Deep Learning-based Speech Source Localization using Binaural and Monaural Microphone Arrays

Peyman Goli[1], Steven van de Par[2]

Acoustics Group, Department of Medical Physics and Acoustics, Oldenburg University, Cluster of Excellence - Hearing4all

[1] *peyman.goli@uol.de*

[2] *steven.van.de.par@uol.de*

Introduction

Speech source localization in hearing aids (HAs) can provide useful information to improve speech intelligibility for hearing-impaired listeners, for example by using this information to steer a beamformer. Interaural time difference (ITD) and interaural level difference (ILD) between the ears are the two major cues which are exploited by the human auditory system to horizontally localize sound sources in an environment. ITDs are estimated on the basis of the difference in arrival time of a sound between the ears, and ILDs are the level difference between the sounds reaching the ears, which is attributed by the head shadow effect. Given the human ability to detect and use the binaural cues for speech source localization, some conventional methods (i.e. non-deep learning-based methods) have used the models of computational auditory scene analysis to localize speech sources [1][2]. For instance, in [1], a peripheral processing based on the human auditory system is applied on the binaural microphone channels to extract the interaural cues. Then, a shallow learning model (e.g. Gaussian mixture models) is used to create a mapping between the interaural cues and corresponding azimuth angles in a limited range between -90° and 90°. Increasing the number of microphone channels in binaural microphone arrays is an efficient approach to improve the performance of localization in a full azimuth angle range and at adverse acoustic conditions.

However, although using more microphone channels can further improve the resolution of the interchannel cues, it causes excessive cue combinations between the channels. Determining the contributions of the excessive interchannel cues in speech source localization and creating a mapping between the cues and corresponding azimuth angles at different acoustic conditions is a challenging task which cannot be handled by shallow learning models. In the proposed method, six channels of a binaural microphone array are used to localize speech sources in a full range of azimuth angles. A peripheral processing is applied to determine the auditory subband signals, and then, a deep neural network (DNN) is used to extract the interchannel cues and exploit the spatial representations throughout the deep layers for localizing speech sources.

Using binaural microphones in HAs has remained a challenging task because of the complexity of hardware and software connections between the microphones. Due to the short distances between the channels in a monaural microphone array, very small time delays between the interchannel signals are found, and the monaural positions of the microphone channels can lead to a low interchannel level difference because of the lack of the head shadow contribution in the level difference. The smaller time delays and level differences between the interchannel signals have led to fewer studies employing monaural microphone arrays for speech source localization in HAs. In this research, a deep model is also trained on the basis of the auditory subband signals of a monaural microphone array (i.e. four channels of the right ear) to extract and exploit the spatial representations of monaural signals for speech source localization.

The paper is organised as follows. Firstly, the proposed auditory peripheral stage is described. Then, the implementation and training of DNNs using the binaural and monaural microphone arrays are explained. Lastly, the performance of the trained deep models in localizing speech sources is presented.

Proposed method

Two deep learning-based models are proposed to horizontally localize speech sources in a noisy and reverberant environment by using binaural and monaural microphone arrays of HAs. A peripheral preprocessing is applied on microphone channels to determine the auditory subband signals for simulating the human auditory system in speech source localization. The three-dimensional (3D) input features are prepared to apply to the deep models. A regression model based on deep convolutional neural network is used to localize the speech source corresponding to the input subband signals.

Preprocessing

A preprocessing stage is applied on the signals received by the microphone arrays to prepare the input features of the DNN. The preprocessing stage consists of a peripheral processing based on the human auditory system and an input-feature preparation stage. In the peripheral processing stage, an auditory filter bank is applied on the signals to simulate the activity of the human cochlea in the inner ear [1]. A fourth-order gammatone filter bank decomposes the signals into $N = 32$ auditory subband signals. The band centre frequencies are distributed uniformly in terms of equivalent rectangular bandwidth between 80 Hz and 5 kHz [3]. Then, the neural transduction process in the inner hair cells is simulated by applying halfwave-rectification and square-root compression on the subband signals. In the input-feature preparation stage, a min–max scaling is firstly applied on the subband signals between 0.1 and 0.9 to prevent the large gradient variance problem when updating the deep model parameters during the training stage. A rectangular window with the size of 20 ms at a sampling rate of 44.1 kHz (i.e. with the length of 882 samples) is applied to segment the scaled subband signals in the time domain.

