DOI: 10.61782/fa.2025.0311

FORUM ACUSTICUM EURONOISE 2025

DOES SOURCE DIRECTIVITY INFLUENCE NEURAL NETWORK-BASED SPEECH POSITION ESTIMATION?

Rhoddy Viveros-Muñoz[1*] **Sebastián Guajardo**[1]
[1] Department of Electronics and Computer Sciences, Universidad Técnica Federico Santa María, Concepción, Chile

ABSTRACT

This study explores the ability of an artificial neural network (ANN) to estimate the position of speech sources—comprising both direction of arrival (DOA) and distance—by addressing the question: "Does the directivity of a speech source influence an ANN's position estimation?"
A comparative analysis was performed between omnidirectional and directional speech sources, with the latter simulating the natural directivity pattern of a human voice. Stimuli were generated using an acoustic virtual reality framework (RAVEN), encompassing multiple conditions: four reverberation times (0.1, 0.3, 0.6, and 0.9 seconds), 12 DOA angles (spaced at 30-degree intervals in azimuth), and two source types (omnidirectional vs. directional).
Findings from this analysis aim to shed light on the role of source directivity in shaping the performance of machine learning models for spatial audio tasks, such as sound source localization and enhancement in complex acoustic environments. The outcomes have promising implications for applications in spatial audio systems, hearing aids, and autonomous systems that rely on auditory information for navigation and interaction.

Keywords: *artificial neural network, direction of arrival, source directivity, acoustic virtual reality, source position estimation.*

[*] ***Corresponding author***: *rhoddy.viveros@usm.cl.*

1. INTRODUCTION

Sound source localization (SSL) refers to the process of estimating the spatial origin of an acoustic signal, specifically described in terms of direction of arrival (DoA) and distance relative to a listener or sensor array. Accurate SSL is critical for various applications, including surveillance, human-computer interaction, robotics, and auditory scene analysis [1-2]. In recent years, increasing attention has been paid to data-driven and learning-based approaches for solving SSL tasks, as traditional signal processing techniques such as time difference of arrival (TDOA), interaural level and time differences (ILD/ITD), and beamforming; often struggle in reverberant or noisy environments [3-4].
Machine learning (ML) methods, particularly deep neural networks (DNNs), have demonstrated considerable promise in overcoming the limitations of classical approaches by learning robust representations from complex acoustic scenes [5]. Interestingly, many of these models draw inspiration from human auditory processing mechanisms. For example, the use of mel-frequency cepstral coefficients (MFCCs) or mel-spectrograms as feature representations mimics the frequency selectivity of the human cochlea [6-7], embedding perceptual priors into the computational models.
Among the broad category of SSL tasks, the localization of speech sources is of particular importance, especially in the context of assistive hearing technologies. For hard-of-hearing individuals or users of hearing aids, the ability to spatially segregate and localize speech sources plays a critical role in auditory scene understanding and effective communication [8-10]. One key but often overlooked property of speech is its acoustic directivity, the way the voice radiates differently across directions depending on frequency and articulation. While traditional SSL models usually assume omnidirectional sources, human speech

exhibits strong directional patterns that can influence spatial cues, particularly under realistic reverberant conditions [11-12]. Despite its potential importance, the influence of source directivity on neural network-based SSL systems remains underexplored.

Therefore, the primary objective of this study is to investigate whether the directivity of speech sources has a measurable effect on the performance of deep neural network-based models for speech position estimation using binaural recordings. We aim to compare two types of spectral feature representations: (i) a Fourier-based spectrogram, which provides a general, physics-driven decomposition of the signal, and (ii) a mel-spectrogram, which mimics the frequency analysis performed by the human auditory system.

By training similar neural network architectures on both representations, we seek to isolate the role of biologically inspired features in learning to resolve spatial information, particularly under conditions that include speech signals with varying directivity patterns.

Understanding the interaction between acoustic directivity and feature representation is expected to provide theoretical insights into the interpretability of ML models for speech localization and inform future designs of assistive hearing technologies that operate in real-world acoustic environments.

2. METHOD

2.1 Dataset and Acoustic Simulation

To investigate the effect of speech source directivity on the performance of neural network-based position estimation, we used a dataset of speech recordings using a controlled simulation framework. The speech material consisted of 700 phonetically balanced sentences adapted from the SHARVARD corpus [13] into Chilean Spanish. These sentences were recorded by 20 native speakers (10 female), all of whom participated in a supervised recording session. The recordings were made in a sound-attenuated booth using a high-quality microphone at a sampling rate of 44.1 kHz.

