Sampo Vesa. 2007. Sound source distance learning based on binaural signals. In: Proceedings of the 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2007). New Paltz, NY, USA. 21-24 October 2007, pages 271-274.

SOUND SOURCE DISTANCE LEARNING BASED ON BINAURAL SIGNALS

*Sampo Vesa**

Telecommunications Software and Multimedia Laboratory
Helsinki University of Technology
P.O. Box 5400, FIN-02015 TKK
sampo.vesa@tml.hut.fi

ABSTRACT

A learning approach for estimating sound source distance based on binaural signals is presented. The frequency-dependent coherence between the left and right ear signals is used as the distance cue. The distance estimation is based on pre-calculated coherence profiles and an energy-weighted likelihood function. The system is evaluated with different speech samples. The accuracy is best in the frontal direction and poorer in the sides, due to shadowing of the head. However, the system shows promise for scenarios where the sound source location is restricted and could be integrated with a two-channel azimuth localization system.

1. INTRODUCTION

In the field of auditory perception research, the term "localization" often refers to determining the direction of sound sources, i.e., the azimuth and/or elevation angles relative to the listener. Computational models for azimuth localization (also termed lateralization) have been developed for decades. However, a complete localization of a sound source includes an assessment of the source distance, in addition to direction. A computational method that could tell the direction and distance of a sound source, based on a binaural signal, could be useful in, e.g., augmented reality audio [1], intelligent hearing aids [2] and audio surveillance — any application, where information about the sound sources in the environment is needed. Investigation of computational estimation of sound source distance is thus well motivated.

In this work, a machine learning approach for estimating sound source distance based on binaural signals is adopted. The frequency-dependent coherence between left and right ear signals is used as the distance cue. The system is first trained using coherence calculated from white noise recorded at a discrete set of source-to-receiver distances in a room. In the testing phase, the most likely distance for each frame of a test signal (speech) is chosen based on maximum likelihood (ML) with energy weighting. The system is evaluated using data recorded in real rooms at five source-to-receiver distances.

The approach is adopted from [3], where the cross-spectral magnitude and phase differences were used for learning azimuth directions and trajectories using a microphone array of two or more microphones. The current research could be combined by that system so that the sound source direction is first calculated using the method described in [3], followed by distance evaluation using the presented method.

According to the knowledge of the author, methods for estimating sound source distance from a speech signal with one or two microphones do not currently exist. There have been attempts at estimating sound source distance from monaural room impulse responses [4].

2. THE METHOD

2.1. Sound source distance features

The direct-to-reverberant ratio has been long recognized as one of the main cues for the perception of sound source distance [5]. A quantity that is related to the direct-to-reverberant ratio is the magnitude-squared coherence (MSC) [6], which measures the linear correlation between two signals as a function of frequency. When the source-to-receiver distance in a room increases, the MSC can generally assumed to decrease, because the sound field at the receiver becomes more diffuse. Therefore, the magnitude-squared coherence was chosen as the distance feature in this work. The MSC is defined as

$$\hat{\gamma}^2_{lr}(f,t) = \frac{|\hat{G}_{lr}(f,t)|^2}{\hat{G}_{ll}(f,t)\hat{G}_{rr}(f,t)} \quad (1)$$

$$\hat{G}_{ll}(f,t) = \langle |X_l(f,t)|^2 \rangle \quad (2)$$

$$\hat{G}_{rr}(f,t) = \langle |X_r(f,t)|^2 \rangle \quad (3)$$

$$\hat{G}_{lr}(f,t) = \langle X^*_l(f,t) X_r(f,t) \rangle \quad (4)$$

where $G_{lr}(f,t)$ is the one-sided cross-spectrum between the left and right ear signals at time t and frequency f. $G_{ll}(f,t)$ and $G_{rr}(f,t)$ are the one-sided power spectra of the left and right ear signals, $x_l(t)$ and $x_r(t)$, respectively. Discretized versions of Eqs. (1)–(4) are used for practical calculations. The averaging operations (denoted by $\langle\cdot\rangle$) in Eqs. (2)–(4) are realized by a first order leaky integrator [6]

$$\langle Q(n) \rangle = \beta \cdot \langle Q(n-1) \rangle + (1-\beta) \cdot Q(n) \quad (5)$$

where $\beta = 0.5$ (time constant of 3.2 ms) defines the amount of smoothing, n is the discrete time index and Q is a generic function.

