BINAURAL DISTANCE PERCEPTION BASED ON DIRECT-TO-REVERBERANT ENERGY RATIO

Yan-Chen Lu and Martin Cooke

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

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

ABSTRACT

The direct-to-reverberant energy ratio has long been recognized as an absolute auditory cue for sound source distance perception in listeners. Traditional methods to extract this energy ratio are based on post-processing of the estimated room impulse response, which is computationally expensive and inaccurate in practice. An alternative is based on estimating the energy arriving from the azimuth of the direct source, under the assumption that reverberant components result in a spatially-diffuse sound field. We propose a binaural equalization-cancellation technique to calculate this energy ratio by locating the source in a delay-line structure, and go on to demonstrate its potential as a distance cue for both simulated and real data. The system is integrated with a Bayesian inference framework, particle filtering, to handle the nonstationary of energy-based measurements. Experiments on simulated room data showed the resulting computational model is capable of estimating source distance based on reverberation information.

1. INTRODUCTION

Which auditory cues underlie distance judgments? While sound source intensity varies with distance, raw intensity information available to a listener conflates intrinsic sound source power variations and the effect of changing source-receiver distance. Variations in intensity cues over time can provide relative distance information to distinguish whether the source is approaching or moving away from the listener. In contrast, the direct-to-reverberant energy ratio (DRR) appears to contain the information necessary for listeners to perform absolute distance judgment, especially for far-field sources [1]. In a reverberant space, the energy contained in late reverberations can be considered as a function of source power and independent of the sound source location relative to reflective surfaces. The ratio between direct energy and reverberant energy can thus remove the source power effect to behave as an absolute distance cue while the former retains the information via its inverse relation to distance.

Listeners' ability to determine source distance from reverberation has been extensively studied [2-5]. Judgments are more accurate in a reverberant space than in an anechoic space and the judgment deviation is low among experiments [2]. Mershon and Bowers [3] further suggested that listeners treat reverberation as an absolute distance cue by giving accurate distance judgments at first stimulus presentation. Zahorik [5] concluded that the principal role of the DRR cue is to provide absolute distance information rather than to support fine distance discriminations. Consequently, DRR is poor as a relative cue. Digital reverberation algorithms which enable virtual auditory displays have been used in a series of psychophysical studies [4, 6, 7], whose outcomes also support the hypothesis that DRR cue has a great influence on listeners' ability to estimate sound source distance.

Bronkhorst and Houtgast [7] proposed a computational model to predict human distance judgment in a controlled condition where the DRR cue is dominant. This model demonstrated an accurate prediction on subjects' perceived distance based on the knowledge of certain acoustical properties of the environment (room volume, reverberation time and source directivity) using monaural data. Other models based on binaural signals used either prior knowledge of environment (e.g. room impulse responses [8]) or extensive training data [9] to estimate source distance. While these studies attempted to demonstrate that distance estimation can be further improved with binaural input, neither emphasized the role of directional information.

The first step in computing DRR is to segregate the direct and reverberant signals from the acoustic mixture. A common approach uses the difference in arrival time of the two components [7], usually applied by specifying an integration window for the room impulse response (e.g. treating 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 [10].

Unlike the temporal domain schemes outlined above, 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 spatial 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. [11] to address a different problem, that of separating multiple sources motivates our proposed approach. Their two-microphone system exploits location information to steer independent nulls that suppress 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 mean of separating direct signal from reverberant signal. The result is transformed into a DRR value which is then used to derive a likelihood function within a Bayesian inference system for sound source distance estimation.

