professor at the University of the Basque Country, Spain. His research interests include robust automatic speech recognition, speech perception, computational auditory scene analysis, active hearing, the effect of noise on speech production and second language acquisition of speech.