Binaural Sound Source Distance Learning in Rooms

Sampo Vesa

***Abstract*—A method for learning the distance of a sound source in a room is presented. The proposed method is based on short-time magnitude-squared coherence between the two channels of a binaural signal. Based on white noise as the training data, a coherence profile is obtained at each desired position in the room. These profiles can then be used to identify the most likely distance of a speech signal in the same room. The proposed approach is compared to a previous method for learning the position of a sound source. The results indicate that the both methods are able to identify the distance of a speech sound source correctly in a grid with 0.5-m spacing in most cases, when the orientation of the listener is 0°, 30°, 60°, 90°, or 180° on the horizontal plane.**

***Index Terms*—Binaural signal, coherence, distance measurement, localization.**

I. Introduction

SOUND source distance is the line-of-sight distance between a sound source, e.g., a loudspeaker, and a receiver, i.e., a human listener or a microphone. Knowledge of the distance of a sound source is useful in any application that could benefit from knowing the spatial location of sound sources in the environment, e.g., augmented reality audio [1], intelligent hearing aids [2], and audio surveillance [3]. In ubiquitous computing, all kinds of context information around the user may be valuable, and sound source distance information could be useful in that context as well. The problem of sound source distance estimation can be solved by using more than two microphones, but in many applications only binaural signals are available, which makes binaural distance estimation a relevant problem.

In this paper, binaural sound source distance estimation is approached as a learning problem. A feature related to sound source distance, namely the frequency-dependent coherence between the left and right ears, is calculated for each signal frame of a training signal, which is a white noise burst. A classification system is then trained using the extracted features. In the testing phase, a speech signal is played back at the same locations and the most likely distance for each short-time signal frame is chosen based on energy-weighted maximum likelihood. The system is tested in two receiver positions in two different rooms and five different orientation angles of the receiver. The generalization performance between the two receiver locations is also investigated. The proposed method is compared to a previous method of sound source position learning presented by Smaragdis and Boufounos [4].

The work described in this paper is a continuation of previous work by the author [5] with improved methods and a larger set of test recordings. In the previous work, there were only five distances, two listener orientations, and one receiver position in each of the two rooms. The larger set of recordings used in this paper has eight discrete distances in two receiver positions, and five listener orientations. The effect of a few key parameters is investigated in more detail, whereas in the previous work these parameters were chosen heuristically. The method itself has been modified by replacing the threshold parameter with a percentage parameter, which indicates the number of frames in percent having the highest likelihood to be included in recognizing distance. Low-passed spectrum estimates are employed in calculating the weighting function instead of instantaneous spectra. A slight modification is also proposed to a previous method of sound source position learning [4], which allows the discrimination of sound source distance, when the direction (azimuth or orientation angle) stays the same. The ability of both methods to learn the orientation angle is also investigated.

II. Previous Work

Human perception in estimating sound source distance has been studied extensively (e.g., [6]–[13]). The most important cues for distance perception in humans have been identified, see [13] for a review. Modeling the results of human perception has also gained some attention (e.g., [14], [15]). In addition, rendering of sound source distance in virtual acoustics has been researched (e.g., [10], [11], [16], [17]). However, there is very little research on actual computational models of distance estimation. This is not the case with azimuth localization, on which there is a lot of research in both human perception and computational models since the 1940s (e.g., [18]–[26]).

One of the reasons why the computational estimation of distance has not been studied is that sound source distance affects several different properties of the received sound signal. Many of these cues can be present in the signal for multiple reasons, the source-to-receiver distance being only one of them. It is therefore hard to measure distance effects from the received signal quantitatively. Humans most likely combine distance cues in a complex manner to get a sensation of a stable distance [12]. The mechanisms involved in this feature fusion are unclear at the moment.

The problem of sound source distance estimation is closely related to the estimation of the energy ratio of direct and reverberant sounds, i.e., the direct-to-reverberant ratio, at the receiver. Changes in the direct-to-reverberant ratio occur for two reasons. One of them is that the properties of the room change, i.e., the reverberation time (T_{60}) or the room volume

Manuscript received April 24, 2008; revised April 10, 2009. First published May 02, 2009; current version published August 14, 2009. This work was supported by the Hecse Graduate School, Nokia Research Center, Nokia Foundation, Tekniikan edistämissäätiö, and the Academy of Finland under Grant 119029. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nakatani Tomohiro.

The author is with the Department of Media Technology, Helsinki University of Technology (TKK), FI-02015 TKK, Espoo, Finland (e-mail: sampo.vesa@tkk.fi).

Digital Object Identifier 10.1109/TASL.2009.2022001

changes. The other reason is that the source-to-receiver configuration changes, i.e., the source-to-receiver distance or the directivity (or orientation) of the sound source changes. Larsen *et al.* [27] present a method for estimating the direct-to-reverberant ratio based on partial knowledge of a monaural room impulse response, i.e., timing and relative amplitudes of the direct sound and the few first reflections are known. They assume that this information is first obtained with other preprocessing methods. A relationship between the direct-to-reverberant ratio and distance is given as

$$r = r_{\mathrm{c}} \cdot 2^{-(D/R)/6} \tag{1}$$

where r is the source distance, r_{c} is the critical distance (or reverberation radius—the distance at which the direct and reverberant energies are equal), and D/R is the direct-to-reverberant ratio. It is assumed that D/R decreases 6 dB per doubling of distance. Knowledge of the direct-to-reverberant ratio and the critical distance is then enough for estimating the distance of a sound source in an enclosed space. In other words, some information on the properties of the room are needed in order to interpret the reverberation effects of the room on the source signal. A learning approach avoids this problem by training the system separately for each room configuration.

Griesinger identified both binaural [28] and monaural [29] cues for sound source distance. The binaural cue is the interaural cross-correlation (IACC) measure, which is basically the strength of correlation between the left and right ear signals. The monaural cue is pitch-coherence, which measures how much the harmonic structure in voiced parts of the speech signal is corrupted by reverberation. However, these cues are not used for actually estimating the distance by computational means.

Crude classification of the distance of hand clap sounds is presented by Lesser and Ellis [30]. In their work, hand claps are classified as near-field or far-field claps based on a few simple metrics. The near-field claps originate close to the microphones, while other claps are regarded as far-field. The features used are based on the energy sequences of the transients: center of mass, slope of decay, and energy compared to background noise. The classification resulted in error rates between 0.4%–5% in the two rooms and two positions that the system was tested in.