In both the training and the evaluation phase of the experiments, the coherence is calculated in overlapping signal frames of length 128 samples (2.9 ms when $f_s = 44.1$ kHz), with 25 % overlap and an FFT size of 1024 samples. A Hanning windowing function was used for windowing the time-domain segments. The time domain window length is much shorter than the FFT length in order to get a smoother coherence profile in the frequency domain.

* The author has received funding from the HeCSE graduate school and Tekniikan edistämissäätiö. The author wishes to thank Prof. Matti Karjalainen, Dr. Kalle Palomäki and Dr. Tapio Lokki for helpful comments. The author also thanks Miikka Tikander and Dr. Ville Pulkki for help with the recordings.

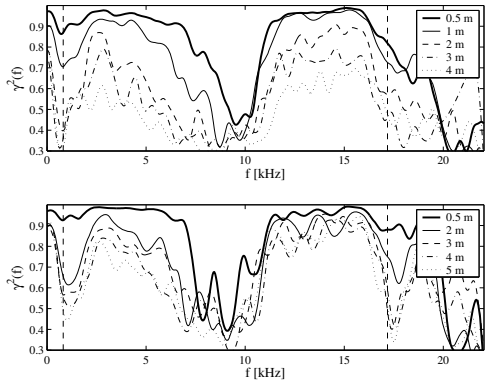

Figure 1: *Top: coherence profile for room 1 (listening room, azimuth angle* 0°*). Bottom: coherence profile for room 2 (meeting room, azimuth angle* 0°*). Vertical dashed lines indicate the range of likelihood evaluation (*K_L *and* K_U *in Eq. (6))*

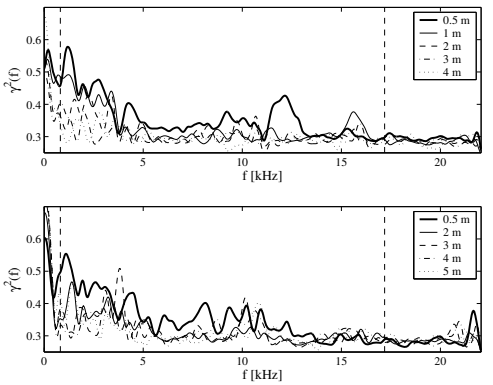

Figure 2: *Same as Fig. 1 but for azimuth angle* 90°.

2.2. Training the system

Eq. (1) is used for calculating the coherence for each frame of the training data, which consists of white noise sequences of 16 s duration convolved with binaural room impulse responses (BRIRs) measured at the set of distances of interest. The average of the coherence over time is calculated to get a coherence profile for each distance.

For the experiments, the coherence profiles were calculated at five discrete source-to-receiver distances in two different rooms. The first of the rooms (room 1) is a standardized listening room with reverberation time of 0.3 s at the 500 Hz third-octave band. The second room (room 2) is a larger meeting room with reverberation time of 0.7 s. The measurement distances in room 1 are 0.5 m, 1 m, 2 m, 3 m and 4 m. There is more space in room 2 so the distances were chosen to be 0.5 m, 2 m, 3 m, 4 m and 5 m. A Cortex binaural manikin was used for measuring the BRIRs. The length of each white noise burst used for training was 15 seconds.

The coherence profiles of the two rooms and the five source-to-receiver distances are plotted in Fig. 1 for an azimuth angle of 0°. There is a clear dip in the coherence profiles around 10 kHz in both rooms. Because the dip is present in profiles of both rooms, its existence has probably something to do with the head-related transfer function (HRTF) of the binaural manikin. When the azimuth angle is 90°, the coherence profiles are very different, as is evident in Fig. 2. Because the sound is now coming from the side, the coherence is generally lower due to the shadowing of the head. The profiles of different distances are now also closer to each other and therefore it is to be expected that the presented method will not work at larger azimuth angles.