Then, four context frames are added to the corresponding input frame (i.e. two past frames and two future frames) to extend the time interval for speech source localization (i.e. the total input duration is 5×20 ms = 100 ms). The number of the time samples is reduced to $M = 2205$ by using a down-sampling step with the rate of 2 to 1. The inputs of the deep model are 3D features with the size of [frequency bands (N), time samples (M), microphone channels (P)].

Desired output

To predict the azimuth angle of speech sources by the deep regression models, the desired outputs are determined on the basis of the likelihood of azimuth angles by using a Gaussian function whose mean value is the target azimuth angle.

$$\phi_m(\theta) = \exp[-(\theta - \theta_m)^2/\sigma^2] \;\; ; \theta = [1, \dots, 360] \qquad (1)$$

where $\phi_m(\theta)$ is the desired output at azimuth angle θ, which corresponds to m^{th} time frame of a speech source positioned at the azimuth angle of θ_m to the head reference. σ^2 is the variance of the Gaussian function that determines the width of the curve around the target angle, as shown in Figure 1. In this experiment, parameter σ is set to 5° based on the human auditory resolution.

Deep model

A deep regression model based on 2D convolution neural network is used to create a complex mapping between the 3D input features and corresponding desired outputs to predict the azimuth angles of speech sources. The mean square error is used as the loss function between the desired and predicted outputs over batches:

$$MSE = \frac{1}{360B} \sum_{m=1}^{B} \sum_{\theta=1}^{360} [\phi_m(\theta) - \hat{\phi}_m^{\boldsymbol{w}}(\theta)]^2 \qquad (2)$$

where $\boldsymbol{w}$ denotes the parameters of the deep model (i.e. weights and biases), which should be learned on the basis of the gradient of the loss function by using backpropagation learning over batches during the training stage. $\hat{\phi}_m^{\boldsymbol{w}}(\theta)$ is the predicted output at m^{th} time frame by the model with the updated parameter $\boldsymbol{w}$, and B is the batch size.

The architecture of the deep model is formed on the basis of four blocks, as shown in Figure 2. In block 1, three convolutional layers are applied on time samples to estimate optimal representations of input features throughout the expanded channels. The dimension of time samples is also reduced by using strided convolutions. In the next block, five residual blocks are applied to improve gradient flow during backpropagation and prevent the vanishing and exploding gradient problems in deeper layers [4]. As illustrated in Figure 3, each residual block is composed of a bottleneck block (i.e. a stack of three convolutional layers with a *bottleneck* at the middle layer) and an identity mapping from input to output of the bottleneck block. In block 3, a 1×1 convolutional layer with 360 channels is firstly applied to extract the optimal representations of previous layer over the azimuth angles. Then, a max-pooling layer with the size of 1×5 is applied on time samples to pool the most prominent samples with more energy. Lastly, an axes swapping is used to change the convolution domain from frequency-sample to frequency-angle in the rest of the model [5]. In the last block, two convolutional layers with the kernel size of 16×5 and 9×5 are applied (i.e. by using zero padding along the azimuth angles) to extract the representations of the frequency-angle domain with filters having expanded channels, and reduce the frequency dimension. In the last layer, a 1×1 convolutional layer with single channel is applied to create the output array with the size of 360×1.

