DOI: 10.61782/fa.2025.0823

FORUM ACUSTICUM EURONOISE 2025

SPEECH POSITION ESTIMATION USING NEURAL NETWORKS FROM ACOUSTIC SIGNALS OF HEARING AIDS

Sebastián Guajardo-Herrera[1*] **Rhoddy Viveros-Muñoz**[1]

[1] Department of Electronics and Computer Sciences, Universidad Técnica Federico Santa María, Concepción 4030000, Chile.

ABSTRACT

The localization of sound events, including speech, has evolved with the use of deep learning algorithms. The inclusion of distance in the task of estimating the direction of arrival allows us to obtain the spatial position of the sound source, which is a variable of interest for the improvement of auditory perception. This study addresses speech direction of arrival and distance estimation using signals generated by simulated hearing aid microphones in acoustic virtual reality environments.

The approach involves the collection of a dataset designed to represent various scenarios, with 96,000 samples distributed at 24 angular positions, varying distances between the source and the listener (ranging from 1 to 3 meters) and 4 reverberation times (0.1, 0.3, 0.6, and 0.9). These data are then used to evaluate the performance of a deep learning algorithm in simulated scenarios that replicate real conditions. The objective of this work is to contribute to the development of techniques that integrate direction of arrival and distance estimation in applications related to hearing aid systems and other acoustic support devices.

Keywords: Direction of Arribal Estimation, Distance Estimation, Deep Learning, Neural Networks, Hearing Aids

1. INTRODUCTION

The perception of the position of a sound event can be decomposed in the estimation of the direction of arrival (DoA), that refers to the angular direction of the event with respect to the listener, and the sound distance estimation (SDE) that is the relative distance between the source and the listener. These parameters allow to characterize spatially one or more sources, allowing to use them in different systems and applications. These parameters have been previously studied using different signal processing and deep learning (DL) techniques [1, 2].

Generally, implemented models for these tasks has been: convolutional neural networks (CNN) [3, 4], residuals networks (ResNet) [5], recurrent convolutional networks (CRNN) [6], Conformers [5, 7], and Transformers [8]. Among the input features used we can find magnitude and phase spectrograms, mel spectrograms, intensity vectors in the case of ambisonic microphones, interaural phase difference, generalized cross-correlation and Spatial Cue-Augmented Log-Spectrogram (SALSA) [9, 10] being particularly prevalent. These features have proven to be effective in obtaining spatial characteristics of sound sources.

A sound source of interest is the voice, which is crucial in communication processes, knowing the spatial position can improve the assistance systems as well as robotic applications [11].

In the case of hearing aids, they have been applied for the estimation of DoA [12], but distance estimation has not been incorporated. In this study we seek to test the effectiveness of deep learning algorithms for the estimation of both: DoA and distance.

The data used in this study have been generated in acoustic virtual reality [13], which enables faster data acquisition and fully accurate labels, avoiding the need for fully measured and controlled environments.

*__Corresponding author__: sguajardo@doctoradoia.cl.

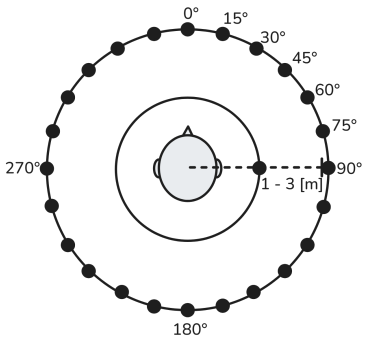

Figure 1. The positions of the simulated sources were configured in azimuth every 15°, elevation 0°, and distance between 1 and 3 meters.

2. METHODOLOGY

2.1 Dataset and Acoustic Simulation

Using a dataset of recorded speech and a controlled simulation framework, we investigated whether a neural network can learn the position of speech using hearing aid signals. The speech signals consisted of 700 sentences adapted to Chilean Spanish from the SHARVARD corpus [14]. These sentences were recorded with high quality microphones at a sampling rate of 44.1 kHz.

The speech signal was positioned at different distances and DoAs, varying in intensity, reverberation and source directivity using the simulation framework Room Acoustic for Virtual Environments (RAVEN) [15]. The listener was simulated as a 4 channel hearing aid. Hearing aid transfer functions (HATF) were obtained from the IHTA-indHARTF dataset [16]. This variability allows the models to generalize to different head shapes and sizes, avoiding the limitation of training to a single type.

The generated data contains the following spatial characteristics: 24 different azimuth angles between 0-360° every 15°, random distance between listener and event in the range of 1-3[m], random directivity of the source between 0-360° (see Figure 1).

