Full-Sphere Binaural Sound Source Localization Using Multi-task Neural Network

Yichen Yang*, Jingwei Xi*, Wen Zhang*†, Lijun Zhang*
* Center of Intelligent Acoustics and Immersive Communications,
School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China
E-mail: {yang_yichen, jingweixi}@mail.nwpu.edu.cn, {wen.zhang, zhanglj7385}@nwpu.edu.cn
† Research School of Electrical, Energy and Materials Engineering,
College of Engineering and Computer Science, The Australian National University, Canberra, Australia

***Abstract*—The accuracy of binaural sound source localization is faced with the challenge of localizing azimuth and elevation simultaneously in noisy and reverberant environments. In this work, a full-sphere binaural sound source localization system is proposed using convolutional neural network and multi-task neural network connected to learn the localization features. The log-magnitudes and interaural phase difference (IPD) of binaural signals are used as inputs to a two-branch convolutional neural network, from which interaural and monaural cues are extracted and combined. Then, the full-sphere localization is formulated as two subtasks of estimating azimuth and elevation separately using multi-task neural network. To reduce reverberation effects, the interaural coherence based pre-processing is used to select the direct-path dominated time-frequency bins for localization. The proposed system is evaluated at a variety of noise and reverberation conditions, in comparison with two baseline systems. The results indicate that the proposed system achieves better localization performance, especially for elevation estimation, at low SNR and strong reverberation conditions.**

I. Introduction

Binaural sound source localization (SSL) aims at achieving the same ability to humans, i.e., identifying the spatial positions of sound sources using two acoustic sensors, by mimicking binaural hearing principles. Comparing with many localisation systems deployed in the audio, radar, and sonar applications, which rely on large arrays of sensors, the major advantages of a two-sensor array are its small size, fast response time and easy calibration. Therefore, SSL in binaural systems, such as in hearing aids, personal sound amplification products (intended to amplify sounds for people who are not hearing impaired) and humanoid robots, has received considerable attention from the audio signal processing community [1], [2], [3], [4], [5].

Humans can achieve source localization in a three-dimensional space mainly by using interaural cues and spectral cues from the signals received at the two ears, known as binaural hearing [6]. Interaural cues refer to interaural time or level difference (ITD/ILD) between signals at the left and right ear and they are used to determine the lateral direction (left, front, right, or frontal horizontal plane) [7]. Spectral cues, caused by scattering and diffraction of sound waves in the pinnae and around human body, are mainly used for localizing the elevation or distinguishing front-back [8]. The head-related transfer function (HRTF) is an acoustic transfer function defined for describing sound propagation from a specific point to the listener's ear and a pair of HRTFs for two ears capture all the localization cues [9].

Early work of binaural SSL focused primarily on azimuth localization, and can be classified into (i) cross-correlation techniques that estimate interaural cues (ITD/ILD) from two microphone signals and by comparing the estimates with the stored dataset to estimate the source azimuth [1], [10], and (ii) model based algorithms that exploit statistics of ITD/ILD through probabilistic models and apply maximum-likelihood estimation for source localisation [2], [11], [12]. While humans actually achieve elevation estimation from the perception of spectral peaks and notches of certain frequencies [13], the method based on spectral differences between the received binaural signals and HRTF data, with spectral pre-processing to remove rapid fluctuations in HRTF data and the spectrum of the sound, has been proposed [5]. Recently, a few systems have been proposed for full-sphere binaural SSL [5], [14], [15]. However, it is still not clear how to combine the interaural and spectral cues in a systematic way for the best results in complex environments.

With the rise of machine learning, neural network based methods have been widely used to solve binaural SSL. HRTFs of 45 subjects in CIPIC [16] database have been analyzed using convolutional neural network (CNN) to exploit the binaural localization cues [3]. Experimental results indicated that CNN can be trained to achieve classification performance comparable to that of humans in a simple sound localization task [17], [18]. The end-to-end system has also been used for binaural SSL [4]. However, when employed in complex environments, binaural SSL still faces a variety of challenges, such as elevation/distance ambiguity, reverberation and interfering sources.