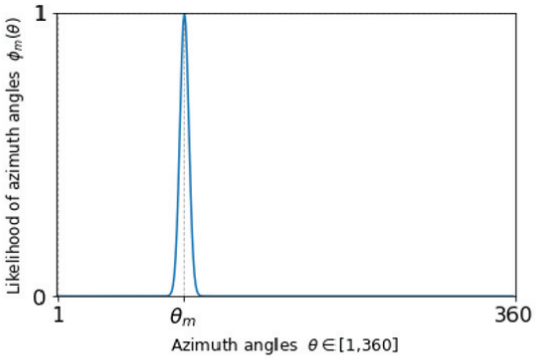

Figure 1: Likelihood of azimuth angles around the target azimuth angle for $\sigma = 5°$

The batch normalisation [6] and Swish activation function [7] are applied after all convolutional layers except the last layers of the residual blocks, where the activation function is applied after the summation unit of the residual blocks. The unboundedness in positive values and the non-monotonicity of Swish activation function [i.e. $Swish(x) = x * Sigmoid(x)$] can further improve the gradient flow in deeper layers in comparison to the conventional activation functions such as RLU (rectified linear unit) and Sigmoid activation functions.

EXPRIMENTS

Microphone arrays

The BKwHA HRTF dataset [8], which is recorded on Brüel & Kjær HATS, is used to simulate the head-related transfer functions of the microphone channels in an 8-channel

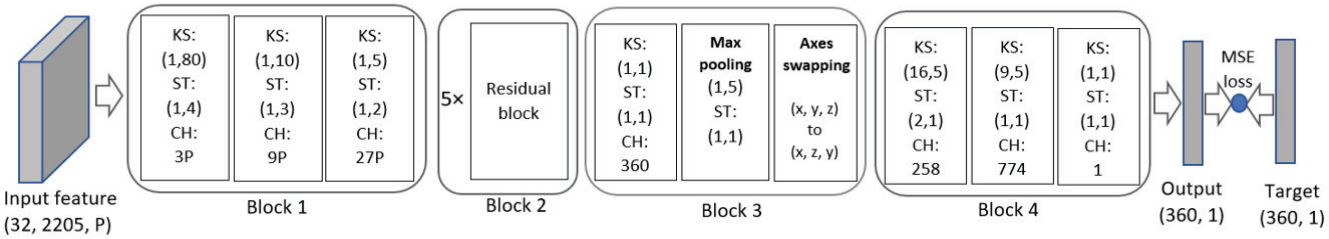

Figure 2: The architecture of the deep models used for SSL. The parameters of the convolutional layers are abbreviated as follows, **KS**: kernel size, **ST**: stride, and **CH**: channels. Parameter P is the number of microphone channels. P is 6 and 4, respectively, in the binaural and monaural methods.

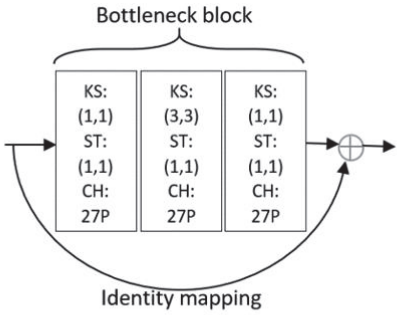

Figure 3: Block diagram of the residual blocks used in the deep models. The parameters of the convolutional layers are abbreviated as Figure 2.

behind-the-ear (BTE) HA [9]. Six microphone channels behind the two ears and four microphone channels behind and in the right ear are respectively used in the binaural and monaural microphone arrays used in the proposed method to capture the audio signals in a noisy and reverberant environment.

Preparing datasets and training models

The impulse responses of an acoustic room are simulated using the perceptually plausible room acoustics simulator (RAZR [10]). The dimensions of the room are 4.25 m length, 3.75 m width and 2.5 m height, and the head is positioned near the room centre of 1.9, 1.8 and 1.8 m. As shown in Figure 4, one speech source and one noise source are located on a circle with the radius of 1.7 m around the head position at random azimuth angles between 1° and 360° with 5° resolution. To prepare the training dataset, 2000 utterances from 200 different speakers (male and female) from TIMIT database [11] and three noise types (Babble, Factory and Car engine noise) from Noisex-92 database [12] are respectively randomly chosen for the speech and noise sources at random SNRs from −5 dB to 30 dB with the step of 5 dB. $T60 = 0.2$ s is chosen for the room reverberation time. The training dataset is composed of 5200 randomised sound source positions with the duration of around 37 h. In the test dataset, 200 utterances from 20 different speakers which are not used in the training dataset are used at 700 randomised sound source positions and SNRs in the same acoustic room.