Recently, a model for binaural distance estimation for the dynamic case, where the receiver is moving, has been proposed by Lu *et al.* [31]. Their model consists of the estimation of two dynamic distance cues, the motion parallax and acoustic tau, which are estimated employing a sequential approach (particle filtering). The model is tested with simulated signals, with which the distance estimation error is reported to be in the range of 1.9–3.4 m when using the particle filter with auxiliary importance resampling in different environments. The results are best in the anechoic case (error of 1.9 m). Adding the direct-to-reverberant ratio as a distance cue was found to improve the results for simulated speech sources [32].

Smaragdis and Boufounos [4] developed an expectation maximization algorithm for learning the amplitude and phase differences of cross-spectra in order to recognize the position of a sound source. In [4], it is stated that the approach can not discriminate well between positions that have the same azimuth angle, but different distances. In the present work, this is indeed found to be the case and a simple modification is proposed, which allows correct recognition of positions that have the same azimuth angle.

III. Method

A practical method for learning the distance of a speech sound source, based on a binaural signal, is presented in this section. A previously published method for sound source position learning is also reviewed.

A. Distance Features

In order to estimate the distance of a sound source, a set of suitable features has to be chosen. While several features that correlate with the perceived distance of a sound source have been identified, for computational models one should choose features that are easily computable and dependent only on sound source distance. Humans most likely combine different features and *a priori* information on the surrounding space and on the sound source in a complex manner, the mechanisms of which are not currently known. As a result, a stable sense of sound source distance is obtained.

The frequency dependent magnitude squared coherence (MSC) between left and right ear signals was chosen as the distance feature in this work. Its use is motivated by the fact that when the source-to-receiver distance increases in a room, the ratio of the direct and reverberant sound energies, i.e., the direct-to-reverberant ratio, decreases. This can be seen as a decrease in the strength of correlation between channels of the received binaural signal.

From the short-time spectra $X_{\mathrm{l}}(f,t)$ and $X_{\mathrm{r}}(f,t)$ of the left and right channels, respectively, the magnitude-squared coherence is calculated using the following set of equations [33], [34]

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

$$\hat{G}_{\mathrm{ll}}(f,t) = \langle |X_{\mathrm{l}}(f,t)|^2 \rangle \tag{3}$$

$$\hat{G}_{\mathrm{rr}}(f,t) = \langle |X_{\mathrm{r}}(f,t)|^2 \rangle \tag{4}$$

$$\hat{G}_{\mathrm{lr}}(f,t) = \langle X_{\mathrm{l}}^*(f,t) X_{\mathrm{r}}(f,t) \rangle \tag{5}$$

where $\hat{G}_{\mathrm{lr}}(f,t)$ is the estimated cross-spectrum between channels, and $G_{\mathrm{ll}}(f,t)$ and $G_{\mathrm{rr}}(f,t)$ are the estimated spectra of the left and right channel signals, respectively. Time-averaging is denoted by $\langle \cdot \rangle$. In practice, the averaging operator is implemented for a generic time series $Q(n)$ using a first-order infinite-impulse response (IIR) with

$$\langle Q(n) \rangle = \beta \cdot \langle Q(n-1) \rangle + (1-\beta) \cdot Q(n) \tag{6}$$

where β is the forgetting factor. It is convenient to define a time constant τ in seconds and calculate the forgetting factor β using

$$\beta = \exp\left(-\frac{N_{\mathrm{h}}}{\tau \cdot f_{\mathrm{s}}}\right) \tag{7}$$

where N_{h} is the hop size of the short-time Fourier transform (STFT) in samples and f_{s} is the sampling rate in Hertz.

The proposed distance estimation method is also compared to the method presented by Smaragdis and Boufounos [4] (the baseline method). They used the logarithmic ratios of the Fourier transforms of the left and right signals as features. After taking the complex logarithm

$$R(f,t) = \log \frac{X_{\mathrm{l}}(f,t)}{X_{\mathrm{r}}(f,t)} \tag{8}$$

the real and imaginary parts contain

$$\Re\{R(f,t)\} = \log \frac{|X_{\mathrm{l}}(f,t)|}{|X_{\mathrm{r}}(f,t)|} \tag{9}$$

$$\Im\{R(f,t)\} = \angle(X_{\mathrm{l}}(f,t) \cdot X_{\mathrm{r}}^{*}(f,t)). \tag{10}$$

The real part has the logarithmic ratio of left and right signal spectrum magnitudes, and the imaginary part has the phase difference between the left and right signal spectra.

B. Training Method

Before the distances of sound sources in a room can be learned, some training data has to be obtained. The training data consists of white noise convolved with binaural room impulse responses (BRIRs) measured at the desired locations in the room. The decision to use white noise as the training signal was based on the results of Backman and Karjalainen [35], who compared different training signals for learning the azimuth angle of a sound source. When compared to pink noise and speech, white noise was found best, resulting in the smallest error of azimuth angle.

For each source-to-receiver distance, a coherence profile is calculated as the mean of the coherence of the training signal over time. This profile is used for recognizing the most likely distance of a speech signal in the testing phase using a Gaussian maximum-likelihood scheme. An actual training algorithm is not needed, because there is only one Gaussian for each frequency at which the coherence is calculated. Since the variance of the coherence of the convolved white noise was found to be very different from the variance of the speech signal in the testing phase, the variance parameter in the Gaussians was set to one. When learning sound source position, Smaragdis and Boufounos [4] also included the variance. The model proposed in the present work is completely specified by the means of the Gaussians, which constitute the coherence profile $\mu(f_k, d)$, where f_k is the frequency corresponding to frequency bin k, and d is the distance.

In the baseline method [4], the features are complex numbers representing the magnitude and phase differences between signals (see Section III-A). The training is based on a wrapped-phase expectation-maximization (EM) algorithm, which takes into account the wrapping of the imaginary part (the phase difference) around the interval $[-\pi, \pi]$ by means of a constrained Gaussian mixture. This results in a complex Gaussian distribution having a complex-valued mean and a variance that is the same for both the real and imaginary parts. The details of the training algorithm are presented in [4].

C. Energy-Weighted Maximum-Likelihood Estimation

Not all parts of the signal carry equal amount of information on the sound source distance. Since the training data is calculated from stationary white noise convolved with the BRIRs, its statistics generally will be different from speech recorded at the same source-receiver configuration. However, it is expected that in some time–frequency regions the coherence of speech will have similar values as the coherence of white noise has. The short-time coherence computed from speech is generally meaningless at most time–frequency bins with low signal energy. Therefore, the signal energy is utilized for giving more emphasis to the information-carrying parts of the signal.