2. PROPOSED SYSTEM

2.1. EC-DRR System Overview

This section describes a system for DRR estimation based on source directional information. It is fundamentally an equalization-cancellation (EC) operation applied on the binaural signal. The EC concept was proposed to explain the masking suppression process for situations in which there is only one noise source [12]. Equalization renders the magnitudes of noise components to be identical between channels, while cancellation subtracts the noise component in one channel from that in 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 DDR estimation system is outlined in Fig. 1. First, successive windowed frames of a binaural signal are processed by a pair of N-channel gammatone filterbanks [13]. Next, individual filter outputs feed two binaural interaction processes, cross-correlation (CC) and equalization-cancellation (EC), operating on an M-element delay line. A cross-frequency integration stage enables robust localization 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 residual of total signal energy S after subtraction of the direct energy component (Eq. 1).

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

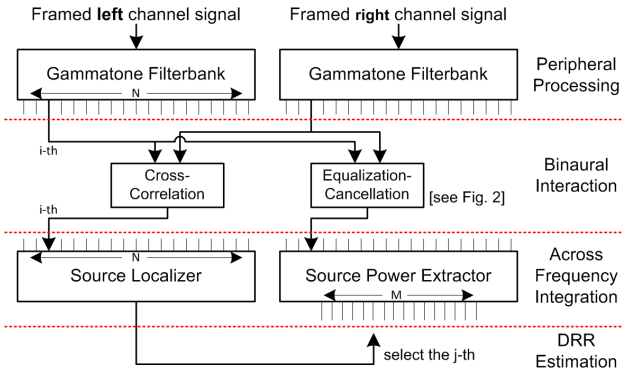

Fig. 1. Schematic diagram of EC-DRR system

2.1.1. Cross-correlation

Cross-correlation (Eq. 2) of the left and right filterbank outputs, $X_L(t)$ and $X_R(t)$, is inspired by Jeffress' coincidence detection model [14]. The coincident position along the delay-line, which is also known as interaural time difference (ITD), helps determine the source azimuthal location. Here, ITD is estimated by identifying the maximum value of averaged CC over frequency (Eq. 3).

$$CC(f,m) = \sum_t X_L(t) X_R(t+m), f = 1 \dots N, m = 0 \dots M-1 \tag{2}$$

$$ITD = \underset{m}{\operatorname{argmax}} \frac{1}{N} \sum_{f=1}^{N} CC(f,m) \tag{3}$$

2.1.2. Equalization-cancellation

A block diagram of the delay-line EC module is shown in Fig. 2. 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 Eq. 4

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

where T is number of samples within a frame. The compensation factor $E(f)$ used by the equalization block is updated every frame to generate Y_e, the compensated delay-channel signal.

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

$$Y_e(f,t) = Y(f,t)E(f) \tag{6}$$

The delayed channel is equalized with respect to the other channel to compensate for difference in intensity 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

$$SR_j(f) = \frac{T}{T-j} \sum_{t=1}^{T} |X(f,t) - Y_e(f,t-j)|^2, j = 0 \dots M-1 \tag{7}$$

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

$$D_j = \sum_{f=1}^{N} D_j(f) = \sum_{f=1}^{N} S_X(f) - SR_j(f) \tag{8}$$

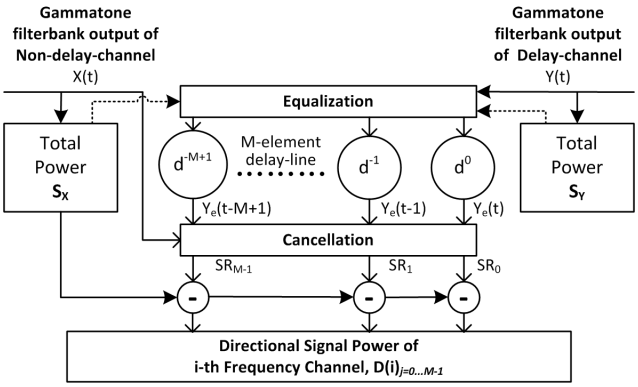

Fig. 2. EC module at i-th frequency channel in EC-DRR system.