In this work, we develop a full-sphere binaural SSL system as illustrated in Fig. 1. The log-magnitudes and IPD of the binaural signals are used as inputs to two parallel CNNs, from which the interaural and monaural localization cues are extracted. The full-sphere SSL is formulated as two subtasks, i.e., estimating the azimuth and elevation separately using multi-task neural network (MNN). The interaural coherence (IC) based pre-processing stage is used to select the time-frequency bins that are dominated by the direct path, so that reverberation

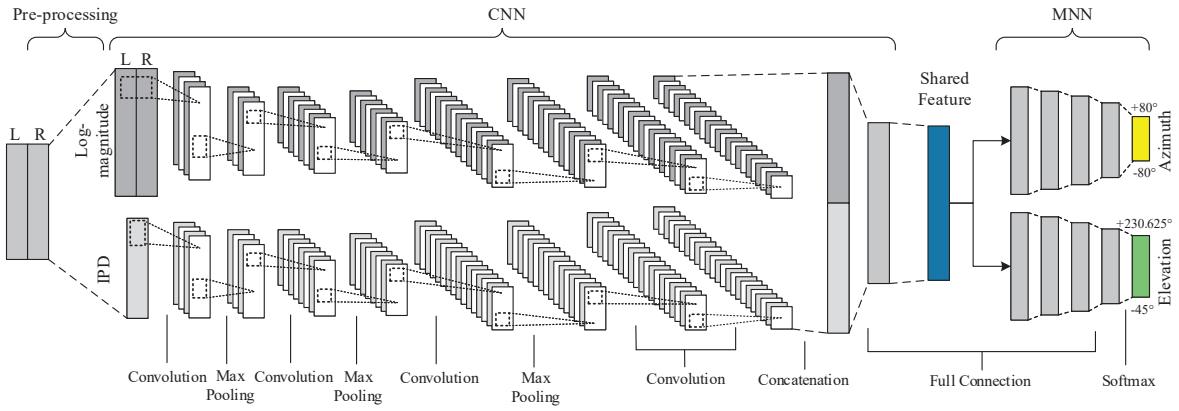

Fig. 1. The proposed full-sphere binaural sound source localization system.

effects can be reduced. The proposed system has been trained at different noise and reverberation conditions, and tested under both trained and untrained conditions. Compared with two baseline systems, the propose system demonstrates more accurate localization results especially in noisy and reverberant environments.

II. Binaural Sound Source Localization Problem

This paper assumes that a single source signal is captured by two microphones, i.e., the left and right ear microphones of a binaural system, in a noisy and reverberant environment. The captured signals at each time-frequency (TF) bin in the short-time Fourier transform domain are written as

$$\begin{aligned} Y_l(t,f) &= S(t,f) \times B_l(f,\Theta) + N_l(t,f) \\ Y_r(t,f) &= S(t,f) \times B_r(f,\Theta) + N_r(t,f), \end{aligned} \tag{1}$$

where t and f indicate time and frequency indices, respectively. $Y(t,f)$, $S(t,f)$, and $N(t,f)$ denote the received binaural signals, the source signal, and the additive noise at a TF point, with l and r representing the left and right ear, respectively. $B_l(f,\Theta)$ and $B_r(f,\Theta)$ are the binaural room transfer functions (BRTFs) of the source position $\Theta \equiv \{\phi, \theta\}$, which can be modelled as a summation of the source position HRTFs, the HRTFs corresponding to the image sources of early reflections, and the acoustic transfer function (ATF) of the late reverberant part. In other words, the BRTFs reduce to the HRTFs, i.e, $H_l(f,\Theta)$ and $H_r(f,\Theta)$, only when it is in a non-reverberant environment, such as in an anechoic chamber. While it is widely adopted to learn or model the HRTF data for binaural SSL, the problem becomes challenging in noisy and reverberant environments.

