forum acusticum 2026
8-12 September · Graz, Austria

DYNAMIC VOICE DIRECTIVITY PREDICTION FROM FACIAL GEOMETRY VIA FEED-FORWARD NEURAL NETWORKS

Yewon Kim[1], Samuel D. Bellows[1], Brian F. G. Katz[2]

[1] *Department of Electrical and Computer Engineering, University of Utah Asia Campus, Republic of Korea*
[2] *Institute Jean le Rond d'Alembert, Sorbonne Université, Paris, France*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

Quantifying voice radiation is essential for applications such as telecommunications, room acoustic modeling, and virtual reality. However, the precise directional characteristics of speech can be complex due to phoneme-dependent variations and movements of the head and torso. This work proposes a feed-forward neural network to predict dynamic voice directivity patterns in the horizontal plane based on extracted facial features from a simultaneous video capture of a singer's face. A static time and phoneme-averaged directivity pattern serves as the baseline for evaluating the model's performance. While the model improves predictions across the majority of frequency bands between 80 Hz and 8 kHz relative to the baseline, the neural network performed best in the mid-frequency bands. The results demonstrate that facial landmarks such as the mouth aperture surface area can provide relevant information in predicting voice radiation characteristics, providing an approach for rendering real-time, dynamic voice directivity.

Keywords: *voice directivity, multi-modal signal processing, computer vision, facial feature extraction*

1. Introduction

Decades of research has analyzed the characteristics of sound radiation from the human voice [1]–[4]. Voice directivity plays an important role in telecommunications [5], architectural acoustics and room acoustic modeling [6], [7], microphone placement [8], and virtual reality. Investigations have shown that voice directivity varies between individuals [9], [10], between phonemes [11]–[13], and with head orientation [14]. The latter two dependencies highlight that voice radiation is a dynamic process, varying both as a function of mouth shape and head pose. While many studies have quantified voice directivity in terms of fixed phonemes or fixed head orientations and postures, how to dynamically model these effects in real-time has received less attention.

Several studies have aimed to link known facial features, such as mouth size, to voice radiation characteristics. For example, theoretical models based on spheres [15] or spheroids [16] provide a direct link between mouth aperture size and radiation patterns. While these models provide a reasonable approximation even to higher frequencies [10], they fail to capture the finer features of voice radiation, in particular due to the non-spherical nature of the human body, including diffraction around the human body and scattering from the torso and legs [14], [17], [18]. In addition, while the models can apply varying mouth aperture areas to imitate phoneme-dependent effects, they are limited in approximating the mouth using simple geometric primitives which deviate from real mouth shapes. Although numerical methods, such as boundary or finite element methods, could conceivably provide realistic predictions based on a known human geometry, the computational expense is too high for real-time applications. Consequently, there remains a need to model dynamic voice directivity using models which are simple enough to be implemented in real-time while maintaining enough complexity to capture the inherent variability in human voice radiation characteristics.

This work proposes applying a feed-forward neural network (FNN) to predict voice radiation patterns based on extracted facial features from a simultaneous video capture. Video-frame-level predictions based on facial geometry provide a straightforward method to build upon theoretical models by incorporating more complex features beyond mouth aperture size alone. Results demonstrate that training the model using extracted facial features such as the mouth area and mouth length-width ratio improve dynamic predictions of voice directivity compared to a static, time-

***Corresponding author**: samuel.bellows11@gmail.com.*

averaged approximation.

2. Methods

2.1. Data Acquisition and Processing

Time-varying voice directivity measurements from [11] served as the experimental data for the study. The data incorporates three simultaneously captured video and audio recordings of a single male singer in an anechoic chamber. The audio recordings provide sound levels at 24 equally spaced angular positions along a 180° arc in the horizontal plane. The video acquisition included high-contrast lip makeup on the singer to improve lip-boundary visibility. Further details on the measurement set-up and data acquisition are available in [11].

