Speaker Distance Estimation using Binaural Hearing Aids and Deep Neural Networks*

Mehdi Zohourian, Jakob Stinner, and Rainer Martin

Ruhr-Universität Bochum, Institute of Communication Acoustics, Bochum, Germany

Abstract

In this paper, we address the problem of speaker distance estimation using binaural hearing aid microphones. The proposed approach is based on deep neural networks which aim to classify discrete distances at different levels of precision. We investigate two types of networks, namely, feedforward and convolutional neural networks which are trained on the binaurally estimated direct-to-reverberant energy ratio. The performance of the proposed algorithm is assessed on speech signals convolved with both real and synthetic binaural room impulse responses for several distances and in different acoustical scenarios. On the one hand the proposed approach offers a reliable classification of coarse distances classes which is also robust against variations of the reverberation time. On the other hand, when a more accurate distance estimation is required, the proposed approach is robust only against small variations of the reverberation time.

Keywords: speaker distance estimation, hearing aid, deep neural networks

1 INTRODUCTION

Acoustic source localization refers to the task of identifying the position of different sources, most commonly in terms of direction and distance. While the estimation of the *direction-of-arrival* (DOA) has been widely studied in the last decades, estimation of sound source distance has received less attention. The knowledge of the sound source distance is desirable for a wide range of applications, e.g. assistive devices for visually-impaired people, human-machine interfaces, and *hearing aids* (HA).

A well-known acoustic parameter which is widely employed to indicate the source distance is the *direct-to-reverberant energy ratio* (DRR). The DRR is a function of room characteristics, distance to a sound source, and the directivities of both sound source and microphones. In a reverberant room, the DRR is proportional to the inverse of the source distance, i.e., every doubling of distance decreases the DRR by 6 dB [1]. In other words, if the sound source distance increases the energy of the direct sound decreases whereas the energy of the reverberant sound remains unchanged. The DRR can be computed directly from the measured *room impulse response* (RIR) which practically requires an intrusive measurement within the room. Alternatively, the DRR is estimated blindly from the incoming signals assuming that the reverberation is modeled as a diffuse sound field. Moreover, in multi-microphone configurations it is assumed that the energy of the direct component is fully represented by the energy of the coherent source.

State-of-the-art distance estimation algorithms can be classified into learning-based approaches and approaches based on prior knowledge of the room properties [2]. Both groups implicitly require a calibration w.r.t. one or multiple points within a room. Learning-based approaches utilize different features for the training phase. The authors in [3] use the *magnitude-squared coherence* (MSC) of binaural signals to train a *Gaussian mixture model* (GMM). The estimated DRR is a common feature used for instance in [4, 5]. Georganti et al. [6] propose to use the spectral power of binaural signals with which two variants of learning algorithms are trained, namely, a *support vector machine* (SVM) and a GMM. *Head-related transfer functions* (HRTFs) are another feature proposed in [7] for source distance estimation of binaural signals in three dimensional space. In almost all these evaluations the reverberation time (T_{60}) of the training and test sets are similar as otherwise large

*This work has received funding from the European Fund for Regional Development, grant no. EFRE-0800372 (RaVis-3D).

errors are to be expected.

In this paper we aim for the estimation of speaker distance using 2x2 binaural HA microphones. Our approach is based on the use of deep feedforward and convolutional neural networks trained on the binaurally estimated DRR. Our algorithm is able to classify discrete distances in simulated and real acoustic environments.

The remainder of this paper is organized as follows. Section 2 describes the estimation of binaural DRR using a coherence based technique. Section 3 explains the topology of the two neural networks used in this paper. The experimental setup and validation results are presented in Section 4. Section 5 concludes this paper.