2.3. Recognizing the most probable distance

The most probable distance is calculated by a maximum likelihood scheme, as in [3]. An energy-based weighting is applied in calculation of the likelihood, since the coherence does not carry any useful information at time-frequency elements not occupied by the sound source. For each frame n of the coherence $\hat{\gamma}^2_{lr}(k,n)$ and each considered distance d, the energy-weighted log-likelihood is formulated as

$$L(n,d) = -\sum_{k=K_L}^{K_U} \left(w(k,n) \frac{\left(\hat{\gamma}^2_{lr}(k,n) - \mu(k,d)\right)^2}{2\sigma^2(k,d)} \right) \quad (6)$$

where K_L is the lower frequency limit, K_U is the upper frequency limit, $\hat{\gamma}^2_{lr}(k,n)$ is the coherence estimated from frame n of the test signal, $\mu(k,d)$ is the mean of the coherence profile (see Fig. 1) for distance index d, and $\sigma^2(k,d)$ is the variance of the coherence profile. The weights $w(k,n)$ are calculated as

$$w(k,n) = -10\log_{10}\left(|X_l(k,n)|^2 + |X_r(k,n)|^2 + \epsilon\right) \quad (7)$$

where $|X_l(k,n)|$ and $|X_r(k,n)|$ are the left and right ear magnitude spectra for frame n, and ϵ is a small constant (`eps` in MATLAB). The minus sign in Eq. (7) is necessary because small energy values result in large negative values when expressed as logarithmic and the likelihood needs to be given a large positive weight $w(k,n)$ at those frequencies, due to the minus sign in Eq. (6).

The most probable distance for each frame is the one that gives the maximum value for Eq. (6). However, not all frames of the signal contain information on the source-to-receiver distance. When there is silence, the likelihood function can not provide any distance information. Therefore, the parts of $L(n,d)$ that carry useful information have to be identified. Due to the energy weighting, $L(n,d)$ should generally have high values when there is high energy in the input signal. In order to identify these parts, $L(n,d)$ is normalized by subtracting its maximum evaluated over all (n,d) using

$$L'(n,d) = L(n,d) - \max_{\forall n,d}\{L(n,d)\} \quad (8)$$

This preserves the mutual order between the likelihoods evaluated at different distances d. All frames for which $L'(n,d)$ is below a certain threshold T_L are then discarded from evaluation of the distance. For each of the remaining frames, the most probable distance is the one that maximizes $L'(n,d)$. Fig. 3 illustrates $L'(n,d)$ calculated from a short speech segment, the spectrogram of which is also depicted. One can see that those parts of the signal

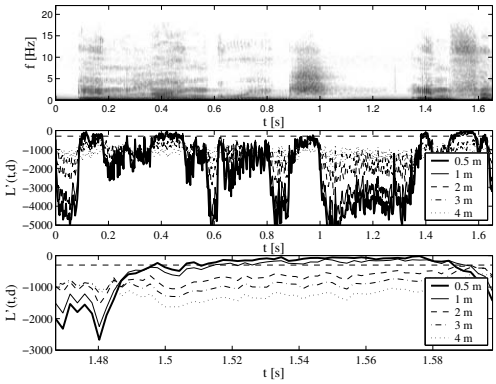

Figure 3: *Top panel: spectrogram of a speech signal segment (average of channels, true distance 0.5 m, room 1, azimuth angle* 90°*). Middle panel: corresponding normalized likelihood function evaluated at different distances. Dashed horizontal line shows the threshold value (*$T_L = -300$*). Bottom panel: zoom into a segment of the likelihood function.*

that have a wideband signal result in a high value of the likelihood and are therefore used for distance estimation.

For the experiments, the evaluation ranges were set to $K_L = 20$ ($\approx$ 818 Hz) and $K_U = 400$ ($\approx$ 17.2 kHz). The lowest frequencies are not included, since the coherence always tends to be large at low frequencies [6]. The upper frequency limit is motivated by the fact that there is no significant energy above that limit in the speech samples that were used as test signals. The means $\mu(k, d)$ are depicted in Fig. 1 and the variances $\sigma^2(k, d)$ are set to a constant value of one. The constant variance was chosen because the variance of the coherence calculated from the training signal, i.e., stationary filtered white noise, is not the same as the variance of the coherence calculated from a non-stationary speech signal. Furthermore, it is hypothesized that the variance of the coherence over time does not give reliable information on the sound source distance. Verification of this hypothesis is outside the scope of this work. When the azimuth angle is 0°, the threshold is set to $T_L = -300$ for room 1 and to $T_L = -200$ for room 2. For azimuth angle 90°, the threshold is set to $T_L = -90$ for room 1 and to $T_L = -100$ for room 2.