The video processing pipeline began by passing each video frame through Google's Mediapipe FaceMesh [19] to extract a mesh composed of 478 facial landmarks. Derived quantities computed from a subset of these facial landmarks served as the input features for the FNN. Theoretical models, such as a radially vibrating cap on a sphere, make use of a region with a fixed surface area to model the human mouth. Consequently, a natural extension to real data is to extract the total mouth aperture size of each frame A_n as an input model feature. This quantity was calculated by using the twenty landmarks forming the inner lip contour and using the shoelace formula [20] to calculate the associated area. A second feature derived from the mouth is the ratio of the width and height for each frame R_n, calculated in each frame by using the four landmarks at the uppermost, lowermost, and sides of the inner lip contour. This feature provides information on the dynamic shape of the mouth, which is ignored in the perfectly symmetric spherical cap models used to approximate speech radiation.

The time-varying directivities were calculated by passing the mth microphone signal $y_m(t)$ through a fourth-order 1/3-octave filter bank to produce the broad band signals $y_{m,b}(t)$, where b indicates the band number. The bands considered in this work were from 80 Hz to 8 kHz. Next, after synchronizing the individually captured audio and video data to a clapperboard strike, centered time windows of approximately 370 ms about the video time served as the boundaries for calculating the RMS signal energy associated for the nth frame, mth microphone, and bth band $y^{(RMS)}_{n,m,b}$. Calibrations between individual microphone sensitives followed by normalizing the data to the 125 Hz band energy of an omnidirectional source placed within the array and thus assumed a flat microphone response over the desired bandwidth.

After extracting the input facial features and the output directivity data, a simple thresholding procedure eliminated any unusable frames due to either little acoustic energy or where the mouth was closed. A total of 1146 frames remained after combining all valid instances across the three recordings. Figure 1 demonstrates the extracted facial landmarks and directivity plots for select frames and 1/3-octave bands. At some frequencies, such as 200 Hz, the radiation pattern appears primarily independent of mouth geometry. However, at higher frequencies, individual mouth shapes due to varying loudness and phonemes lead to differing directivity patterns [21].

2.2. Feed-Forward Neural Network

A feed-forward neural network (FNN) provided a nonlinear mapping between the extracted facial features and the measured radiated patterns. The input data was the two features of mouth area S_n and the shape parameter R_n. From the facial features, there are several possible methods to calculate the dynamic directivity of the source, such as directly predicting the frame RMS quantities $y^{(RMS)}_{n,m,b}$. However, separating the frame RMS quantity into a time-averaged (and consequently phoneme-averaged) value and dynamic value yielded better results. First, the time-averaged RMS quantities for a given microphone and band follow by taking the mean value across all frames as

$$\bar{y}^{(RMS)}_{m,b} = \frac{1}{N}\sum_{n=1}^{N} y^{(RMS)}_{n,m,b} \tag{1}$$

where $N = 1146$ is the total number of frames. For a given frame, the residual variation is

$$\tilde{y}^{(RMS)}_{n,m,b} = y^{(RMS)}_{n,m,b} - \bar{y}^{(RMS)}_{m,b}. \tag{2}$$

This value represents the variation in the radiation pattern from the mean value. The output prediction of the FNN are these twenty-four residual variations for each array microphone.

Each 1/3-octave band utilized a separate FNN. The FFNs consisted of three hidden layers with 64 and 32 units, respectively, each connected by a ReLU activation function (Fig. 2). The model was trained using the Adam optimizer with ten-fold cross validation. For a given frame and band, the baseline error is the RMSE error between the measured and mean directivity pattern:

$$\mathrm{RMSE}^{(base)}_{n,b} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\bar{y}^{(RMS)}_{m,b} - y^{(RMS)}_{n,m,b}\right)^2}, \tag{3}$$

where $M = 24$. This value represents the error occurring when one neglects the dynamic nature of voice radiation. The error of the FNN, $\mathrm{RMSE}^{(\mathrm{FNN})}_{n,b}$, is computed identically with $\mathrm{RMSE}^{(base)}_{n,b}$ but with the FNN prediction in place of the time-averaged values $\bar{y}^{(RMS)}_{m,b}$. In addition, the baseline and FNN RMSE values may be averaged across the total number of frames to provide a singular metric for the entire band as

$$\overline{\mathrm{RMSE}}^{(\mathrm{base,FNN})}_{b} = \frac{1}{N}\sum_{n=1}^{N}\mathrm{RMSE}^{(\mathrm{base,FNN})}_{n,b}. \tag{4}$$