Based on the coherence computed using (2)–(5), the energy-weighted log-likelihood is then calculated as

$$L(t_n, d) = -\sum_{f_k=f_L}^{f_H} \left(w(f_k, t_n) \frac{\left(\hat{\gamma}_{\mathrm{lr}}^2(f_k, t_n) - \mu_{\mathrm{c}}(f_k, d)\right)^2}{2\sigma_{\mathrm{c}}^2(f_k, d)} \right) \tag{11}$$

where n, d and k denote the frame, distance, and frequency indices, respectively, t_n is the time corresponding to n, f_k is the frequency corresponding to k, f_{L} and f_{H} are the minimum and maximum frequencies of evaluation (bins from L to H), respectively, and $w(f_k, t_n)$ is a weighting function. The coherence profile at distance d is the mean $\mu_c(f_k, d)$, and the variance $\sigma_{\mathrm{c}}^2(f_k, d)$ is set to one for reasons discussed in Section III-B.

The energy weighting is calculated as

$$\begin{aligned} &w(f_k, t_n) \\ &\quad = -10 \log_{10} \left(\frac{\langle |X_{\mathrm{l}}(f_k, t_n)|^2 \rangle + \langle |X_{\mathrm{r}}(f_k, t_n)|^2 \rangle}{M} + \epsilon \right) \end{aligned} \tag{12}$$

where $X_{\mathrm{l}}(f_k, t_n)$ and $X_{\mathrm{r}}(f_k, t_n)$ are the short-time spectra of the left and right ear signals, respectively, $M = \max_{f_k, t_n}\{\langle |X_{\mathrm{l}}(f_k, t_n)|^2 \rangle + \langle |X_{\mathrm{r}}(f_k, t_n)|^2 \rangle\}$ is a normalizing constant, and ϵ is a very small constant for avoiding logarithms of zero. Notice the extra minus sign that is necessary to decrease the likelihood (11) at time–frequency elements with low energy. In order to ensure that the weights are positive, the sum of the spectra is normalized in (12). The logarithm of energy was chosen as it outperformed uniform weighting and non-logarithmic energy in the experiments conducted.

In previous work [5], instantaneous STFTs were used to calculate the weights. It was found that when the low-passed spectrum estimates from (3) and (4) are used instead, smoother weights, and therefore smoother likelihood tracks are obtained. The time period from which information is integrated is then the same when calculating the coherences and the weights of the likelihoods.

In the baseline method, the most likely distance d at each time instant t_n corresponding to frame index n is recognized by calculating the likelihood of each model as (adapted from [4])

$$\begin{aligned} L(R(t_n), d) = &\prod_{f_k=f_L}^{f_H} \sum_{m=m_{\min}}^{m_{\max}} \frac{1}{\pi\sigma(f_k, d)^2} \\ &\cdot \exp\left(\frac{-|R(f_k, t_n) - \mu(f_k, d) + m2\pi\sqrt{-1}|^2}{\sigma(f_k, d)^2} \right) \end{aligned} \tag{13}$$

where $R(f_k, t_n)$ is the complex-valued feature vector obtained using (8) at time t_n corresponding to time index n and at frequency f_k, $\mu(f_k, d)$ is the complex valued mean of the model at frequency f_k and distance d, $\sigma(f_k, d)^2$ is the real-valued variance of the model, $m_{\min} = -2$, and $m_{\max} = 2$. In practical calculations, the log-likelihood corresponding to (13) is used.

D. Identifying the Most Likely Distance

After computing the likelihoods of different distances, the most likely distance is recognized. In order to do this, the parts of the signal that are the most reliable for distance estimation have to be identified. The energy-weighting already helps to some extent, but there are parts in the signal where the short-time coherence does not carry information on distance. For example, when the sound source is silent or has energy in a narrow frequency band, the likelihoods calculated with (11) will not reflect the true distance. A method for choosing the informative frames is therefore necessary.

The method for identifying the usable signal frames is based on the likelihood magnitudes. When the likelihood of certain distance is high, there is a good match between the coherence of the test signal and the coherence profile of the distance. It seems, then, that the frames where the likelihoods get the highest values over some time interval should be chosen. Unlike in the previous work where a fixed threshold was used [5], P percent of frames having the highest likelihoods are chosen. In practice, for each frame n the maximum of likelihood $L(t_n, d)$ in (11) is taken over all distances d, yielding $L_{\max}(t_n)$. The frames within P percent of the largest values of $L_{\max}(t_n)$ are then included in the analysis. This actual percentage P of frames used is termed the percentage parameter from now on. In this paper, the time interval at which the likelihood track statistics are considered is the whole input signal. For real-time implementations one should keep track of the statistics of the likelihoods and then choose the highest values.

Finally, when the informative frames are selected, the most likely distance at each frame is determined by comparing the mutual order of the likelihood magnitudes and choosing the distance that most frames are assigned to. In the following sections, the percentage of frames assigned to the correct distance is termed the frame-level result. The percentage of correctly classified distances is referred to as the clip-level result, since the whole speech sample is used for deciding the correct distance.

E. Likelihood Behavior With Changing Distance and Changing Angle

Fig. 1 illustrates likelihood tracks calculated using the compared methods. In the first and third plots of Fig. 1, examples of likelihood tracks resulting from the proposed and baseline methods, respectively, are shown. A speech sound source is located at a distance of 0.5 m directly in front of a dummy head. The likelihood tracks in the plots are calculated for models trained using white noise at the same location with the same dummy head orientation angle, at distances from 0.5 m to 4.0 m with 0.5-m spacing. The likelihood of the model trained at the true distance, 0.5 m, is plotted with a thick solid line, while the other likelihoods are plotted with different plot styles of regular thickness. Similarly, the second and fourth plots of Fig. 1 show the likelihoods for the same sound signal at the same location, but now the models are trained with varying orientation angles (0°, 30°, 60°, 90°, and 180°) at a fixed distance of 0.5 m. The dummy head is turned counterclockwise to these orientation angles when training the system. The true orientation angle is 0° (thick line). Note that the second plot has only four likelihood tracks, because the likelihood for one angle is so low that it is left outside the y-axis range. The fifth plot is the spectrogram of the speech signal.