2.2. Qualitative Investigation of EC-DRR

2.2.1. Evaluation of DRR concept

Ideally, DRR is a quantity that varies only with source distance and is independent of source power. These assumptions can be used to examine the effectiveness of the proposed EC-DRR system. A pink noise source with constant power was used to generate simulated audio sequences in an 18m by 18m by 2.75m rectangular space using Roomsim [15]. The distance between a simulated listener (KEMAR head model) and the noise source was increased from 2m to 11m with 0.31s reverberation time. The resulting binaural signals were converted into DDR values by the EC-DRR system. Calculated DDR increases as distance decreases in this simulated case shown at the upper part in Fig. 3, suggesting that the EC-DDR system does generate a distance related feature. The lower part in Fig. 3 displays total energy and estimated direct energy and corresponding reverberant energy as a function of source distance. The total and direct energy increased with decreasing distance. However, the predicted constant reverberation energy was not obtained. Instead, the estimated reverberant component also increased with decreasing distance, 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 reflected surfaces employed in Roomsim.

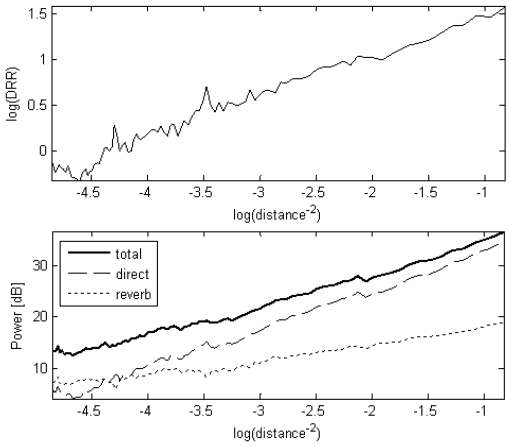

Fig. 3. DRR calculated by EC-DRR system (upper) and its segregated components (lower).

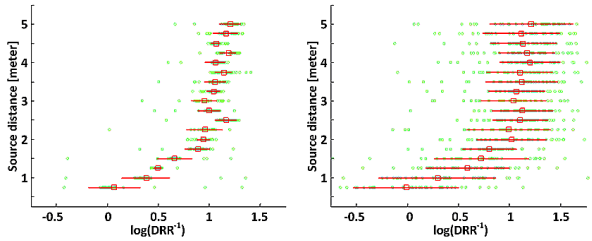

Fig. 4. DRR for real data; noise source (left); speech source (right).

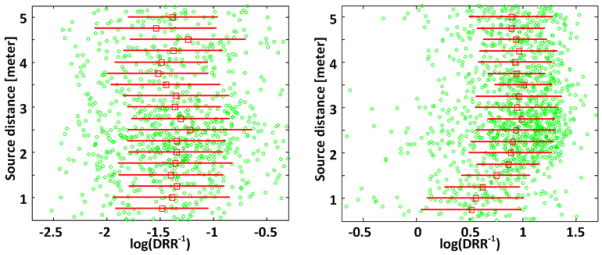

Fig. 5. DRR for simulated data; anechoic (left), reverberant (right).

2.2.2. Real data

Real test sequences collected from a 9m x 6m x 4m studio room were processed to compare with the simulated case. Two different static sources, pink noise or speech, were placed in front of a pair of microphones which were 10.6 cm apart. Eighteen different distances were chosen from 0.75m to 5m at 0.25m intervals. Recorded sequences for each distance were 10 s long and processed in 200 ms frames to obtain 50 DDR values per distance, and 900 in total. These estimates are depicted in Fig. 4 with respect to their distance to 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/DDR) for distances up to around 2.5 m. Thereafter, DDR shows less dependence on distance. The noise source has a narrower DDR distribution than that of speech source, and the width of the distribution narrows with distance. The somewhat wider DRR distribution for the speech source shows that the output of our EC-DDR system is not perfectly independent of source power.