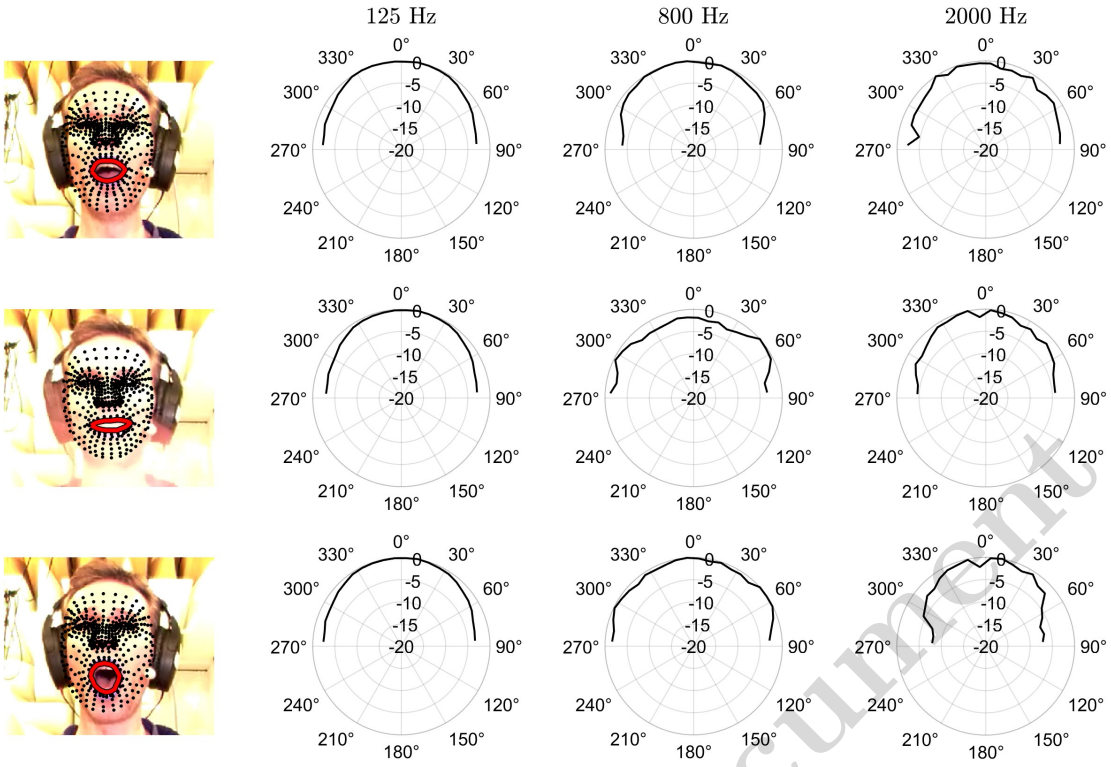

Figure 1: Sample video frames with extracted facial landmarks and associated horizontal plane directivity patterns for the 200 Hz, 800 Hz, and 2000 Hz 1/3-octave bands on a 20 dB scale.

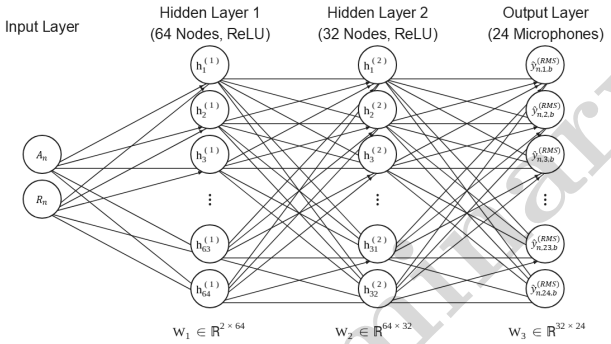

Figure 2: Feed-forward neural network architecture used in predicting horizontal-plane directivities.