III. Proposed Approach

A. Localization Feature Extraction by CNN

The log-magnitudes and IPD of binaural signals are used as inputs fed to two parallel convolutional neural networks to extract localization features. That is, the input data of one branch are log-magnitudes of binaural signals,

$$\begin{aligned} E_l(t,f) &= 20\log_{10}|Y_l(t,f)| \\ E_r(t,f) &= 20\log_{10}|Y_r(t,f)|, \end{aligned} \tag{2}$$

and the other branch is fed by the IPD of binaural signals

$$\text{IPD}(t,f) = \angle \frac{Y_l(t,f)}{Y_r(t,f)}. \tag{3}$$

In this work, the signals are sampled at 16 kHz and the window length is 100 ms with 50 ms overlap. The log-magnitude features are extracted as 801 samples in each frame, which represent the frequencies up to 8 kHz.

For the IPD feature, early studies have shown that IPD is more dependable for frequency below 1.5 kHz due to the phase wrapping in high frequency [6], while some recent work has proven that high-frequency IPD cues will contribute to more accurate full-sphere localization results [14]. Therefore, two different frequency range schemes for IPD are investigated in this work. The model that uses only low-frequency IPD is called the LIPD model, which contains 82 samples of the IPD feature for frequency below 1.5 kHz. A second model that uses full-frequency IPD is called the FIPD model, which contains 801 samples of the IPD feature for the full frequency band. In each frame, the left-ear and right-ear log-magnitudes are stacked together to form the magnitude input matrix of size 801×2, and the IPD form another input matrix of size 82×1 or 801×1.

Two independent CNNs are adopted to find the localization features from log-magnitudes and IPD for full-sphere sound localization. As shown in Fig. 1, the IPD are followed by layers of 32 kernels of shape 3×1 to extract interaural features, and log-magnitudes are followed by layers of 32 2D kernels of shape 3×2 to extract interaural and monaural features from inputs simultaneously.

In each branch, the first convolutional layer is followed by 2×1 max pooling, and four more convolutional layers are employed to search for features suitable for localization. The

first two of the four layers use 64 kernels of size 3×1. And the last two use 128 kernels of size 3×1. The first two layers are followed by 2×1 max pooling and all convolutional layers are followed by batch normalization (BN) operations with rectified linear unit (ReLU) activation function respectively.

Finally, the outputs of the two branches are flattened and concatenated together to unite the magnitude feature and IPD feature, and the united feature is fed into two fully-connected hidden layers with 8192 and 4096 units, respectively, to generate the shared feature for the following multi-task sound source localization.

B. Full-sphere Localization by Multi-task Neural Network

In NN-based learning, the typical way of solving a problem is to build a single model for a particular task and optimize the parameters of this model based on certain criterion. However, given only optimized for a single task, the network cannot achieve optimality when several related tasks need to be completed simultaneously.

One appropriate method is to use the shared representations between several related tasks, thus the entire model can be trained together and give the optimal performance for each single task. This kind of approach is called multi-task learning (MTL) [19], which has been used successfully in many applications including sound localization [15], [20].

In the architecture of the proposed multi-task neural network (MNN) as shown in Fig. 1, a hard-parameter sharing strategy has been used. There are two branches in this MNN representing the azimuth and elevation estimation, respectively. Each branch has five fully-connected layers and two parallel output layers with softmax activation.

The cross-entropy loss function for a single azimuth estimation or elevation estimation task is used to train the network. The total loss function L for full-sphere source localization network is designed by the weighted sum of the azimuth loss function L_a and elevation loss function L_e, that is

$$L = \alpha L_a + (1-\alpha) L_e, \tag{4}$$

where α represents the weight with the value ranging from [0, 1]. In this way, the two single-task branches are trained simultaneously by minimizing the total loss function L.

C. Interaural Coherence Pre-Processing

As stated, binaural SSL in reverberant environments is challenging because the signals received at two ears are contaminated by early reflections and late reverberation, which leads to a decrease in the coherence of received binaural signals.