Spatialization of the speech recordings was performed using the Room Acoustics for Virtual Environments (RAVEN) simulation framework [14], which allowed precise control over source characteristics, spatial configuration, and room acoustics. Two types of source directivity were modeled: an omnidirectional source, radiating energy uniformly in all directions, and a directional speech source, simulating the typical radiation pattern of the human voice. For the directional condition, two speaker orientations were simulated to reflect contrasting spatial alignments relative to the listener: (i) facing the listener, and (ii) facing directly away from the listener (i.e., 180° rotation). This manipulation aimed to capture the acoustic variability introduced by speech directivity and head orientation.

The directional speech source was modeled using a voice directivity pattern available in the RAVEN directivity database. Each selected utterance was rendered in multiple acoustic scenes, varying systematically across:

- Four reverberation times (T30): 0.1 s, 0.3 s, 0.6 s, and 0.9 s.
- Twelve DoA: 0°, 30°, 60°, 90°, 120°, 150°, 180°, 210°, 240°, 270°, 300°, and 330°.
- Three source conditions: omnidirectional, directional-facing, and directional-reversed.

Binaural renderings were generated using head-related transfer functions (HRTFs) obtained from the IHTA-indHARTF database, which contains high-resolution acoustic measurements from 30 adult participants recorded [15]. Their use ensured that the binaural stimuli reflected perceptually valid spatial information suitable for human-oriented localization models.

The final dataset comprised 144,000 binaural audio samples. The data were split into training (80%), validation (10%), and test (10%) sets.

2.2 Feature Extraction

To investigate the role of feature representation in speech source position estimation, we extracted two types of time-frequency features from the binaural audio signals: Short-Time Fourier Transform (STFT) spectrograms and mel-spectrograms.

For the STFT spectrograms, each audio signal was segmented using a Hamming window of 512 samples and a hop size of 256 samples. The STFT was independently applied to the left and right channels, yielding complex-valued spectrograms that were converted to magnitude representations. These were subsequently concatenated along the channel axis to form a two-channel input capturing the interaural spectral differences.

For the mel-spectrograms, a Hamming window of 1024 samples and a hop size of 512 samples were used to better align with perceptual frequency resolution. The resulting magnitude spectra were passed through a mel filterbank composed of 64 filters spaced according to the mel scale, approximating cochlear frequency selectivity. The resulting mel-spectrograms were log-compressed

(log-mel) to enhance dynamic range compression and facilitate neural network training.
Both representations were stored as input tensors of shape [C × F × T], where C = 2 (left and right channels), F corresponds to the number of frequency bins (129 for STFT, 64 for mel), and T denotes the number of time frames. This formulation allows for a consistent comparison of biologically inspired and physics-based spectral front-ends within the same deep learning framework.

2.3 Neural Network Architecture

We implemented a convolutional neural network (CNN) designed to estimate the position of speech sources—specifically, the DoA and distance—from binaural time-frequency representations. The network accepts as input either STFT spectrograms or mel-spectrograms, both formatted as two-channel tensors that encode spectral features from the left and right ears.
The architecture consists of the following components:

- **Input layer**: Accepts input tensors of shape [2×F×T], where F is the number of frequency bins, and T is the number of time frames, depending on the duration of the audio segment and the hop size used.
- **Convolutional blocks**: The feature extractor comprises four convolutional layers with increasing filter depths (64, 128, 256, 512), each followed by ReLU activation, 2D max-pooling, and dropout (rate = 0.1). These layers are responsible for learning hierarchical patterns across time and frequency.
- **Flattening and fully connected layers**: The output of the final convolutional block is flattened and passed through a series of dense layers with 64 and 32 hidden units, respectively. This stage maps the high-level feature representations to a compact encoding suitable for regression.
- **Output layer**: The final dense layer contains two units, corresponding to the predicted azimuth angle (in degrees) and distance (in meters).

The total number of trainable parameters was approximately 2 million, and the model was implemented using PyTorch version 2.4.1. To ensure a controlled comparison, both models were trained using the same neural network architecture and optimization procedure, although the input representations differed not only in perceptual scaling (mel vs. linear) but also in time-frequency resolution due to distinct window and hop size configurations.