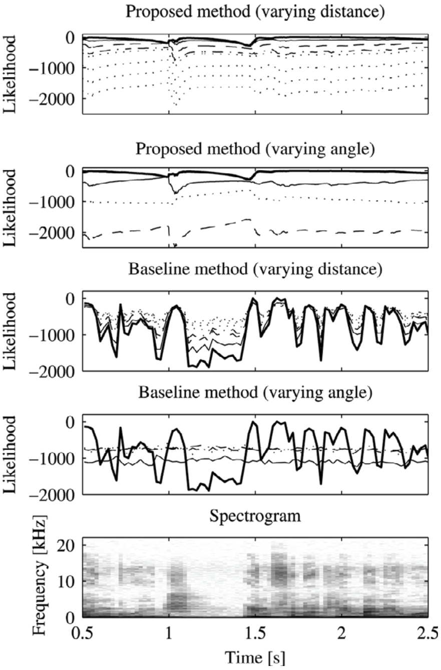

Fig. 1. Example likelihoods calculated from a speech signal with a varying distance or a varying angle. The likelihood corresponding to the true distance/angle (0.5 m/0°) is plotted with a thick solid line, and the likelihoods of other distances/angles are plotted with other plot styles.

When estimating the distance with the proposed method (first plot), it can be seen that the true model has the highest likelihood at the parts of the signal where there is energy according to the spectrogram. However, the marginal between the true model and the other models is very small. When using the baseline method for distance learning (third plot), it can be seen that a similar phenomenon is taking place, but now the likelihood tracks follow the dynamics of the signal more clearly because a time constant is not used when calculating the features. It is worth noting that the proposed method results in smoother likelihood tracks because of the time constant $\tau = 205$ ms applied in calculating the coherence, which is used as the feature. When estimating the angle with the proposed method (second plot) and the baseline method (fourth plot), the true likelihood clearly gets the highest score for most of the time when there is speech present. The silent frames which do not carry information on sound source location can be simply discarded based on their energy, as is seen when comparing the likelihood plots to the spectrogram.

The purpose of this comparison is to show that when learning sound source distance, one cannot simply discard the silent

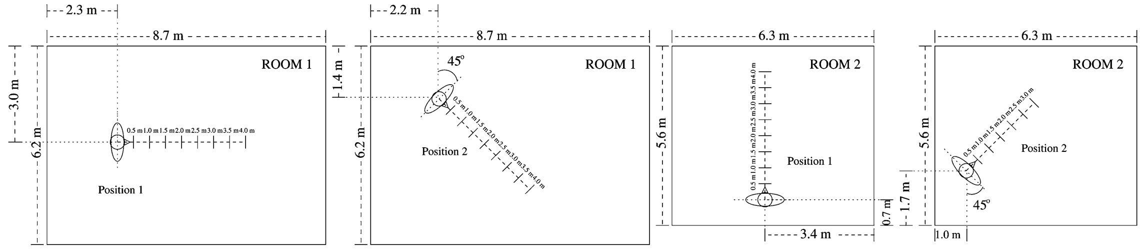
2.3 m
8.7 m
3.0 m
6.2 m
ROOM 1
0.5 m 1.0 m 1.5 m 2.0 m 2.5 m 3.0 m 3.5 m 4.0 m
Position 1
2.2 m
8.7 m
1.4 m
6.2 m
45°
ROOM 1
0.5 m 1.0 m 1.5 m 2.0 m 2.5 m 3.0 m 3.5 m 4.0 m
Position 2
6.3 m
5.6 m
ROOM 2
0.5 m 1.0 m 1.5 m 2.0 m 2.5 m 3.0 m 3.5 m 4.0 m
Position 1
0.7 m
3.4 m
6.3 m
1.7 m
5.6 m
ROOM 2
0.5 m 1.0 m 1.5 m 2.0 m 2.5 m 3.0 m
Position 2
45°
1.0 m

Fig. 2. Placement of the source and the receiver in the two rooms and the two positions.

TABLE I
PROPERTIES OF THE RECORDING ROOMS

Room name	Dimensions	Volume	T_{60} (500 Hz)
Room 1 (meeting room)	870 cm × 620 cm × 360 cm	194 m^3	0.6 s
Room 2 (listening room)	625 cm × 560 cm × 295 cm	103 m^3	0.3 s

frames and get good results, as is the case when the orientation or azimuth angle varies between positions. For example, between approximately 0.75 and 0.9 s, the distance is misclassified by the baseline method, as can be seen in Fig. 1. It is probable that both the proposed features and the features used in the baseline method have smaller variations between positions when the distance changes, compared to the situation where the orientation or azimuth angle changes. This is especially true for the features employed in the baseline method, because those features can be seen as interaural level and phase differences, which most strongly correlate with changing angles of arrival and not with the distance. If there is distance information in these cues, it is due to position-dependent fluctuations of these cues caused by room reflections. This may be the reason why the authors of the baseline method report that in cases where the azimuth angle stays the same between positions, it is difficult to correctly discern between the positions [4]. Later in the present study it is shown that when the usable signal frames are chosen based on the likelihood magnitudes, both methods can recognize the distance correctly.

IV. MATERIALS

In order to test the effectiveness of the proposed distance learning method, audio material was collected from two real rooms.

A. Recordings

A set of test recordings were recorded in two rooms. One of the rooms was a meeting room (room 1) and the other a standardized listening room (room 2). Properties of the two rooms are presented in Table I, while Fig. 2 shows the placement of the recording setup in the rooms. A Cortex MK1 binaural manikin was used for recording the binaural signals. A 5-s-long anechoic sentence in English ("In language, infinitely many words can be written with a small set of letters.") spoken by both a male and a female speaker was emitted by a Genelec 1029A loudspeaker at each location. BRIRs were also measured at each location using the sine sweep excited system identification method [36]. The sampling rate of the recordings was $f_s = 44.1$ kHz.

The manikin (listener) was located at two different positions in each room. Position 1 is close to the center line of the room and the line connecting the source and the listener is parallel to the side walls. In position 2, the listener is close to one of the walls and the line connecting the source and the listener is at 45° angle with respect to the closest wall. The goal was to make the impulse responses at the locations clearly dissimilar between the locations in the same room in order to test the generalization ability of the method.

There are therefore four different cases (two rooms, two locations). In three of these cases the distance of the sound source varied from 0.5 to 4.0 m in 0.5 m increments. An exception was location 2 in room 2, where the distance varied from 0.5 to 3.0 m in 0.5 m increments. This was because there was no space due to furniture present in room 2.