2.2.3. Simulated data

To verify that reverberation contributed to the variation of DDR with distance, the same set of spatial configurations as that used in the real room recordings were simulated using Roomsim to evaluate the effect of reverberant (T60=0.76s) versus anechoic conditions. A speech source was used, with responses collected at the ears of simulated KEMAR head model. 1000 DDR values were collected with distance arbitrarily from 0.75 m to 5 m. The distance space was discretized into 18 states. Fig. 5 confirms the presence of a systematic DRR effect in the reverberant condition but not in the anechoic space. The observation of an effect up to 2.5 m suggests that the room volume might impose a constraint upon the effective DDR operating range.

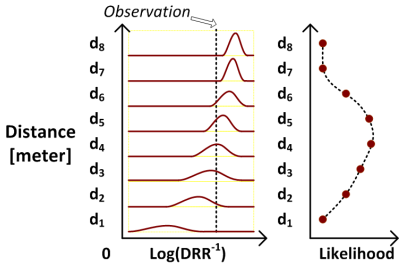

Fig. 6. EC-DRR likelihood function for source distance estimator. The left panel shows a Gaussian mixture model (GMM) where each mixture component is mapped to a discrete distance. The right panel shows the distance likelihood function derived from the GMM for the DDR value shown by the dotted line.

2.3. EC-DRR GMM

Given a DRR observation measured from the current frame, it is possible to obtain a likelihood function (Fig. 6) which estimates source-listener distance based on the previously collected training data in the same environment. Training data is stored in the form of a Gaussian mixture model (GMM). Each Gaussian is mapped to a discretized distance value with mean and variance describing the distribution of DRR measurements around this distance range. In the schematic example of Fig. 6, the distance space is discretized into 8 segments and forms an 8-element GMM used to derive the distance likelihood function.

The proposed EC-DRR GMM parameters need to adapt to changes in reverberation properties, e.g. reverberation time and the location of reflectant surfaces. Given a set of training data for the current environment, GMM parameters can be learned through EM [16]. There is no guarantee that EM will converge to a global maximization unless appropriate initial conditions are used. Equations 9-13 describe how the means $\mu(i)$ and variances $\sigma^2(i)$ of the EC-DRR GMM are initialized based on the statistics ($\gamma_{range/max/min}$) of L training data items for a K-element GMM. Note that DRR is expressed as its log-inverse, log(1/DRR). The distance space is uniformly discretized into K states. The smallest sampled distance state is mapped to the first Gaussian element.

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

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

$$\gamma_{range} = \gamma_{max} - \gamma_{min} \quad (11)$$

$$\mu(i) = \gamma_{max} - 0.582 \cdot \gamma_{range} \left(e^{(K-i/K)} - 1 \right), i = 1 \ldots K \quad (12)$$

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

The effect is to initialize components at smaller distances with higher variances, while component means increase logarithmically with distance. An example can be seen at the left panel of Fig. 6.

3. EVALUATION

In previous work [17], we have used dynamic cues to distance such as motion parallax and acoustic time-to-contact (or τ), within a particle filtering framework [18]. Here, we investigate the integration of the EC-DRR likelihood function into that framework. The resulting system then uses both binaural cross-correlation and monaural intensity measurements to infer source distance. The intensity-based acoustic τ cue is only applicable for sound sources with constant power, e.g. white noise, since it assumes that all changes in observed intensity are due to changes in distance and not in intrinsic source power. The introduction of the EC-DRR cue relaxes the constant power constraint, allowing the processing of nonstationary sources such as speech and music.

Particle filtering is an iterative application of the operations which alter the state variables and associated particle weights based on models of the sound source dynamics and the likelihood of the current observations. Particles represent hypotheses distributed in the target state space. Each iteration of the PF algorithm has three stages: prediction, update and resampling. At the update stage of particle filtering, cues to distance are fused by multiplying the likelihood functions for each independent cue (i.e. EC-DRR, acoustic τ and motion parallax) for renewing each particle's likelihood weight.