The improvement of the FNN relative to the baseline for a given band follows on a logarithmic scale as

$$\Delta_b = 20 \log_{10} \left(\frac{\overline{\mathrm{RMSE}}_b^{(\mathrm{FNN})}}{\overline{\mathrm{RMSE}}_b^{(\mathrm{base})}} \right). \tag{5}$$

Comparing the baseline and FNN prediction values for these metrics provides validation of the proposed method.

3. Results

Table 1 summarizes the RMSE values averaged for each band. The baseline RMSE across frequency bands provides insight into frequency-dependent variations between averaged and dynamic voice radiation. The bands between 250 Hz and 4 kHz exhibited the highest variations between the static assumption and the dynamic reality, exceeding an RMSE of 10% in four of these bands. In contrast, the lowest and highest frequency bands tended to have lower RMSE.

These variation may be due to several physical reasons. At low frequencies, the acoustic wavelength is much larger than the mouth aperture area, leading to monopolar radiation independent of any specific mouth geometry [22]. This result is consistent with directivity plots in Fig. 1 for the 125 Hz 1/3-octave band, which showed little variation in the voice radiation pattern despite differences in the facial features. At high frequencies, the radiation becomes more specular and concentrated in front of the talker, which could likewise lead to less variation between phonemes and mouth shapes. However, intermediate frequencies, where the acoustic wavelength is on the order of the mouth size, can exhibit high variations between radiation patterns between phonemes [11], [21].

Similar to the baseline RMSE, the performance of the FNN varied across bands. First, for frequency bands less than 200 Hz, the neural network performed only marginally better than the static baseline. This result may be linked to the little variability between radiation patterns at these frequencies. The best performing frequency bands of the FNN fell between the 400 Hz and 4 kHz 1/3-octave bands, peaking at the 800 Hz band. In this band, the FNN decreased the RMSE by about 3% relative to the baseline, repre-

Band (Hz)	Baseline (%)	FNN (%)	Δ_b (dB)
80	5.46	5.41	-0.1
100	4.57	4.53	-0.1
125	7.28	7.13	-0.2
160	8.82	8.57	-0.3
200	7.48	7.33	-0.2
250	10.22	9.81	-0.3
315	14.41	14.35	0.0
400	11.98	11.28	-0.5
500	5.35	5.21	-0.2
630	3.04	2.57	-1.4
800	11.07	8.24	-2.6
1000	5.23	4.35	-1.6
1250	8.08	7.27	-0.9
1600	5.34	5.11	-0.4
2000	5.15	4.90	-0.5
2500	6.07	5.42	-1.0
3150	8.80	7.85	-1.0
4000	4.19	3.91	-0.6
5000	1.17	1.16	-0.1
6300	1.57	1.56	-0.1
8000	1.93	1.90	-0.1
Average	6.54	6.09	-0.6

Table 1: RMSE averaged over the ten-fold cross validation for the static baseline and the neural network. Negative improvement Δ_b indicates that the neural network reduced the error relative to the baseline. The average row reports the mean RMSE across all bands.

senting a 2.6 dB reduction. As noted in the baseline RMSE, these bands exhibit high variability in directivity patterns, which would allow the FNN to determine stronger correlations between radiation trends and facial features.

Above 4 kHz, the improvement of the dynamically predicted directivity relative to the baseline was marginal. Several factors could lead to this result, including more detailed and intricate radiation patterns [10] that would require a model with more complexity, spatial aliasing errors due to insufficient sampling positions [23], or increased measurement noise and uncertainty.

Averaged over all bands, the neural network achieved a mean RMSE of 6.09%, compared with 6.54% for the static baseline. This corresponds to an overall relative improvement of 0.6 dB when using the FNN to dynamically predict the radiation pattern.