In order to get different test cases and in order to test the angle recognition performance of the compared algorithms, at each distance the orientation of the listener was varied by turning the dummy head counterclockwise 0°, 30°, 60°, 90°, and 180°. The loudspeaker always stayed on the same line facing the dummy head at each position (see Fig. 2) and only the dummy head orientation was changed. The effect of this may not be exactly the same as if the sound source was moved instead (when the azimuth angle of the source changes), but there was not enough space in the rooms to do that, because sufficiently large variations in distance, relative to the room dimensions, were preferred. The changing angle is used for comparing the performance of the proposed and baseline algorithms when the angle of the sound source relative to the listener changes instead of distance. It might also give an idea of how the performance of the method changes with sound source direction.

In the case of room 1 and position 1, an extra set of recordings was made where the location of the sound source was slightly offset so that the projection of the location vector to the original axis was 0.1 m longer than at the original locations, i.e., 0.6 m, 1.6 m, etc. The source was also moved to the left or right so that azimuth angles from 9.5° to 2.6° were formed from the shortest distance to the longest. The purpose of this test was to evaluate the system in conditions with small mismatches in distances and azimuth angles between training and testing.

B. Training Signals

The training signals for each source and receiver configuration were constructed by convolving a white noise signal with the BRIR of a particular location. The training signal was 4-s long, because it was found that there was a negligible difference between a coherence profile calculated from a signal of that length and one calculated from a much longer signal. Because longer time constants are used, the fast variations in the short-time coherence stabilize, and therefore there is no need to use a longer training signal. This also saves computing time when calculating the profiles at different source-receiver configurations.

V. Results

In this section, the choice of parameters is first discussed, followed by a brief investigation into the behavior of the trained models. Finally, the distance and angle estimation performances of both studied methods are presented.

A. Choice of Parameter Values

A few different parameters have to be chosen in order to test the methods. One could systematically try to find an optimal combination of parameters. For both of the tested approaches, it was decided that the window length parameters are kept fixed and an optimal value for the percentage parameter P (between 0% and 100%) and the frequency range of evaluation $f_L - f_H$ was chosen by global optimization using the simulated annealing algorithm [37]. The goal of the optimization was to maximize the frame-level performance (see Section III-D). The optimal value for the time constant τ was also sought for the proposed algorithm. The time constant was constrained between 5 and 700 ms with 5-ms resolution, as the models have to be precalculated in order to save time. Following suggestions for the baseline method [4], the time-domain Blackman window length was set to 1024 samples, there was no overlap between successive windows, and the FFT size was 1024. For the proposed method, the time-domain Hanning window size was set to 128 samples, hop size to 32 samples, and the FFT size to 1024. When using the proposed method it was found that the coherence profiles turn out to be smooth enough when the FFT size is considerably higher than the time-domain window size.

The time constant has an effect on how much fast changes affect the coherences, consequently also affecting the likelihood tracks—a larger time constant results in smoother likelihood tracks, but some critical information may also be lost. A range of time constants from 5 to 700 ms corresponds to integrating information from time periods shorter than the typical time window of 20 ms, in which speech is usually considered to be stationary, to larger time scales where this condition does not hold. The larger the time constant, the less fast variations in the spectrum estimates have an effect on the coherence.

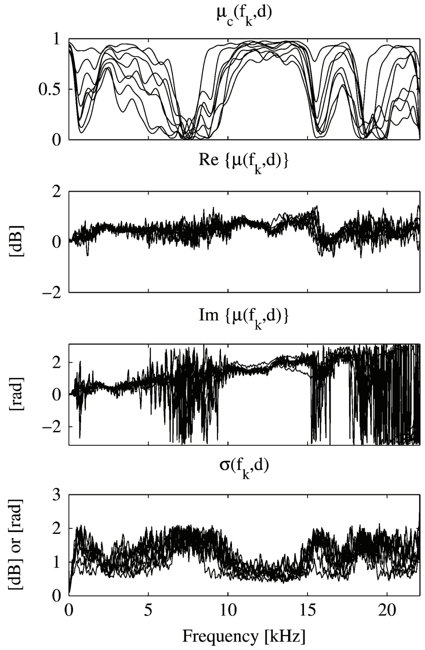

Fig. 3. Examples of models learned from a white noise signal with varying distance (room 1, position 1, angle fixed at 0°). First panel: coherence profile of the proposed method. Second panel: real part of the mean of model used in the baseline method. Third panel: imaginary part of the mean of model used in the baseline method. Fourth panel: variance of the model used in the baseline method.

B. Behavior of the Trained Models

In order to gain insight into the nature of the problem, properties of the models trained using the compared methods are investigated in more detail. Fig. 3 illustrates how the models trained with white noise (see Section IV-B) behave when the source distance changes in room 1, position 1. The top plot shows the coherence profiles of the proposed method. Each solid line plots the mean coherence of white noise played back at one of the eight distances. It is seen that distance clearly affects the coherence. There is a dip located approximately between 5 and 10 kHz. It is probably caused by the head-related transfer function (HRTF) of the manikin. Hartmann *et al.* [38] have investigated binaural coherence in rooms and there is a dip at the same frequency range in their results too. The three lower plots show how the baseline method behaves when distance changes. These three plots show the real and imaginary parts of the means, and their common variances, respectively. All the three plots look as if some distance-dependent variation is superimposed on a basic shape, which behaves more smoothly. This basic shape corresponds to the direction-dependent cues, and the variations are then mostly caused by unique reflection patterns at each distance. It seems that the phase difference cues $\Im\{\mu(f_k, d)\}$ in the baseline method exhibit fast variations as a function of frequency in between 5 and 10 kHz, which is the same area where the coherence dips.

Fig. 4 is a similar plot where the distance stays fixed at 0.5 m while the listener orientation angle changes (0°, 30°, 60°, 90°,

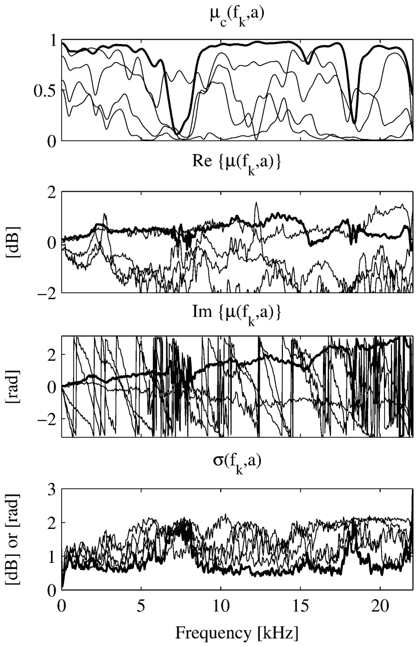

Fig. 4. Same as Fig. 3 but with varying angle (distance fixed at 0.5 m).

and 180°). The 0° case is plotted with a thicker line. It is evident from the plots that the changing orientation angle has a greater effect on the location cues of both methods, compared to the case where the distance changes. The means and the variance are clearly different between the cases, and there is systematic frequency-dependent behavior. This is especially true for the imaginary part of the mean of baseline method model, where one can see wrapping of the phase from π to $-\pi$ at several frequencies. The location features used in the baseline method roughly correspond to interaural level differences (ILD, the real part) and interaural time or phase differences (ITD/IPD, the imaginary part). Since these cues are most strongly affected when the orientation angle of the listener (or the azimuth angle of a sound source) changes, it is consistent that clear changes in the features are observed.