3. EVALUATION

The distance learning method was evaluated using recordings made in the two different rooms described previously. The coherence profile was calculated for five different distances (see the legends in Fig. 1) and two different angles (0° and 90°). The test signals were made by playing back two anechoic speech samples of 5.7 s duration, using a Genelec 1029A loudspeaker at the same locations as the BRIRs were measured at, and recorded by the Cortex binaural manikin. The test signals were recorded at two different sound source levels to ensure that variation in the sound source level does not affect the algorithm performance.

Figs. 4 and 5 show four frame-level confusion matrices for rooms 1 and 2, with azimuth angle 0°, calculated with the four different source signal and level combinations. Each matrix element (i, j) indicates the percentage of frames classified to distance j, when the true sound source distance is i. In other words, the rows indicate the true distance and the columns the estimated distance. The percentages have been rounded to the nearest integer for convenience. Thus, all of the rows do not sum to exactly 100 %.

It is evident in Figs. 4 and 5 that the correct distance is found in each case, even though on the frame level, the classification percentage seems to vary. It is notable that the shortest and longest distances are classified the best on frame level. In room 1, the confusion of individual frames seems to happen mostly with distances closest to the true distance (e.g., 0.5 m and 2 m for true distance of 1 m). There seems to be more confusion with the longer distances in room 2. The reason for this probably has something to do with the coherence profiles for room 2 (Fig. 1) being quite close to each other, for distances $\geq$2 m. A +5 dB boost in the playback level has an effect on the classification on frame-level — some of the percentages are higher and some are lower when the gain is increased.

When the manikin head was oriented differently, i.e., the azimuth angle is 90° (Figs. 6 and 7), the performance of the method is much worse. Some of the distances are correctly classified but with a much smaller margin. There is also a lot more confusion between different distances. This is due to the coherence profiles being very close to each other (see Fig. 2). Also, quite a small percentage of frames (average 2.6 %), is used for classifying the distances compared to the zero azimuth case (average 6.9 %). This also explains why there are large differences in classification between the two different playback levels. A small change in classification significantly affects the percentages when a small amount of frames is used for evaluating the likelihood.

In these experiments, all the parameters were chosen manually. For a real-world algorithm, the threshold T_L should be adjusted by an adaptive algorithm. Other approaches for finding the informative parts of the likelihood function could also be considered. The current way of normalizing the likelihood by its maximum is also not suitable for a real-time implementation as such, since the global maximum is needed.

4. CONCLUSIONS

A learning approach for estimating the sound source distance from a binaural signal is proposed. The algorithm is suitable for situations where the microphones are stationary and when the sound source is in frontal direction related to the listener. A more thorough investigation into the behavior of the coherence profiles in different source-to-receiver configurations could be conducted as future work. Real-head recordings should also be tried, since human localization performance is better with real-head recordings than with recordings made using a binaural manikin [7]. Other cues besides the coherence should also be investigated, as well as more powerful learning algorithms that could take the temporal evolution of the features into account.

5. REFERENCES

[1] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, "Augmented Reality Audio for Mobile and Wearable Appliances," *Journal of the Audio Engineering Society*, vol. 52, no. 6, pp. 618–639, 2004.