2 Direct-to-reverberant energy ratio

In our scenario we consider a noiseless acoustic environment with a point source signal s received by 2x2 microphones distributed on a pair of HAs. Using the convolution operator $*$, the microphone signal x may be expressed as

$$x_m(n) = h_m(n,\theta_s) * s(n), \quad m \in \{1,\ldots,4\}, \tag{1}$$

where n denotes the sampling index and θ_s determines the DOA of the source. In this equation, h_m indicates the binaural RIR (BRIR) for the corresponding source/receiver positions which is composed of $h_m(n,\theta_s) = h_m^{(\mathrm{d})}(n,\theta_s) + h_m^{(\mathrm{er})}(n) + h_m^{(\mathrm{lr})}(n)$, where $h_m^{(\mathrm{d})}(n,\theta_s)$, $h_m^{(\mathrm{er})}(n)$, and $h_m^{(\mathrm{lr})}(n)$ denote the direct path, the early reflections, and the late reverberation, respectively. Then, the DRR for each microphone can be directly computed from the RIR as

$$\eta_{\mathrm{true}} = \frac{\sum_{n=1}^{T_m+N} |h_m(n,\theta_s)|^2}{\sum_{n=T_m+N+1} |h_m(n,\theta_s)|^2}, \tag{2}$$

where T_m indicates the index of the first peak in the BRIR for each microphone. Moreover, N denotes the range of samples in the neighborhood of the direct component and early reflections used to compute the energy of the direct path. Practically, we consider $N/f_S = 2$ ms in the neighborhood of the highest peak in the RIR for the computation of the energy of the direct path.

In practice, however, the RIR is not available and thus the DRR is blindly estimated from the multi-microphone signal. For this purpose, the effect of early reflections is commonly not considered and thus the signal model in (1) may be written as

$$y_m(n) = x_m(n,\theta_s) + v_m(n), \tag{3}$$

where $x(n,\theta_s) = h_m^{(\mathrm{d})}(n,\theta_s) * s(n)$ indicates the non-reverberated microphone signal and $v_m(n) = h_m^{(\mathrm{lr})}(n) * s(n)$ denotes the reverberated part of the microphone signal which is modeled as an ideal diffuse sound field.

Applying the *short-time Fourier transform* (STFT) to the signals leads to

$$Y_m(k,\mu) = X_m(k,\mu,\theta_s) + V_m(k,\mu), \tag{4}$$

where k and μ indicate frequency and time frame indices, respectively.

One solution to estimate DRR is to compute the coherence $\Gamma_{Y_m Y_{m'}}$ between two microphone signals defined as

$$\Gamma_{Y_m Y_{m'}}(k) = \frac{\Phi_{Y_m Y_{m'}}(k)}{\sqrt{\Phi_{Y_m Y_m}(k)\Phi_{Y_{m'} Y_{m'}}(k)}}, \tag{5}$$

where $\Phi_{Y_m Y_{m'}}(k) = E\left\{Y_m(k)Y_{m'}^*(k)\right\}$ determines the cross *power spectrum* of the microphone signals which is estimated by using the first-order recursive temporal smoothing filter as

$$\Phi_{Y_m Y_{m'}}(k,\mu) = \alpha\Phi_{Y_m Y_{m'}}(k,\mu-1) + (1-\alpha)Y_m(k,\mu)Y_{m'}^*(k,\mu), \tag{6}$$

where $\alpha = 0.92$ denotes the smoothing factor.

Based on the signal model (3) and the assumption that the non-reverberated and reverberated components of the microphone signals are mutually uncorrelated, we may express

$$\Phi_{Y_mY_{m'}}(k,\mu) = \Phi_{X_mX_{m'}}(k,\mu,\theta_s) + \Phi_{VV}(k,\mu)\Gamma_{V_mV_{m'}}(k). \tag{7}$$

In this equation, Φ_{VV} indicates the power of the reverberated components of the microphone signals and $\Gamma_{V_mV_{m'}}$ denotes its coherence which is estimated assuming the spherically isotropic (diffuse) sound field as

$$\Gamma_{V_mV_{m'}}(k) = \frac{\sin(2\Omega_k f_s a c^{-1})}{2\Omega_k f_s a c^{-1}}, \tag{8}$$

with $a = 8.5$ cm indicating the radius of the head and $c = 343$ m/s denoting the speed of sound. Here, $\Omega_k = 2\pi k/M$ where M is the number of *discrete Fourier transform* (DFT) bins.