C. Performance With Optimal Parameter Values

Table II shows the mean distance recognition performances of the proposed [Table II(a)] and baseline [Table II(b)] methods, and the corresponding optimal parameter values derived using the optimization procedure described in Section V-A. Fig. 5 illustrates how the performances change when the optimal parameters of Table II are used but the percentage parameter P is varied. The mean distance recognition performance is calculated by taking the mean of the frame-level or clip-level results (Section III-D) over 40 test cases (two rooms, two speech samples, and two positions) for each of the five angles (0°, 30°, 60°, 90°, and 180°). For the proposed method, the parameters are optimized also when excluding angle 180° at position 2 in both rooms, because those cases were found to be problematic due to a combination of strong wall reflections and diffraction around the earlobes causing abrupt behavior in the coherences. The performance is also optimized when the roles of distance and orientation angle are switched, i.e., the listener orientation angle is recognized while the distance stays fixed (see Section III-E for examples of likelihood behavior when either the distance or the angle changes). This gives perspective into the difficulties inherent in recognizing the distance. Fig. 6 is a plot similar to that of Fig. 5 for the case where the angle varies instead of distance.

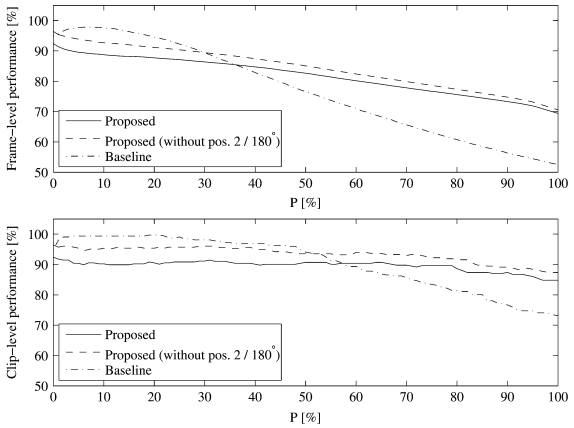

Fig. 5. Effect of the percentage parameter P in the proposed algorithm and the baseline algorithm when learning the distance. All other parameters are set to values in Table II while P is varied. Top panel: the frame-level result. Bottom panel: clip-level result.

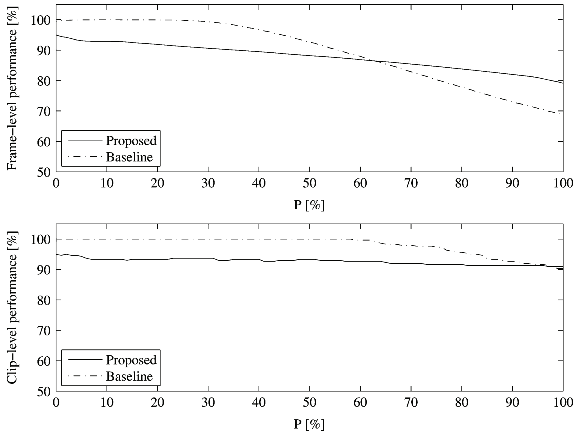

Fig. 6. Same as Fig. 5 but when learning the orientation angle.

For the proposed method, the optimal values [Table II(a)] show some interesting phenomena. The best distance recognition performance is reached when the percentage parameter P is zero. This means that only one single frame where the likelihood is at its maximum is analyzed. When P is increased, the frame-level performance drops almost linearly (top panel of Fig. 5). In practice, using just a single frame is not very robust, so more frames should actually be included in deciding the distance, even if the frame-level results would be worse. The

TABLE II
OPTIMAL PARAMETER VALUES AND PERFORMANCE. (a) PROPOSED METHOD (COHERENCE PROFILE). (b) BASELINE METHOD (WRAPPED-PHASE COMPLEX GAUSSIAN MODEL OF CROSS-SPECTRUM MAGNITUDE AND PHASE)

Case	P (%)	f_L (Hz)	f_H (Hz)	τ (ms)	Frame-level perf. (%)	Clip-level perf. (%)
distance (all)	0	215	13566	205	92.4	92.4
distance (w/o pos. 2 / 180°)	0	215	13351	185	96.4	96.4
angle	0	0	15461	300	95.0	95.0

(a)

Case	P (%)	f_L (Hz)	f_H (Hz)	Frame-level perf. (%)	Clip-level perf. (%)
distance	7.5	215	17959	97.9	99.4
angle	11.5	2326	12575	100	100

(b)

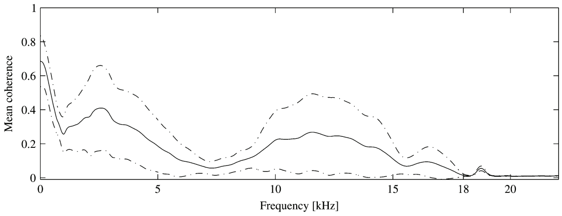

Fig. 7. Statistics of the coherence of the 40 speech samples used for testing ($\tau = 180$ ms). The mean is plotted with solid line, while the dash-dotted lines plot the standard deviation range.

bottom panel of Fig. 5 shows that when including the entire data set (solid line), the percentage of correctly recognized distances (clip-level performance) stays above 85% as P increases, while the frame-level result drops to 70% for $P = 100\%$. The optimal values for the lower and upper frequency limits of evaluation are also interesting. The frequency range is approximately 200–13 000 Hz, whether the problematic cases are excluded or not. Fig. 7 plots the mean coherence calculated over all test speech samples. One can see that the coherence drops to a small value near 18 kHz and varies only very slightly after that. The upper limit f_{H} is about 5 kHz below this frequency limit, above which the coherence cue becomes unreliable with the test material used. The time constants are close to 200 ms, which indicates that information from a time span 10 times as long as the usual stationary period of speech (20 ms) is used for recognizing the distance. Removal of the problematic cases increases the performance somewhat (4%), as was expected.