2.4 Evaluation Metrics

The performance of the neural network models was evaluated using regression-based metrics for both spatial dimensions: DoA and distance. Given the continuous nature of these output variables, the Mean Absolute Error (MAE) was used as the evaluation metric on the test set. For both DoA and distance, MAE was computed as the average absolute difference between the predicted values and the ground truth. MAE in DoA was expressed in degrees, while MAE in distance was expressed in meters. This metric provides an intuitive measure of average estimation accuracy and is robust to outliers.
All metrics were computed independently for the STFT-based and mel-based models under each experimental condition. Performance was further analyzed by grouping the results according to source directivity (omnidirectional, directional-facing, directional-reversed) to assess whether speech radiation patterns affected localization accuracy.

3. RESULTS AND DISCUSSION

3.1 Effect of Source Directivity on Direction of Arrival (DoA) Estimation

The impact of source directivity on DoA estimation was analyzed for both STFT-based and mel-based models across the 12 tested azimuth angles. Figures 1 and 2 show polar plots of the Mean Absolute Error (MAE) in azimuth estimation as a function of angle, grouped by directivity condition (omnidirectional, directional-facing, and directional-reversed). Figure 1 corresponds to the mel-spectrogram-based model, while Figure 2 displays results from the STFT-based model.

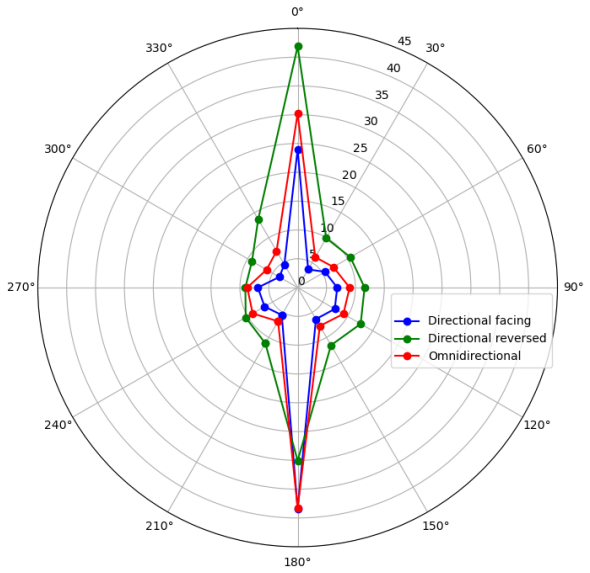

Figure 1: Mean Absolute Error (MAE) in DoA estimation for the mel-spectrogram-based model, plotted as a function of source azimuth angle. The polar plot shows results for three source directivity conditions: omnidirectional (red), directional-facing (blue), and directional-reversed (green).

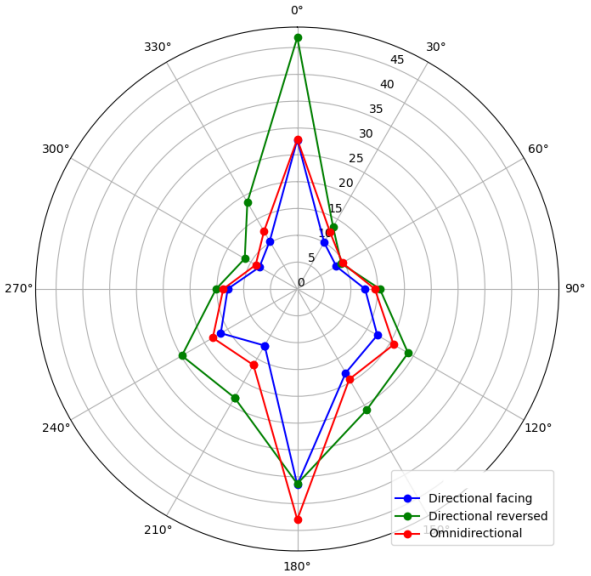

Figure 2: Mean Absolute Error (MAE) in DoA estimation for the STFT-based model, plotted as a function of source azimuth angle. The polar plot includes results for omnidirectional (red), directional-facing (blue), and directional-reversed (green) conditions.

For both feature types, source directivity strongly influenced DoA estimation accuracy. The lowest errors were consistently observed for the directional-facing condition, followed by the omnidirectional condition. In contrast, the directional-reversed configuration yielded the highest errors, especially at frontal (0°) and rear (180°) angles, suggesting a prominent front-back confusion effect.

When comparing the two feature representations, the mel-based model exhibited a more robust DoA estimation across all azimuth degrees and source directivity conditions. Specifically, mel features reduced angular MAE by up to 10 degrees in the most challenging reversed condition at 0° and 180°, relative to the STFT model. This suggests that the perceptually motivated mel scale better captures direction-dependent spectral variations, particularly under conditions of direct-path cues.