Two deep regression models based on the binaural and monaural microphone arrays are trained to predict the azimuth angles of speech sources in the acoustic room. Parameter P has values of 6 and 4 in the deep models using binaural and monaural microphone arrays, respectively. Adam optimiser [13] is used as an adaptive optimiser by the backpropagation algorithm to update the model parameters during the training stage. The mini-batch size B is 32 and the learning rate is set to 10^{-4} for all epochs in both deep models. Each epoch requires about 3 days by using one GPU Tesla P100 and three CPUs Intel Xeon E5-2650 provided by the Carl cluster of Oldenburg University.

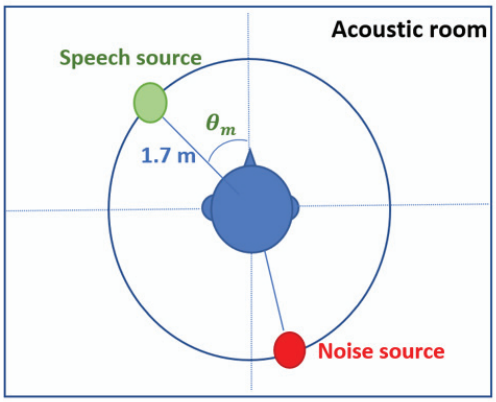

Figure 4: Positions of speech and noise sources in the acoustic room

Results

The models localize the speech sources by finding the maximum of the estimated output as the likelihood of speech source localization. The performance of the models is evaluated on three time intervals at different SNRs. The model output corresponding to 20 ms (i.e. the length of the input frame) and the average of the outputs corresponding to 80 ms and 160 ms (i.e. four and eight successive input frames) are considered the estimated outputs. The correct prediction occurs when the maximum argument of the estimated output (i.e. the estimated azimuth angle) is within the true azimuth angle range with 5° resolution:

$$\theta_m - 2 \leq \text{argmax}[\hat{\phi}_m(\theta)] \leq \theta_m + 2 \qquad (3)$$

The average of the correct prediction scores over the available noises is illustrated in Figure 5. The performance of the models in speech source localization improves as the time intervals of output estimation increase, and higher

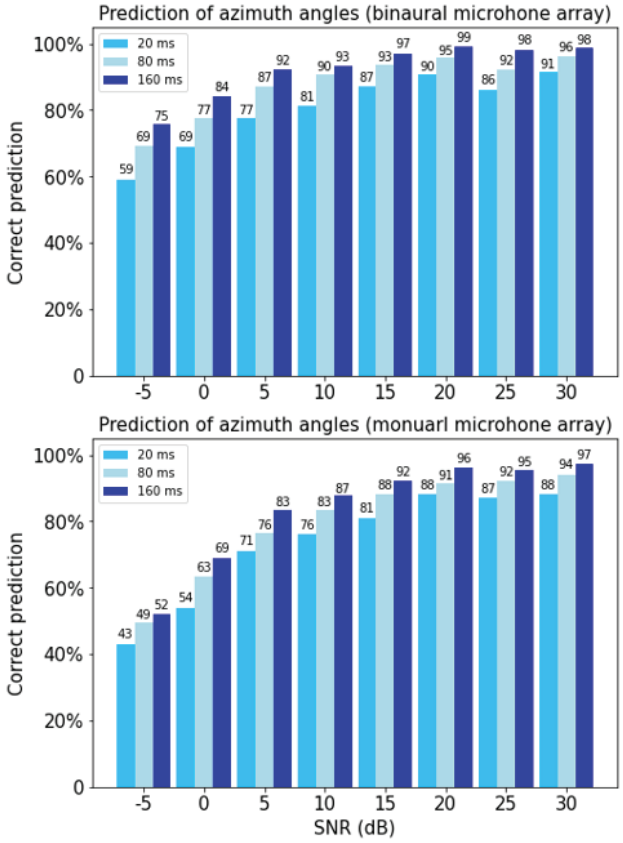

Figure 5: Performance of the proposed method in speech source localization by using binaural and monaural microphone arrays

prediction scores at each SNR are achieved in 180 ms time intervals in both models.