Moreover, taking the microphone m as a reference we may write

$$\Phi_{X_mX_{m'}}(k,\mu,\theta_s) = \frac{D_{m'}(k,\theta_s)}{D_m(k,\theta_s)}\Phi_{XX}(k,\mu), \tag{9}$$

where Φ_{XX} denotes the PSD of the non-reverberated component of microphone m (the reference microphone) and $D_m(k,\theta_s)$ indicates the Fourier transform of $h_m^{(\mathrm{d})}(n,\theta_s)$. In this work, we only consider the coherence between binaural microphones and thus $m \in \{1,3\}$ and $m' \in \{2,4\}$. Therefore, we may write

$$\Phi_{X_mX_{m'}}(k,\mu,\theta_s) = G(k,\theta_s)\exp\left(j\Omega_k f_s \Delta\tau(\theta_s)\right)\Phi_{XX}(k,\mu) \tag{10}$$

where $G(k,\theta_s) = \left|\frac{D_{m'}(k,\theta_s)}{D_m(k,\theta_s)}\right|$ and $\Delta\tau(\theta_s) = \tau_{m'}(\theta_s) - \tau_m(\theta_s)$ are the *interaural level difference* (ILD) and *interaural time difference* (ITD) for the source angle θ_s, respectively.

By inserting (7) in (5) and introducing $\eta(k,\mu) = \frac{\Phi_{XX}(k,\mu)}{\Phi_{VV}(k,\mu)}$ as the narrowband DRR of microphone m we may achieve

$$\Gamma_{Y_mY_{m'}}(k,\mu) = \frac{G(k,\theta_s)e^{j\Omega_k f_s \Delta\tau(\theta_s)}\eta(k,\mu) + \Gamma_{V_mV_{m'}}(k)}{\sqrt{(\eta(k,\mu)+1)(G^2(k,\theta_s)\eta(k,\mu)+1)}}. \tag{11}$$

In principle, $\Gamma_{Y_mY_{m'}}$ is a complex value whereas η is a positive real-valued quantity. We thus need a meaningful mapping of the complex coherence to the real-valued DRR. In our previous work [8] we propose to take the MSC to estimate the distance of speech signals coming from different azimuths. In this work we employ the MSC approach to estimate the speaker's distance at $\theta_s = 0$. Therefore, $G(k,0) = 1$, $\Delta\tau(0) = 0$, and (11) is simplified to

$$\Gamma_{Y_mY_{m'}}(k,\mu) = \frac{\eta(k,\mu) + \Gamma_{V_mV_{m'}}(k)}{\eta(k,\mu)+1}. \tag{12}$$

Taking the magnitude square of (12) and rearranging the equation leads to the following MSC based DRR estimation approach

$$\hat{\eta}(k,\mu) = \frac{\Gamma_{V_mV_{m'}}(k) - |\Gamma_{Y_mY_{m'}}(k,\mu)|}{|\Gamma_{Y_mY_{m'}}(k,\mu)| - 1}. \tag{13}$$

Furthermore, inspired by the human auditory system, we apply 40 mel-scaled triangular bandpass filters to the estimated narrowband DRR and average across time frames as

$$\hat{\eta}_{\mathrm{mel}}(c) = \frac{1}{B}\sum_{\mu=0}^{B-1}\sum_{c=0}^{C-1} L(c,\mu)\hat{\eta}(k,\mu), \tag{14}$$

where $L(c,\mu)$ indicates the filter response of auditory channel c.

3 Distance estimation using classification techniques

In order to estimate the distance of a sound source we propose to use a classification technique using two neural network architectures, namely, feedforward and convolutional neural network that aim to classify discrete distances. The networks are trained on estimated DRR and will be evaluated on several acoustic scenarios.

3.1 Feedforward neural network

The first classifier is based on a *feedforward neural network* (FNN) which is also known as a *multilayer perceptron* (MLP). The structure of the network consists of an input layer, an output layer, and multiple fully connected hidden layers. The network weights are computed by minimizing a loss function using the gradient descent approach. In this work we design a FNN comprising 2 hidden layers with 44 hidden neurons, each of which compute a non-linear activation function of the weighted sum of the output of the previous layer. The final layer uses the *softmax* rule to estimate the posterior probability for each distance class using the *cross-entropy* loss function. The network is trained in full batch mode with *scaled conjugate gradient backpropagation* (SCG) [9] and early stopping.

3.2 Convolutional neural network

The second classifier is based on a *convolutional neural network* (CNN) which typically consists of convolutional and pooling layers followed by a fully connected layer. The convolutional layer applies the convolution operation to the input and each neuron receives input from only a restricted subarea (receptive field) of the previous layer. The pooling operation reduces the size of the input data and the fully connected layer aggregates local information to provide class discrimination. This topology is widely used in practice since it significantly reduces the computational complexity for large sizes of input data.