In this work, we adopt interaural coherence based (IC-based) pre-processing to select TF-bins that are dominated by the direct path. The IC is defined as follows

$$\Gamma(t,f) = \frac{\Phi_{l,r}(t,f)}{\sqrt{\Phi_{l,l}(t,f) \times \Phi_{r,r}(t,f)}}, \tag{5}$$

where $\Phi_{l,r}(t,f)$ represents the cross-power spectral density (CPSD), $\Phi_{l,l}(t,f)$ and$\Phi_{r,r}(t,f)$ represent the auto-power spectral density (APSD) of the time aligned signals received at the left and right ear, respectively. A recursive smoothing is applied to estimate the CPSD and APSDs from the received binaural signals processed in frames as in [21], that is

$$\begin{aligned} \Phi_{l,l}(t,f) &= \beta \Phi_{l,l}(t-1,f) + (1-\beta)|Y_l(t,f)|^2 \\ \Phi_{l,r}(t,f) &= \beta \Phi_{l,r}(t-1,f) \\ &\quad + (1-\beta) Y_l(t,f) \times \overline{Y_r(t,f)}, \end{aligned} \tag{6}$$

where β represents the smoothing parameter. $\Phi_{r,r}(t,f)$ is obtained in a similar way of calculating $\Phi_{l,l}(t,f)$.

IV. Experiments and Analyses

A. Experimental Setup

Binaural signals are simulated by convolving the binaural room impulse response (BRIR) with the speech signals from the DARPA TIMIT database [22]. The BRIRs are generated using the HRTF of subject 003 in CIPIC database and the room impulse responses (RIR) [23] simulated with the image-source method. A simulated room of size 5 m × 5 m × 3 m is created and the subject head is located at the room center, i.e., (2.5, 2.5, 1.5) m. The sound source is positioned at 1250 directions on the sphere around the subject, i.e., 25 azimuth angles ranging from $-80°$ to $80°$ and 50 elevation angles ranging from $-45°$ to $230.625°$, which is the same position scheme as used in the CIPIC database.

At each direction, the training dataset is obtained by selecting randomly 24 different speech signals from the TIMIT train set, while another 6 different speech signals form the validation dataset and 10 more speech sentences are selected to create the test dataset.

The model is trained at a variety of reverberation and noise conditions, and tested under both trained and untrained conditions. In the reverberation test, the corresponding T_{60} varies from 200 to 500 ms approximately by manipulating the absorption coefficients of the walls. As for the noise test, additive white gaussian noise is used to generate data with SNRs varying from 10 to 30 dB. Multi-conditional Training (MCT) is applied to allow the network to learn the various features and enhance the performance under different conditions. The data with T_{60} of 150, 250, 350, 450 ms and SNRs of 5, 15, 25, 35 dB are used for training, and the data with T_{60} of 200, 300, 400, 500 ms and SNRs of 10, 20, 30 dB are used for validation and test.

The *adam* optimizer and a decreasing learning rate initialized at 1e-3 are adopted for training the network. If the error on the validation set does not decrease in 2 epochs, the learning rate is multiplied by 0.5. The early stopping is applied after at least 5 epochs. The weight α for MNN is set to 0.5.

The performance of the proposed system is compared with two state-of-the-art baselines. One uses the composites feature vector of the IPD and ILD that are selected based on the analysis of mutual information to improve the 3D localization performance in complex conditions [14]. The other system uses IPD and ILD directly as inputs to a time-frequency

TABLE I
COMPARISON OF AZIMUTH AND ELEVATION ESTIMATION ACCURACY [%] AT DIFFERENT SNRS.