The total data was generated is 96,000 samples of 4 channel audio, with a uniform distribution for both DoA and distance. For training the network, this was divided into training and validation sets with an 80-20 ratio.

2.2 Feature Extraction

The feature used for training and inference was a Mel Spectrogram obtained using Torchaudio library, to extract frequency and time features from the 4 microphones channels. We used a Hanning window with 1024 samples with a hop size of 512 samples, to get a magnitude spectrogram and applied a mel filterbank with 64 filters. The resulting input tensor has the shape [C x F x T], where C = 4 (left/right front and rear microphones), F corresponds the frequency bins (64) and T denotes the number of time frames.

2.3 Neural Network Model

The model used was a CNN designed to estimate the position (DoA + Distance) of a sound source, in this case speech, from a 4-channel time-frequency representation of the hearing aid. The architecture consists of:

1. **Input Layer**: Mel spectrogram per channel with shape [4 x F x T].
2. **Convs Layers**: Extract features from time-frequency representation, increasing filter depths (64, 128, 256, 512), each layer followed by ReLU activation function and 2D Max-Pooling. In the training process we used a Dropout (rate = 0.1).
3. **Flattening and Fully Connected Layer**: Flattened features passed through a dense layers with 64 and 32 hidden units.
4. **Output Layer**: Use a two neurons output, corresponding to azimuth angle (in scaled degree) and distance (in scaled meters).

The size of the network was approximately 2 million of parameters, and the model was implemented using PyTorch version 2.4.1.

3. RESULTS AND DISCUSSIONS

The overall results of the network for all validation data can be seen in Table 1, where we can observe the errors of the predictions in terms of DoA (in degrees) and Distance (in meters), measured in Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

The effect of the different reverberation times can be seen for DoA in Figure 2 and for Distance in Figure 3, in both cases we can observe an exponential increase of the error.

Table 1. MAE and RMSE calculated using the validation data set. Azimuth DoA is displayed in degrees and distance between listener and event is displayed in meters.

	MAE	RMSE
Distance [m]	0.40	0.49
DoA [°]	24.91	35.1

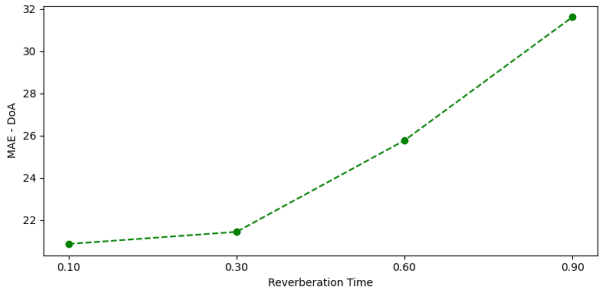

Figure 2. MAE of DoA estimation plotted as a function of reverberation time.

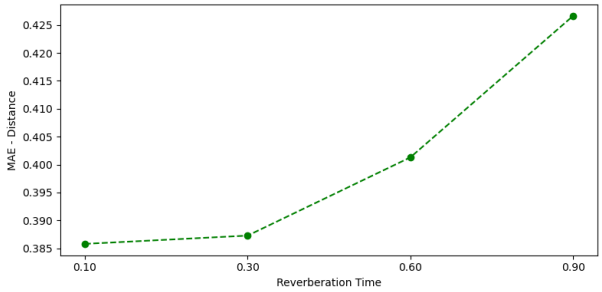

Figure 3. MAE of Distance estimation plotted as a function of reverberation time.

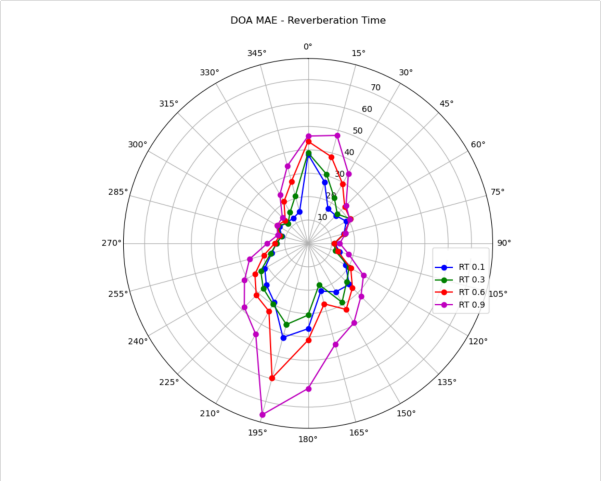

Figure 4. MAE of DoA plotted as a function of source azimuth angle. The polar plot includes the result for 4 reverberation times: 0.1(blue), 0.3(green), 0.6(red), 0.9(magenta).