In this work, we propose a six-layer CNN including five convolutional and one fully-connected layers. The size of kernels (filters) in the convolutional layers is set to 3×3. The number of filters (depth) in the convolutional layers is set to 16 in the first layer and increases by a factor of 2 after each pooling up to 128 in the final layer. The convolutional stride is fixed to 1 and a padding of 1 is designed to maintain the resolution after convolution. We use max-pooling with a 3×3 kernel, stride of 2 to bisect the input dimensions. Furthermore we use *rectified linear unit* (ReLU) activation functions and batch normalization. Global pooling over time followed by a dropout of 50% prior to the final fully connected layer enforces approximate time-translation invariance. The final layer uses the *softmax* rule to estimate the posterior probability for each distance class using the *cross-entropy* loss function. Moreover, for optimizing the learning process we employ the Adam algorithm [10].

4 Experimental results

In our experiment we use speech signals convolved with synthetic and real BRIRs at eight discrete distances $d_{\text{true}} \in [0.5 : 0.5 : 4]$ m.

In order to generate synthetic BRIRs we integrate the *head-related impulse responses* (HRIRs) in [11] into the *auditory virtual environment* (AVE) software [12]. The HRIRs [11] are recorded for 6 microphones of a pair of *behind-the-ear* (BTE) HAs from which we only use 4 channels corresponding to the front and rear HA microphones. The AVE software uses a modified image source model [13] and an improved statistical reverberation model [14] to synthesize acoustics of multiple sound sources in different rooms. The algorithm also provides a smooth transition between the early reflections and late reverberation. With this software we simulate a room with dimensions of 4 m × 5.5 m × 3 m and three reverberation times of $T_{60} = 0.4, 0.5$, and 0.9 s.

We also measure BRIRs using 2x2 microphones of a pair of BTE HAs attached to a *Head and Torso Simulator* (HATS) [15]. BRIRs are recorded in a reverberant room with dimensions of 7.5 m × 6.3 m × 3.3 m, three reverberation times of $T_{60} = 0.4, 0.7$, and 0.9 s and for the various distances mentioned earlier.

We convolve the BRIRs with speech signals taken from the TIMIT database [16]. For the training set we

Table 1. Accuracy (in percent) of the proposed FNN-based distance estimation algorithm when the network is trained and tested on signals generated in the same specific reverberation time $T_{60} = \{0.4, 0.5, 0.7, 0.9\}$. Results are shown for both simulated (Synth.) and measured (Real) BRIRs.

	T_{60}	Distance (cm)								
		50	100	150	200	250	300	350	400	Overall
Synth.	0.4s	100	100	100	100	100	100	100	100	100
	0.5s	100	100	100	100	100	100	100	100	100
	0.9s	100	100	100	100	100	100	100	100	100
Real	0.4s	100	100	100	100	100	100	100	100	100
	0.7s	100	100	100	100	100	100	100	100	100
	0.9s	100	100	100	100	100	100	100	100	100

use 462 speakers including 326 males and 136 females each of which speak 10 sentences, forming a total set of 4620 training sentences. Each sentence has a duration of approximately 2 s. For the test set we use 168 speakers including 112 males and 56 females each of which speak 10 sentences forming a total set of 1680 test samples. The training and test sets are created such that they contain different speakers.

For the estimation of DRR features which are processed in the STFT domain, we segment signals into overlapping frames of 32 ms using a Hamming window with a frame advance of 8 ms. The DFT size is equal to $M = 512$.

The DRR is estimated for front and rear binaural microphones resulting in 80 = 2 × 40 values for each time frame. Since FNN requires only one-dimensional inputs, we take the mean of the estimated narrowband DRRs across all time frames leading to the feature vector of dimension 80×1. For CNN, however, we compute the estimated narrowband DRRs across 40 consecutive time frames and for both front and rear microphones which result in a feature matrix of dimension $40 \times 40 \times 2$. We train both networks on a set of 4620 feature vectors for each distance randomly split into 85% training and 15% validation sets. Each input vector is normalized to the range of $[-1, 1]$.

The performance of the proposed algorithm is measured in terms of accuracy of the classification which is defined as the percentage of examples for which the model predicts the correct outputs. It is mathematically expressed as

$$\mathrm{A} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \tag{15}$$

where TP and TN indicate the sum of true positive and true negative predictions and FP and FN denote the sum of false positive and false negative predictions, respectively.