(a) Azimuth Accuracy Comparison								
SNR	No Noise	35dB	30dB	25dB	20dB	15dB	10dB	5dB
[14]	**99.44**	**-**	**98.88**	-	97.20	-	94.40	-
[15]	98.08	97.87	97.46	96.88	95.57	93.05	88.48	79.87
Proposed-LIPD (without IC)	98.07	98.03	98.01	97.98	97.83	97.27	94.99	86.24
Proposed-LIPD	97.91	97.86	97.80	97.71	97.45	96.47	93.02	82.38
Proposed-FIPD (without IC)	98.10	98.10	98.10	**98.10**	**98.09**	**98.07**	**97.94**	**96.95**
Proposed-FIPD	98.03	98.02	98.02	98.01	97.99	97.96	97.75	96.44
(b) Elevation Accuracy Comparison								
SNR	No Noise	35dB	30dB	25dB	20dB	15dB	10dB	5dB
[14]	96.08	-	89.60	-	72.64	-	37.04	-
[15]	**100**	97.67	95.73	92.42	86.93	78.37	65.77	48.47
Proposed-LIPD (without IC)	92.87	91.24	90.43	89.10	86.69	81.44	70.33	50.02
Proposed-LIPD	92.37	90.52	89.47	87.78	84.55	78.10	64.91	43.50
Proposed-FIPD (without IC)	99.21	**98.83**	**98.63**	**98.28**	**97.59**	**96.17**	**93.06**	**85.25**
Proposed-FIPD	99.18	98.71	98.45	98.02	97.26	95.75	92.42	84.56

TABLE II
COMPARISON OF AZIMUTH AND ELEVATION ESTIMATION ACCURACY [%] AT DIFFERENT REVERBERATION TIMES (T_{60}).

(a) Azimuth Accuracy Comparison								
T_{60}	150ms	200ms	250ms	300ms	350ms	400ms	450ms	500ms
[14]	-	94.32	-	91.44	-	89.44	-	78.88
[15]	**99.05**	95.95	96.19	91.60	92.12	87.44	90.40	83.64
Proposed-LIPD (without IC)	96.37	95.38	96.18	94.04	95.43	91.67	94.09	87.48
Proposed-LIPD	96.91	96.30	96.98	95.58	96.58	93.94	95.37	90.34
Proposed-FIPD (without IC)	97.26	95.53	96.21	94.23	95.77	92.57	94.98	90.02
Proposed-FIPD	98.83	**97.47**	**98.23**	**97.06**	**98.17**	**96.18**	**97.70**	**94.45**
(b) Elevation Accuracy Comparison								
T_{60}	150ms	200ms	250ms	300ms	350ms	400ms	450ms	500ms
[14]	-	75.84	-	68.48	-	55.52	-	42.64
[15]	**99.06**	96.26	95.96	91.76	91.73	86.93	89.57	81.70
Proposed-LIPD (without IC)	95.56	94.75	96.37	93.21	95.51	90.14	93.89	84.61
Proposed-LIPD	96.89	96.47	97.68	95.76	97.18	93.44	95.70	88.49
Proposed-FIPD (without IC)	95.00	94.32	95.34	93.09	95.08	91.13	94.43	87.47
Proposed-FIPD	98.66	**98.46**	**98.91**	**97.99**	**98.73**	**96.69**	**98.28**	**93.87**

convolutional neural network for full-sphere binaural SSL. [15]. In order to test the performance of this system in different reverberant conditions, the same MCT setting is used to train the network. The same descending learning rate scheme is used for training, while other parameters are consistent with the original paper. The performance measure is the localization accuracy with the angular error tolerance of 0°.

B. Results and Discussion

Table I and II show the performance comparisons between the proposed method and baseline systems in different noise and reverberation conditions. In each table, (a) represents the azimuth localization results and (b) represents the elevation localization results, respectively. Compared with two baseline systems, the proposed method achieves the best performance under most conditions, especially for low SNR (SNR$\leq$ 25 dB) and strong reverberation ($T_{60} \geq$ 200 ms) conditions. Note that the proposed system has roughly the same localization accuracy over the 1250 points on the full sphere, however to be consistent with results of the baseline systems, only horizontal-plane azimuth estimation results and median-plane elevation estimation results are shown here.

In the proposed system, the log-magnitudes and IPD of the binaural signals, which include both the interaural cues and monaural cues, are fed into the CNN for feature selection. This operation leads to more accurate source localization, especially in terms of elevation estimation under noisy and reverberant conditions, as shown in Table I and II. The two baseline systems perform better only at high SNR and low reverberation conditions. As shown in Table II, the baseline system of [14] is severely contaminated by reverberation. Compared with the approach using the same methodology of MNN for full-sphere localization [15], the proposed method

maintains reasonably accurate estimations even without the IC pre-processing step for $T_{60} \geq 250$ ms. These experimental results confirm that when employed in complex environments, using binaural magnitudes rather than using the interaural magnitudes can more accurately preserve localization cues. The IC-based pre-processing that are used to select the direct-path dominated TF bins can bring further improvement for localization in reverberant environments as shown in Table II.