Finally, the results of the error by DoA are presented in Figure 4 we can see an increase of the error in the rear zone of the listener, since the best estimation angles are the lateral ones.

4. CONCLUSIONS

This work investigate whether DL architectures, such as CNN, are capable of estimating the position of a speech sound source from signals coming from hearing aids. In this work, it is possible to demonstrate the architecture capability to perform this work from using only spectrograms of Mel as feature extraction, as well as characterize how the estimation varies according to the reverberation and position of the source.

Given the results, as future work we will analyze the possibility of incorporating more features and measure the impact that these have on improving the results, as well as incorporate improvements to the proposed model, thinking about applications in hearing aid systems or other acoustic support systems.

5. REFERENCES

[1] G. Jekateryńczuk and Z. Piotrowski, “A Survey of Sound Source Localization and Detection Methods

and Their Applications," *Sensors*, vol. 24, p. 68, Jan. 2024. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[2] M. U. Liaquat, H. S. Munawar, A. Rahman, Z. Qadir, A. Z. Kouzani, and M. A. P. Mahmud, "Localization of Sound Sources: A Systematic Review," *Energies*, vol. 14, p. 3910, Jan. 2021. Number: 13 Publisher: Multidisciplinary Digital Publishing Institute.

[3] K. Van Der Heijden and S. Mehrkanoon, "Goal-driven, neurobiological-inspired convolutional neural network models of human spatial hearing," *Neurocomputing*, vol. 470, pp. 432–442, Jan. 2022.

[4] H. Yu, "DOA AND EVENT GUIDANCE SYSTEM FOR SOUND EVENT LOCALIZATION AND DETECTION WITH SOURCE DISTANCE ESTIMATION," 2024.

[5] Q. Wang, Y. Dong, H. Hong, R. Wei, M. Hu, S. Cheng, Y. Jiang, M. Cai, X. Fang, and J. Du, "THE NERC-SLIP SYSTEM FOR SOUND EVENT LOCALIZATION AND DETECTION WITH SOURCE DISTANCE ESTIMATION OF DCASE 2024 CHALLENGE," 2024.

[6] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," May 2019. Issue: arXiv:1905.08546 arXiv:1905.08546.

[7] Q. Wang, Y. Jiang, S. Cheng, M. Hu, Z. Nian, P. Hu, Z. Liu, Y. Dong, M. Cai, J. Du, and C.-H. Lee, "THE NERC-SLIP SYSTEM FOR SOUND EVENT LOCALIZATION AND DETECTION OF DCASE2023 CHALLENGE," 2023.

[8] S. Kuang, J. Shi, K. v. d. Heijden, and S. Mehrkanoon, "BAST: Binaural Audio Spectrogram Transformer for Binaural Sound Localization," Aug. 2024. Issue: arXiv:2207.03927 arXiv:2207.03927.

[9] T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "SALSA: Spatial Cue-Augmented Log-Spectrogram Features for Polyphonic Sound Event Localization and Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1749–1762, 2022. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[10] T. N. Tho Nguyen, D. L. Jones, K. N. Watcharasupat, H. Phan, and W.-S. Gan, "SALSA-Lite: A Fast and Effective Feature for Polyphonic Sound Event Localization and Detection with Microphone Arrays," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 716–720, May 2022. ISSN: 2379-190X.

[11] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems*, vol. 96, pp. 184–210, Oct. 2017.

[12] P. Goli and S. van de Par, "Deep Learning-Based Speech Specific Source Localization by Using Binaural and Monaural Microphone Arrays in Hearing Aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1652–1666, 2023. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[13] M. Vorländer, *Auralization*. Springer, 2020.

[14] V. Aubanel, M. L. G. Lecumberri, and M. Cooke, "The sharvard corpus: A phonemically-balanced spanish sentence resource for audiology," *International journal of audiology*, vol. 53, no. 9, pp. 633–638, 2014.

[15] D. Schröder and M. Vorlaender, *RAVEN: A real-time framework for the auralization of interactive virtual environments*. Jan. 2011.

[16] F. Pausch, S. Doma, and J. Fels, "Ihta-indhartf–database of individual behind-the-ear hearing-aid-related transfer functions with high spatial resolution," *Institute for Hearing Technology and Acoustics, RWTH Aachen*, 2022.