We investigate the performance of our approach in different scenarios. The first scenario considers signals generated in a room with the same T_{60} for training and testing. Results for both FNN and CNN and for both synthetic and real BRIRs are presented in Tables 1 and 2. It can be observed from these tables that when the networks are trained and tested on signals generated with the very same reverberation time both networks achieve reliable distance estimation accuracies.

In the second scenario we assess the generalization capability of the proposed learning-based approach. Hence, we train the networks on signals generated with a particular T_{60} and test on signals generated with different T_{60}s. Here, we train the networks on signals generated with $T_{60} = 0.5$ s for synthetic and with $T_{60} = 0.7$ s for measured BRIRs and test on signals generated using all the available T_{60}s. Results for both FNN and CNN classifiers and for both synthetic and real BRIRs are illustrated in Tables 3 and 4, respectively. According to these tables when the reverberation time between the training and the test signals varies slightly, the accuracy of both networks is fairly robust. However, when the network has not been properly trained on the acoustical

Table 2. Accuracy (in percent) of the proposed CNN-based distance estimation algorithm when the network is trained and tested on signals generated in the same reverberation time T_{60} of the room. Results are shown for both simulated (Synth.) and measured (Real) BRIRs.

	T_{60}	Distance (cm)								
		50	100	150	200	250	300	350	400	Overall
Synth.	0.4s	100	100	99.7	100	99.9	100	100	99.9	99.9
	0.5s	100	100	99.6	99.8	99.5	99.4	99.9	99.5	99.7
	0.9s	100	100	99.4	99.2	98.9	99.6	99.4	99.7	98.4
Real	0.4s	100	100	100	100	100	100	100	100	100
	0.7s	100	100	100	100	100	100	100	100	100
	0.9s	100	100	100	100	100	100	100	100	100

Table 3. Accuracy (in percent) of the proposed FNN-based distance estimation algorithm when the network is trained on signals generated with $T_{60} = 0.5$ s for synthetic (Synth.) and with $T_{60} = 0.7$ s for measured (Real) BRIRs and is tested on signals generated using all the available T_{60}s.

	T_{60}	Distance (cm)								
		50	100	150	200	250	300	350	400	Overall
Synth.	0.4s	100	90.4	24.3	37.6	0.1	94.8	45.1	95.3	61
	0.5s	100	100	100	100	100	100	100	100	100
	0.9s	0.1	77.7	43.7	97.0	7.3	10.8	0.1	11.8	31.1
Real	0.4s	100	93.6	1.1	91.4	32.8	1.3	88.6	96.3	63.1
	0.7s	100	100	100	100	100	100	100	100	100
	0.9s	100	100	100	94.4	100	97.6	99.6	100	99

properties of the room, e.g., the reverberation time the deep learning approaches can not estimate the source microphone distances reliably. The outliers for some particular distances shown in these tables may result from either the low sensitivity of these distance classes to the estimated DRR or from the small size of the network. Overall, the DNN-based speaker distance estimation technique will be a reliable candidate provided that the network is trained on signals from different acoustic situations.

In the third scenario we investigate the performance of the proposed DNN-based distance estimation algorithms on the classification of coarse distance classes. For this experiment we train the networks on signals generated for distances in the range of $d_{\text{true}} \in \{1\,\text{m}, 2\,\text{m}, 3\,\text{m}, 4\,\text{m}\}$ which implicitly reduces the resolution of the distance estimation. Similar to the second scenario, we train the networks on signals generated with $T_{60} = 0.5$ s for synthetic and with $T_{60} = 0.7$ s for measured BRIRs and test on signals generated using all the available T_{60}s. Results for both synthetic and real BRIRs and for FNN and CNN are presented in Table 5. Comparing the overall results of Tables 3 and 4 with the overall results of Table 5 shows that the proposed learning based distance estimation approaches can be more reliable for the estimation of coarse distance classes, e.g., in steps of 1 m. Higher accuracies of acoustic distance estimation indeed require more advanced training schemes or employing other modalities provided by, for instance, a radar sensor which is beyond the scope of this paper.