As for the two proposed IPD models, the FIPD model, which uses the full frequency range IPD information, demonstrate better localization performance than that of the LIPD model, which only uses IPD for frequency below 1.5k Hz. These results indicate that there are some key localization cues in the high-frequency IPD spectrum for full-sphere SSL, which should be included to achieve more accurate localization results.

V. Conclusions

This paper proposed a full-sphere binaural sound localization system that uses the log-magnitudes and IPD of binaural signals as inputs and combines two parallel cascades of CNNs with a multi-task neural network to learn the shared localization features. The full-sphere localization problem is formulated as two subtasks of estimating azimuth and elevation localization in the MNN, with an IC-based pre-processing to reduce reverberation influence. The proposed system are validated in a variety of noise and reverberation conditions, and demonstrated significant improvement of localization accuracy compared with two baseline systems.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (NSFC) funding scheme under Project No. 61831019.

References

[1] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, pp. 68–77, 2010.

[2] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, pp. 1503–1512, 2012.

[3] E. Thuillier, H. Gamper, and I. J. Tashev, "Spatial audio feature discovery with convolutional neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6797–6801, Calgary, AB, April 2018.

[4] P. Vecchiotti, N. Ma, S. Squartini, and G. J. Brown, "End-to-end binaural sound localisation from the raw waveform," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 451–455, Brighton, United Kingdom, May 2019.

[5] B. R. Hammond and P. J. Jackson, "Robust full-sphere binaural sound source localization using interaural and spectral cues," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 421–425, Brighton, United Kingdom, May 2019.

[6] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[7] E. A. Macpherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *J. Acoust. Soc. Am.*, vol. 111, pp. 2219–2236, 2002.

[8] F. Asano, Y. Suzuki, and T. Sone, "Role of spectral cues in median plane localization," *J. Acoust. Soc. Am.*, vol. 88, pp. 159–168, 1990.

[9] C. I. Cheng and G. H. . Wakefield, "Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency and space," *J. Audio Eng. Soc.*, vol. 49, pp. 231–249, 2001.

[10] R. Parisi, F. Camoes, and A. Uncini, "Cepstrum prefiltering for binaural source localization in reverberant environments," *IEEE Signal Processing Letters*, vol. 19, pp. 99–102, 2012.

[11] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization soure separation and localization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, pp. 382–394, 2010.

[12] T. May, S. V. D. Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, pp. 1–13, 2011.

[13] M. M. Van Wanrooij and A. J. Van Opstal, "Contribution of head shadow and pinna cues to chronic monaural sound localization," *Journal of Neuroscience*, vol. 24, pp. 4163–4171, 2004.

[14] X. Wu, D. S. Talagala, W. Zhang, and T. D. Abhayapala, "Individualized interaural feature learning and personalized binaural localization model," *Applied Sciences*, vol. 9, no. 13, p. 2682, 2019.

[15] C. Pang, H. Liu, and X. Li, "Multitask learning of time-frequency cnn for sound source localization," *IEEE Access*, vol. 7, pp. 40725–40737, 2019.

[16] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipic hrtf database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, pp. 99–102, IEEE, 2001.

[17] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 12, pp. 2444–2453, 2017.

[18] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Localizing speakers in multiple rooms by using deep neural networks," *Computer Speech & Language*, vol. 49, pp. 83–106, 2018.

[19] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[20] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2386–2390, Calgary, AB, April 2018.

[21] I. Kossyk, M. Neumann, and Z.-C. Marton, "Binaural bearing only tracking of stationary sound sources in reverberant environment," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pp. 53–60, IEEE, 2015.

[22] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[23] E. A. Lehmann and A. M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1429–1439, 2009.