Overall, these results confirm that speech directivity patterns, especially when facing away from the listener, can impair azimuth estimation, and that mel-based models are more resilient to such degradations.

3.2 Effect of Source Directivity on Distance Estimation

Figures 3 and 4 show the Mean Absolute Error (MAE) in distance estimation as a function of distance bins (ranging from 1.0 to 4.0 meters), grouped by source directivity condition. Figure 3 presents results for the mel-spectrogram-based model, while Figure 4 displays results for the STFT-based model.

Both the STFT-based and mel-based models demonstrated clear performance degradation with increasing distance, consistent with the expected weakening of spatial cues at longer ranges.

In terms of directivity effects, the directional-facing condition yielded the most consistent and lowest distance errors across all bins, for both models. The omnidirectional source, in contrast, showed increasing error with distance, particularly above 2.5 meters, likely due to more uniform spatial energy distribution and reduced spectral asymmetries. The directional-reversed condition exhibited variable performance: at close distances, its error was relatively low, but it increased sharply beyond 3 meters, especially for the STFT model.

Again, mel-spectrograms provided superior distance estimation, with lower overall MAE in all distance bins and for all directivity types. The advantage of mel features was particularly pronounced for far-field estimation (>3 m), where STFT-based predictions showed significantly larger variance and error, especially for omnidirectional sources.

These findings reinforce the idea that biologically inspired spectral representations enhance robustness in challenging

acoustic conditions, especially when direct-path information is weak or distorted by reverberation and source radiation patterns.

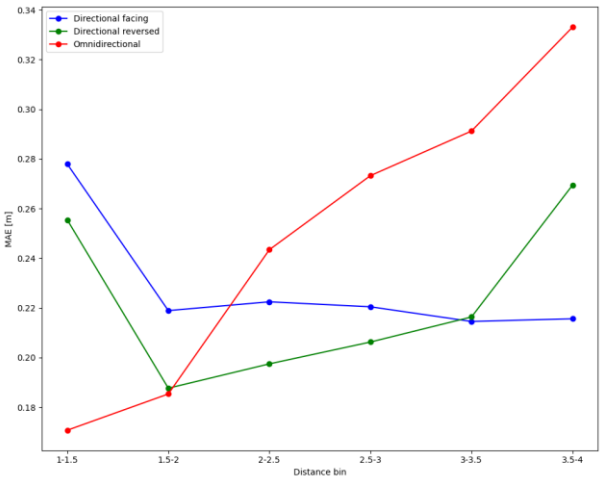

Figure 3: Mean Absolute Error (MAE) in distance estimation for the mel-spectrogram-based model, shown across distance bins from 1.0 to 4.0 meters. Each curve represents a different source directivity condition: omnidirectional (red), directional-facing (blue), and directional-reversed (green).

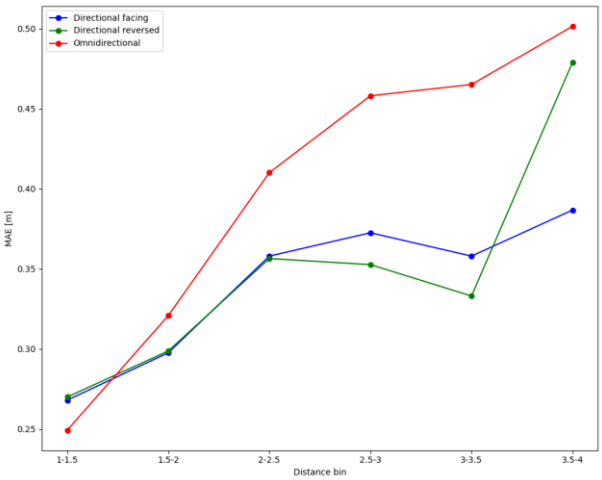

Figure 4: Mean Absolute Error (MAE) in distance estimation for the STFT-based model, plotted across distance bins from 1.0 to 4.0 meters. Results are shown for three source directivity conditions: omnidirectional (red), directional-facing (blue), and directional-reversed (green).

4. CONCLUSION

This study investigated whether the directivity of speech sources affects the performance of a deep neural network in estimating source position from binaural audio. The findings provide a clear affirmative answer to the central question: yes, source directivity significantly influences neural network-based speech position estimation.