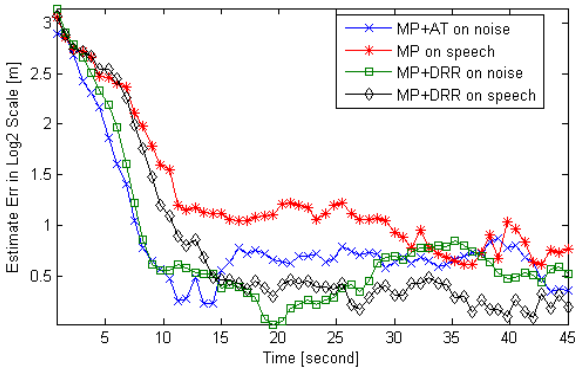

Fig. 7. Average estimation error in log2(metres) for various auditory distance estimation algorithms within a particle filtering framework. MP: motion parallax; AT: acoustic τ.

A similar experimental setup as that used in [17] was employed, viz. a single static sound source, either pink noise or speech, was placed in the centre of a 18m by 18m by 2.75m room space with a 0.76s reverberation time. At each time step, the simulated listener moved either straight ahead or $\pm 20^{\circ}$ forward (or more when necessary to avoid collision with the room boundary). Fig.7 presents the average distance estimation error along time over 30 different listener trajectories, each of which has 60 time steps (advancing in 0.75 s), 45 s in total.

We compared four different simulation settings in the experiments. All used the motion parallax cue which is based on triangulating the successive azimuth values relative to a known listener movement baseline. Two conditions additionally used the DRR cue, while one employed acoustic τ. This latter condition was applied only for the noise case due to the aforementioned issues with acoustic τ for non-stationary sources.

In general, all four methods gave more accurate estimates when information was pooled over more observations. The addition of the DRR cue produced a clear benefit for the speech source, but the benefit was less obvious in the case of noise. The method with DRR activated showed a better rate of convergence to the stable phase with a lower estimation error.

4. CONCLUSIONS

A new scheme to use reverberation information in estimating sound source distance is proposed based on binaural equalization-cancellation in a delay-line structure. The direct-to-reverberant energy ratio, which can serve as an absolute auditory distance cue, is estimated. The direct energy component of the source is estimated based on the azimuthal location of the source while the remainder is attributed to the diffuse reverberant component. Integrated alongside motion-based cues to distance in a particle-filtering framework, the addition of the direct-to-reverberant energy ratio cue leads to significant improvements in distance estimation for speech sources in a moderate-to-high reverberant space. Future work will develop the multiple source subtraction technique in delay-line structure to remove strong early reverberation or interfering sources.

5. REFERENCES

[1] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: a summary of past and present research," *Acta Acustica united with Acustica*, vol. 91, pp. 409-420, May/Jun. 2005.

[2] 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.

[3] D. H. Mershon and J. N. Bowers, "Absolute and relative cues for the auditory perception of egocentric distance," *Perception*, vol. 8, pp. 311-22, 1979.

[4] C. W. Sheeline, An investigation of the effects of direct and reverberant signal interaction on auditory distance perception, Ph.D. Thesis, Stanford University, 1984.

[5] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *J Acoust Soc Am*, vol. 112, pp. 2110-7, Nov. 2002.

[6] H.-G. Moon and Y. Choi, "Auditory depth control using reverberation cue in virtual audio environment," *IEICE Trans. Fund Elec, Comm & Comp Sci*, vol. E91-A, pp. 1212-1217, 2008.

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

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

[9] S. Vesa, "Sound source distance learning based on binaural signals," in *Proc. WASPAA*, Oct. 2007.

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

[11] C. Liu, B. C. Wheeler, W. D. O'Brien, Jr., C. R. Lansing, R. C. Bilger, D. L. Jones, and A. S. Feng, "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, pp. 3218-31, Dec. 2001.

[12] 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.

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

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

[15] 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, pp. 48-51, 2005.

[16] 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, pp. 1-38, 1977.

[17] Y.-C. Lu, M. Cooke, and H. Christensen, "Active binaural distance estimation for dynamic sources," in *Proc. Interspeech*, Aug. 2007.

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