With the baseline algorithm [Table II(b)] the situation is different, as the best results are obtained with a larger percentage value of $P = 7.5\%$. The performances are also higher. Increasing P causes a drop in the performance (Fig. 5). The drop is steeper than with the proposed method. This is most likely due to time constant used in calculating the coherence feature in the proposed method, which causes the likelihoods to respond more slowly to changes in the input signal. The likelihood track of the correct distance will then tend to be the highest one also during short gaps in the speech signal. The frequency range is close to the proposed method, even though the upper frequency is higher.

It is interesting to compare the distance recognition performance to the angle recognition performance. The optimal parameters for the case where the roles of distance and orientation angle were switched so that the angle was recognized, is shown as the last entries of Tables II(a) and (b), for the proposed and baseline methods, respectively. The performance is 95% and 100% for the proposed and baseline methods, respectively. Fig. 6 shows how the percentage parameter P affects angle recognition performance. The percentage is not as critical now with the baseline method, since even when all frames are used in the analysis ($P = 100\%$), the correct angle is identified in 90% of the cases. When learning the distance, only 73% of distances were correctly recognized at $P = 100\%$. With the proposed method at $P = 100\%$, there is only 6% difference to distance learning as 91% of the cases are correctly recognized when learning the angle. The higher time constant of $\tau = 300$ ms probably reflects the larger differences between the models when the angle changes instead of distance. It seems that it is acceptable for the features to respond more slowly to changes in the signal when learning the angle, compared to learning the distance. It was manually investigated that the average percentage of speech in the anechoic test clips (male and female speech) is approximately 72%. Interestingly, the clip-level performance of the baseline method (bottom panel of Fig. 6) starts to drop just after $P = 60\%$. This indicates that when the angle changes, one could simply discard the silent frames from the analysis, as was also discussed in Section III-E.

D. Generalization Ability

To test the limits of both tested approaches, the effect of the training data being taken from a position other than the testing data is investigated. This is done by swapping the training data between positions 1 and 2 in both rooms, i.e., in position 1, training data from position 2 is used and vice versa. When using the parameters listed in Table II, the frame-level/clip-level distance recognition results then drop to 28.8%/27.0% and 31.0%/32.5% for the proposed and baseline methods, respectively. This is to be expected, as this approach

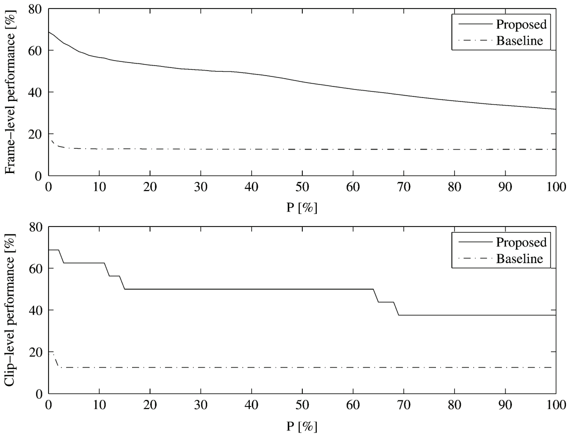

Fig. 8. Same as Fig. 5 when the source location is slightly perturbed (only in room 1, position 1, angle 0°).

to distance learning is dependent on the particular ways that the room affects the localization cues. When the source and receiver move to a different position inside the room, the effects change. This would happen even more severely when moving to a different room.

Fig. 8 is similar to Fig. 5, but with the difference that the test data was recorded with the sound source at slightly offset locations in room 1, position 1, and angle 0° (see Section IV-A). The mean offset of the eight positions was 18 cm, with a standard deviation of 2.4 cm. The parameters used are those listed in first rows of Tables II(a) and (b), for the proposed and baseline methods, respectively. Now the best frame-level and clip-level performance for both methods is obtained with $P = 0\%$, i.e., only the highest likelihood value is used for deciding the correct distance. The proposed and baseline methods reach 68.8% and 18.8% correctly recognized distances, respectively. The 50% higher performance of the proposed method is surprising, as in [4] the authors of the baseline method used offsets up to 20 cm in synthetic experiments (where a room model was used) and found that the algorithm is able to correctly recognize the distance, even when there is a slight mismatch between testing and training. Regarding the real-world experiments it was just stated that the test recordings were made in approximately the same positions [4], from which could be concluded that the behavior of the recognition was not tested with an offset as was done in this work. It is possible that when both the distance and the azimuth angle change when moving from the training position to the offset position, as in the present experiment, the distance cues of the baseline method behave in ways that the models are not able to generalize to. The coherence features probably behave more consistently as the distance changes and therefore can generalize better in this case.

VI. Discussion

Based on the results, both of the investigated methods for binaural sound source distance learning can successfully learn the distance of a speech sound source when the orientation angle of the listener is known. Because the location features employed in the methods are position-dependent, the models have to be trained for each listener location separately. This is practically impossible for a moving listener and, therefore, the methods are only practical for a fixed microphone array.

Optimal values for three crucial parameters, the time constant of coherence calculation τ, the percentage of frames used for analysis P, and the frequency range $f_L - f_H$ were sought. It was found that for optimal performance in the proposed method, the time constant generally should be of the order of hundreds of milliseconds as opposed to the previous work [5], where a very short time constant of $\tau = 3.2$ ms was used. A longer time constant is feasible, since fast, signal-dependent fluctuations have less effect then. It is also shown that frequencies up to as high as 13 or 18 kHz for the proposed and baseline methods, respectively, are useful for distance recognition.

It was also shown that binaural distance learning is a special case of binaural sound source location learning (the baseline method previous presented in [4]), where the room-dependent variations to interaural cues are utilized to discern between sound sources at different distances. Because the variations are subtle, only the frames that have the highest likelihoods can be used when deciding the most likely distance. When the direction of the sound source relative to the listener changes, there are larger variations in the interaural cues, permitting easier recognition. This is especially true with the baseline method. With the proposed method, the time constant used in calculating the features lessens this effect.

Compared to human performance, the evaluated methods are more accurate. Humans tend to underestimate large sound source distances and overestimate distances shorter than 1 m [13]. The same phenomenon is not observed in the learning methods presented here. However, the generalization abilities of the algorithms are poorer than with humans, since the classification systems have to be trained for each room and listener position separately.