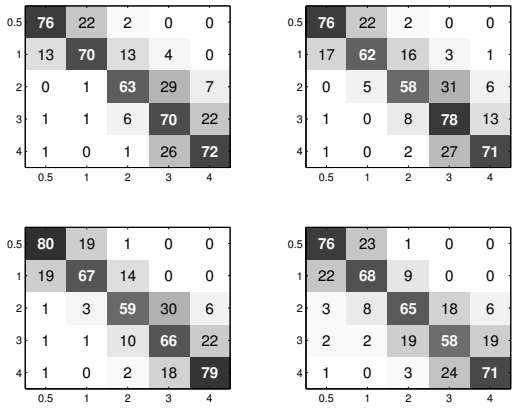

Figure 4: *The frame-level confusion matrices of distance learning for room 1 (listening room, azimuth angle* 0°*). Top row: male speech. Bottom row: female speech. Left column: gain +0 dB. Right column: gain +5 dB. The values have been rounded to nearest integers for clarity.*

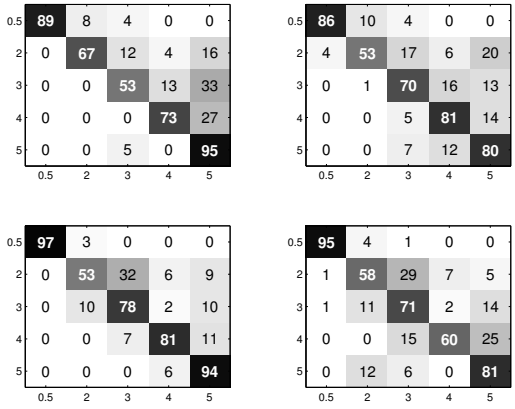

Figure 5: *Same as Fig. 4, but for room 2 (meeting room, azimuth angle* 0°*).*

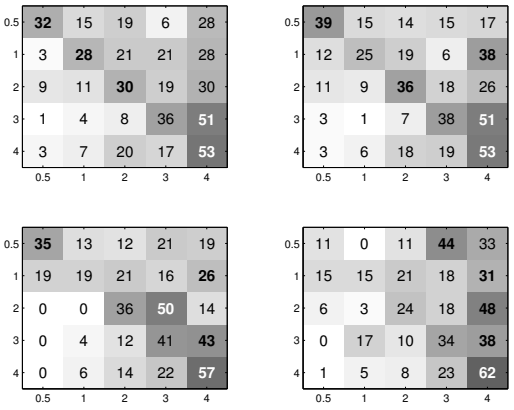

Figure 6: *Same as Fig. 4, but with azimuth angle* 90°.

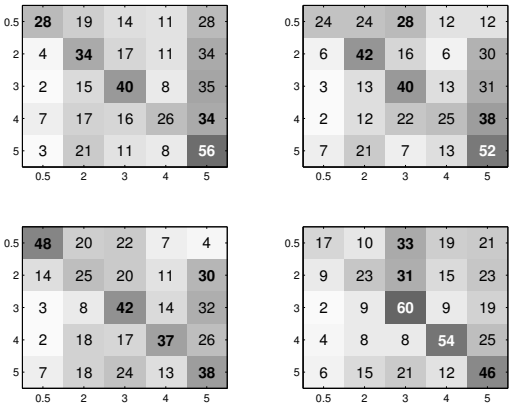

Figure 7: *Same as Fig. 4, but for room 2 (meeting room, azimuth angle* 90°*).*

[2] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal Processing in High-End Hearing Aids: State of the Art, Challenges, and Future Trends," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 18, pp. 2915–2929, October 2005.

[3] P. Smaragdis and P. Boufounos, "Position and Trajectory Learning for Microphone Arrays," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 358–368, 2007.

[4] E. Larsen, C. D. Schmitz, C. R. Lansing, W. D. O. Jr., B. C. Wheeler, and A. S. Feng, "Acoustic Scene Analysis Using Estimated Impulse Responses," in *Proceedings of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, November 2003.

[5] P. Zahorik, "Auditory Distance Perception in Humans: A Summary of Past and Present Research," *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 409–420, 2005.

[6] T. Wittkopp, "Two-Channel Noise Reduction Algorithms Motivated by Models of Binaural Interaction," Ph.D. dissertation, Carl von Ossietzky University Oldenburg, March 2001.

[7] P. Minnaar, S. K. Olesen, F. Christensen, and H. Møller, "Localization with Binaural Recordings from Artificial and Human Heads," *Journal of the Audio Engineering Society*, vol. 49, no. 5, pp. 323–336, 2001.