Across both spatial dimensions, DoA and distance, models trained on signals from directional sources yielded different levels of accuracy depending on the orientation of the speaker. The directional-reversed condition, in which the source was facing away from the listener, consistently led to higher estimation errors, especially under reverberant conditions. This demonstrates that the directionality of speech radiation alters the availability and quality of spatial cues, which are crucial for both angular and distance localization.

Furthermore, models trained with mel-spectrogram features showed greater robustness to variations in source directivity and acoustic conditions compared to those using STFT-based representations. This suggests that biologically inspired, perceptually motivated features may better capture the spatial information encoded in directional speech signals.

Overall, the results emphasize that source directivity should not be neglected in the development and evaluation of spatial hearing models. Accounting for realistic radiation patterns is essential for building deep learning systems capable of generalizing complex and ecologically valid environments. These insights are particularly relevant for applications such as assistive hearing devices, speech enhancement systems, and auditory scene analysis in autonomous agents.

While the use of simulated environments allowed precise control over acoustic and spatial parameters, future work should explore real-world recordings and dynamic source behaviors—such as head movements—to further enhance model generalizability and ecological validity.

5. REFERENCES

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

[2] D. Desai and N. Mehendale, 'A Review on Sound Source Localization Systems', *Arch. Comput. Methods Eng.*, vol. 29, no. 7, pp. 4631–4642, Nov. 2022, doi: 10.1007/s11831-022-09747-2.

[3] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, 'Acoustic echoes reveal room shape', *Proc. Natl. Acad. Sci.*, vol. 110, no. 30, pp.

12186–12191, Jul. 2013, doi: 10.1073/pnas.1221464110.
[4] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, 'A survey of sound source localization with deep learning methods', *J. Acoust. Soc. Am.*, vol. 152, no. 1, pp. 107–151, Jul. 2022, doi: 10.1121/10.0011809.
[5] A. Carlini, C. Bordeau, and M. Ambard, 'Auditory localization: a comprehensive practical review', *Front. Psychol.*, vol. 15, p. 1408073, Jul. 2024, doi: 10.3389/fpsyg.2024.1408073.
[6] M. Slaney, 'Auditory toolbox: A MATLAB toolbox for auditory modeling work'. Interval Research Corporation, Tech. Rep, 1998.
[7] Y. Tokozume and T. Harada, 'Learning environmental sounds with end-to-end convolutional neural network', in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA: IEEE, Mar. 2017, pp. 2721–2725. doi: 10.1109/ICASSP.2017.7952651.
[8] V. Best, S. Carlile, C. Jin, and A. Van Schaik, 'The role of high frequencies in speech localization', *J. Acoust. Soc. Am.*, vol. 118, no. 1, pp. 353–363, Jul. 2005, doi: 10.1121/1.1926107.
[9] J. Zaar and L. H. Carney, 'Predicting speech intelligibility in hearing-impaired listeners using a physiologically inspired auditory model', *Hear. Res.*, vol. 426, p. 108553, Dec. 2022, doi: 10.1016/j.heares.2022.108553.
[10] A. Francl and J. H. McDermott, 'Deep neural network models of sound localization reveal how perception is adapted to real-world environments', *Nat. Hum. Behav.*, vol. 6, no. 1, pp. 111–133, 2022.
[11] B. B. Monson, E. J. Hunter, A. J. Lotto, and B. H. Story, 'The perceptual significance of high-frequency energy in the human voice', *Front. Psychol.*, vol. 5, Jun. 2014, doi: 10.3389/fpsyg.2014.00587.
[12] A. Ihlefeld and B. G. Shinn-Cunningham, 'Effect of source spectrum on sound localization in an everyday reverberant room', *J. Acoust. Soc. Am.*, vol. 130, no. 1, pp. 324–333, Jul. 2011, doi: 10.1121/1.3596476.
[13] V. Aubanel, M. L. G. Lecumberri, and M. Cooke, 'The Sharvard Corpus: A phonemically-balanced Spanish sentence resource for audiology', *Int. J. Audiol.*, vol. 53, no. 9, pp. 633–638, Sep. 2014, doi: 10.3109/14992027.2014.907507.
[14] D. Schröder and M. Vorländer, 'RAVEN: A real-time framework for the auralization of interactive virtual environments', *Forum Acusticum*, pp. 1541–1546, 2011.
[15] F. Pausch, S. Doma, and J. Fels, 'Hybrid multi-harmonic model for the prediction of interaural time differences in individual behind-the-ear hearing-aid-related transfer functions', *Acta Acust.*, vol. 6, p. 34, 2022, doi: 10.1051/aacus/2022020.