VII. Conclusion

A binaural method for learning the distance of a speech source is presented and evaluated. The proposed method is compared to a previous method [4]. It is shown that both methods can accurately recognize the distance of a short speech sample in most cases, even though the proposed method has difficulties when the sound source is located directly behind the listener when the listener is located close to a wall. Future work includes developing a real-time version of the algorithm and an extension to dynamic situations, where the sound source and/or the listener may be moving. New distance cues are probably needed for these extensions. The use of different learning methods should also be investigated.

Acknowledgment

The author wishes to thank Dr. T. Lokki, Dr. K. Palomäki, Mr. S. Tervo, and Prof. M. Karjalainen for helpful discussions and comments, the anonymous reviewers for insightful comments, and Dr. V. Pulkki and Mr. M. Tikander for help with the recordings.

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," *J. Audio Eng. Soc.*, vol. 52, no. 6, pp. 618–639, 2004.

[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 J. Appl. Signal Process.*, vol. 2005, no. 18, pp. 2915–2929, Oct. 2005.

[3] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, Oct. 2005, pp. 158–161.

[4] P. Smaragdis and P. Boufounos, "Position and trajectory learning for microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 358–368, Jan. 2007.

[5] S. Vesa, "Sound source distance learning based on binaural signals," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, Oct. 2007, pp. 271–274.

[6] S. Nielsen, "Distance perception in hearing," Ph.D. dissertation, Aalborg Univ., Aalborg, Denmark, May 1991.

[7] B. G. Shinn-Cunningham, "Localizing sound in rooms," in *Proc. ACM SIGGRAPH and EUROGRAPHICS Campfire: Acoust. Rendering for Virtual Environ.*, Snowbird, UT, May 2001.

[8] D. S. Brungart and K. R. Scott, "The effects of production and presentation level on the auditory distance perception of speech," *J. Acoust. Soc. Amer.*, vol. 110, no. 1, pp. 425–440, 2001.

[9] M. Naguib and R. H. Wiley, "Estimating the distance to a source of sound: Mechanisms and adaptations for long-range communication," *Animal Behaviour*, vol. 62, no. 5, pp. 825–837, 2001.

[10] L. Ottaviani, F. Fontana, and D. Rocchesso, "Recognition of distance cues from a virtual spatialization model," in *Proc. Int. Conf. Digital Audio Effects*, Hamburg, Germany, Sep. 2002.

[11] P. Zahorik, "Auditory display of sound source distance," in *Proc. Int. Conf. Auditory Display*, Kyoto, Japan, Jul. 2002.

[12] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *J. Acoust. Soc. Amer.*, vol. Ill, no. 4, pp. 1832–1846, 2002.

[13] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acust. United With Acust.*, vol. 91, no. 3, pp. 409–420, May/Jun. 2005.

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

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

[16] B. G. Shinn-Cunningham, "Distance cues for virtual auditory space," in *Proc. IEEE Pacific-Rim Conf. Multimedia*, Sydney, Australia, Dec. 2000, pp. 227–230.

[17] D.-Y. Jang, H.-G. Moon, K. Kang, and H.-K. Jung, "Room impulse response shaping for enhancement of perceived spaciousness and auditory distance," in *Proc. Int. Conf. Digital Audio Effects*, Naples, Italy, Oct. 2004, pp. 315–318.

[18] L. A. Jeffress, "A place theory of sound localization," *J. Comp. Physiol. Psychol.*, vol. 41, pp. 35–39, 1948.

[19] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals," *J. Acoust. Soc. Amer.*, vol. 80, pp. 1608–1622, 1986.

[20] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front," *J. Acoust. Soc. Amer.*, vol. 80, pp. 1623–1630, 1986.

[21] R. K. Clifton, "Breakdown of echo suppression in the precedence effect," *J. Acoust. Soc. Amer.*, vol. 82, no. 5, pp. 1834–1835, 1987.

[22] W. Gaik, "Combined evaluation of interaural time and intensity differences: Psychoacoustical results and computer modeling," *J. Acoust. Soc. Amer.*, vol. 94, no. 1, pp. 98–110, 1993.

[23] C. Liu, B. C. Wheeler, W. D. O'Brien, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," *J. Acoust. Soc. Amer.*, vol. 108, no. 4, pp. 1888–1905, 2000.

[24] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Hong Kong, May 2003, vol. 5, pp. 149–152.

[25] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, 2003.

[26] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Amer.*, vol. 116, no. 5, pp. 3075–3089, 2004.

[27] E. Larsen, C. D. Schmitz, C. R. Lansing, W. D. O'Brien, B. C. Wheeler, and A. S. Feng, "Acoustic scene analysis using estimated impulse responses," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, Nov. 2003, vol. 1, pp. 725–729.

[28] D. Griesinger, "Measurement of acoustic properties through syllabic analysis of binaural speech," in *Proc. Int. Congr. Acoust.*, Kyoto, Japan, Apr. 2004, vol. 1, pp. 29–32.

[29] D. Griesinger, "Pitch coherence as a measure of apparent distance in performance spaces and muddiness in sound recordings," in *Proc. Audio. Eng. Soc. Conv.*, San Francisco, CA, Oct. 2006, preprint 6917.

[30] N. Lesser and D. Ellis, "Clap detection and discrimination for rhythm therapy," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, Mar. 2005, vol. 3, pp. 37–40.

[31] Y.-C. Lu, M. Cooke, and H. Christensen, "Active binaural distance estimation for dynamic sources," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 574–577.

[32] Y.-C. Lu and M. Cooke, "Auditory distance perception based on direct-to-reverberant energy ratio," in *Proc. Int. Workshop Acoust. Echo Noise Contr.*, Seattle, WA, Sep. 2008.

[33] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Amer.*, vol. 62, no. 4, pp. 912–915, Oct. 1977.

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

[35] J. Backman and M. Karjalainen, "Modelling of human directional and spatial hearing using neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Minneapolis, MN, Apr. 1993, vol. 1, pp. 125–128.

[36] S. Müller and P. Massarani, "Transfer-function measurement with sweeps," *J. Audio Eng. Soc.*, vol. 49, no. 6, pp. 443–471, Jun. 2001.

[37] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[38] W. M. Hartmann, B. Rakerd, and A. Roller, "Binaural coherence in rooms," *Acta Acust. United With Acust.*, vol. 91, no. 3, pp. 451–462, 2005.

Sampo Vesa was born in Helsinki, Finland, in 1979. He received the M.Sc. degree in acoustics and audio signal processing from the Helsinki University of Technology (TKK), Espoo, Finland, in October 2004. He is currently pursuing the Ph.D. degree at TKK.

His research interests include audio signal processing, particularly the analysis of binaural signals and its practical applications.