The model using binaural microphone array with considerable interchannel cues can localize speech sources in a full azimuth angle range with the accuracy over 90% at the SNRs more than 5 dB. The deep model using monaural microphone array can achieve the correct prediction scores above 95% at higher SNRs. This achievement is important in speech source localization based on monaural microphone arrays in HAs despite of the short distances and monaural positions of the microphone channels, which lead to inconsiderable interchannel time delays and level differences. However, the masked speech sources at lower SNRs cannot provide detectable ITDs and ILDs for localization in monaural microphones compared with the binaural microphones.

Conclusion

New deep learning-based methods for speech source localization in full range azimuth angles are proposed using binaural and monaural microphone arrays in HAs. An auditory peripheral stage is used to prepare the auditory subband signals. Then, a convolution neural network is utilised as a regression model to create a complex mapping between the auditory subband signals and corresponding azimuth angles. The model using six channels of the binaural microphone array has acceptable performance in speech source localization at noisy and reverberant conditions because of the remarkable interchannel cues. Despite of the monaural positions of the microphone channels in the monaural microphone array, which leads to very small ITD and ILD values, the model can localize the speech sources with the accuracy more than 95% at higher SNRs. This achievement is vital in speech source localization in monaural microphone arrays in HA devices.

Acknowledgement

The research has been founded and supported by the cluster excellence-Hearing4All in Oldenburg University.

References

[1] T. May, S. van de Par, and A. Kohlrausch. “A probabilistic model for robust localization based on a binaural auditory front-end.” *IEEE Transactions on audio, speech, and language processing* 19.1 (2010): 1-13.

[2] N. Roman, D.L. Wang, and G.J. Brown. “Speech segregation based on sound localization.” *The Journal of the Acoustical Society of America* 114.4 (2003): 2236-2252.

[3] N. Yalta, K. Nakadai, and T. Ogata. “Sound source localization using deep learning models.” *Journal of Robotics and Mechatronics* 29.1 (2017): 37-48.

[4] W. He, P. Motlicek, and J.M. Odobez. “Neural Network Adaptation and Data Augmentation for Multi-Speaker Direction-of-Arrival Estimation.” *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2021).

[5] B.R. Glasberg, and B.C.J. Moore. “Derivation of auditory filter shapes from notched-noise data”. *Hearing research* 47.1-2 (1990): 103-138.

[6] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition.” *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[7] P. Ramachandran, B. Zoph, and Q.V. Le. “Searching for activation functions.” *In: Proc. Int’l Conf. on Learning Representations (Workshop track)*. 2018.

[8] S. Ioffe, and C. Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” *International conference on machine learning.* PMLR, 2015.

[9] J. Thiemann, and S. van de Par. “A multiple model high-resolution head-related impulse response database for aided and unaided ears.” *EURASIP Journal on Advances in Signal Processing* 2019.1 (2019): 1-9.

[10] H. Kayser, S.D. Ewert, and J. Anemüller. “Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses.” *EURASIP Journal on Advances in Signal Processing* 2009 (2009): 1-10.

[11] T. Wendt, S. van de Par, and S.D. Ewert. “A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation.” *Journal of the Audio Engineering Society* 62.11 (2014): 748-766.

[12] J.S. Garofolo, L.F. Lamel, and W.M. Fisher. “Darpa Timit acoustic-phonetic continuous speech corpus cd-rom.” *nist speech disc* 1–1.1. 1993. p. 27403

[13] H.J.M. Steeneken, and F.W.M. Geurtsen. “Description of the RSG.10 noise database.” *TNO Institute for Perception*, 1988. (Technical Report IZF 1988-3)

[14] D.P. Kingma, and J. Ba. “Adam: A method for stochastic optimization.” *arXiv preprint arXiv*:1412.6980 (2014).