Table 4. Accuracy (in percent) of the proposed CNN-based distance estimation algorithm when the network is trained on signals generated with $T_{60} = 0.5$ s for synthetic (Synth.) and with $T_{60} = 0.7$ s for measured (Real) BRIRs and is tested on signals generated using all the available T_{60}s.

	T_{60}	Distance (cm)								
		50	100	150	200	250	300	350	400	Overall
Synth.	0.4s	93	85	96.2	98.5	16.5	40.5	13.6	55.8	62.4
	0.5s	100	100	99.6	99.8	99.5	99.4	99.9	99.5	99.7
	0.9s	0	4.2	100	0	4.4	14.3	0	7.9	16.4
Real	0.4s	28.9	0	0.1	73.8	93.7	4	32.6	1.3	29.3
	0.7s	100	100	100	100	100	100	100	100	100
	0.9s	100	100	99.1	99	97.3	82.1	21.6	54.8	81.7

Table 5. Accuracy (in percent) of the proposed FNN-based and CNN-based distance estimation algorithm when the network is trained on signals generated with $T_{60} = 0.5$ s for synthetic (Synth.) and with $T_{60} = 0.7$ s for measured (Real) BRIRs and is tested on signals generated using all the available T_{60}s. Here a coarse level of precision for distance classification is tested.

		FNN					CNN				
	T_{60}	Distance (cm)									
		100	200	300	400	Overall	100	200	300	400	Overall
Synth.	0.4s	98.1	69.1	99.5	57.1	81	100	90.0	56.3	75.2	80.4
	0.5s	100	100	100	100	100	100	99.9	99.3	99.3	99.7
	0.9s	91.2	83.3	3.7	24.9	50.8	85.7	47.2	9.4	38.8	45.3
Real	0.4s	100	76.7	18.2	95.2	72.5	89.7	62.5	59.8	100	75.5
	0.7s	100	100	100	100	100	100	100	100	100	100
	0.9s	100	100	97.0	100	99.2	100	99.3	100	79.9	94.8

5 CONCLUSIONS

In this work, we presented a learning-based algorithm for the estimation of a speaker's distance using HA microphones. Our algorithm aims to classify discrete sound source distances using deep neural networks. The networks are trained on estimated DRRs which are computed using the magnitude squared coherence taking the binaural configuration into account. We tested the accuracy of our algorithm on speech signals convolved with synthetic and measured BRIRs at fine and coarse distance classes and under different reverberation times.

Our results show that the speaker's distances can be reliably estimated if the network is trained properly on the acoustical properties of the room. The method achieves a reliable accuracy when the reverberation times of the signals in the training and test sets varies only slightly. For large variations in the reverberation time of the signals under test the estimation is sufficiently accurate only for the classification of coarse distance classes. For the classification of fine distance classes, however, apparent mismatch of training and test data occurs in specific conditions which requires an additional regularization.

References

[1] J. Blauert, *Communication Acoustics*, vol. 2, Springer, 2005.

[2] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using D/R spatial correlation matrix model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2374–2384, 2011.

[3] S. Vesa, "Sound source distance learning based on binaural signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2007, pp. 271–274.

[4] Y. C. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1793–1805, 2010.

[5] A. Brendel and W. Kellermann, "Learning-based acoustic source-microphone distance estimation using the coherent-to-diffuse power ratio," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 61–65.

[6] E. Georganti, T. May, S. Van De Par, and J. Mourjopoulos, "Sound source distance estimation in rooms based on statistical properties of binaural signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1727–1741, 2013.

[7] F. Keyrouz, "Binaural range estimation using head related transfer functions," in *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, Sept 2015, pp. 89–94.

[8] M. Zohourian and R. Martin, "Direct-to-reverberant energy ratio estimation based on interaural coherence and a joint ITD/ILD model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[9] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural networks*, vol. 6, no. 4, pp. 525–533, 1993.

[10] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[11] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 6, 2009.

[12] Christian Borss, *An Improved Parametric Model for the Design of Virtual Acoustics and its Applications*, Verlag Dr. Hut, 2011, PhD. Thesis.

[13] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[14] J. A. Moorer, "About this reverberation business," *Computer Music Journal*, pp. 13–28, 1979.

[15] ITU-T Recommendation P.58, *Head and Torso Simulator for telephonometry*, HEAD acoustics, 2018.

[16] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech communication*, vol. 9, no. 4, pp. 351–356, 1990.