4. Discussion

Predicting dynamic voice radiation patterns for real-time applications presents several practical challenges, including accounting for phoneme-dependent variations. This work explored the effectiveness of a FNN to dynamically predict voice radiation patterns from facial features alone as a pilot investigation into the feasibility of the approach. A contrasting first theoretical approach could consider using table look-ups by extracting phoneme-dependent information from acoustic signals alone using automatic speech recognition (ASR) systems. However, the latency associated with buffering sufficient audio samples could be too prohibitive for real-time constraints. For example, streaming ASR systems typically must buffer audio in chunks on the order of 100–300 ms to produce a reliable estimate [24], whereas lightweight video-based facial landmark extractors such as MediaPipe FaceMesh have been reported to run with per-frame latencies of only a few milliseconds [25]. Since the trained FNN itself required well under 1 ms per inference, the practical bottleneck for the video-based approach is expected to be the camera frame rate rather than model inference time: standard video cameras operate at roughly 33 ms per frame (30 fps), whereas dedicated high-speed cameras can reach frame rates of several hundred to several thousand frames per second, corresponding to intervals on the order of 1 ms or less. Consequently, an image-based system could provide results faster than an ASR-based system, but at the additional cost of more needing both expensive equipment and significantly higher processing power.

The results of this work suggest that while there is potential to develop real-time voice directivity predictions from a video capture alone, there are also several limitations and difficulties. First, the FNN did improve predictions relative to the baseline in many frequency bands, in particular at mid-frequencies exhibiting high inter-phoneme variations. This result suggests that the FNN is able to learn relevant correlations between facial features and radiation characteristics. However, the FNN was less effective at lower and higher frequency bands, suggesting that mouth geometry alone may be insufficient to predict these variations. In fact, several studies have shown that the head, torso, and legs have an impact on voice directivity [14], [17], [18], indicating that taking further geometric features of the body into account could improve the model's accuracy. Additionally, a model using mouth shape alone neglects radiation from the nose and face, which is important for certain phonemes such as /n/ or /m/, leading to further errors.

Lastly, the results of this work employed a single male talker with radiation patterns measured at 24 fixed locations. Because the input feature extraction uses a pre-trained image-processing model (MediaPipe), the model can readily be applied to other video inputs, with the limitation that the output directivity will be that of the single subject. Generalizing the radiation results to more diverse talkers would require training the model on more synchronized video-audio data sets.

Improvements to the model could include deriving other features from video data to further decrease errors. In addition, a multi-modal signal pipeline utilizing both video and audio data could improve results. Other variations could incorporate a multi-band model or time-dependence between frames to better exploit correlations between the input features and output radiation patterns.

5. Conclusions

This work developed a feed-forward neural network to predict dynamic voice directivity patterns in the horizontal plane from facial features extracted from a video input. The model, which predicts RMS pressure variations from a phoneme-averaged radiation pattern using the mouth aperture area and mouth length-width ratio, improved upon the static baseline in the majority of the twenty-one 1/3-octave bands investigated. The best results occurred in the mid-frequency region characteristic of high inter-phoneme variations, with improvements exceeding 2.0 dB. While the FNN did show notable improvements in these frequency regimes, the effectiveness was less pronounced at low and high frequencies, with improvements of less than 0.5 dB typical. Future work includes incorporating other features, either acoustic or visual, which can decrease prediction errors across all frequencies and exploring other deep-learning architectures. In addition, the results should be extended to full 360° 2D patterns as well as 3D spherical data. Future work should also include training on a multi-speaker dataset to evaluate and improve the model's cross-speaker generalization.

6. Acknowledgments

This work was supported by funding from the Undergraduate Research Opportunities Program at the University of Utah awarded to Yewon Kim.

References

[1] H. Dunn and D. W. Farnsworth, "Exploration of pressure field around the human head during speech," *J. Acoust. Soc. Am.*, vol. 10, no. 3, pp. 184–199, Jan. 1939. DOI: https://doi.org/10.1121/1.1915975.

[2] A. Moreno and J. Pretzschner, "Human head directivity in speech emission: A new approach," *Acoustics Letters*, vol. 1, pp. 78–84, 1978.

[3] G. Studebaker, "Directivity of the human vocal source in the horizontal plane," *Ear and Hearing*, vol. 6, no. 6, pp. 315–319, 1985.

[4] A. H. Meyer and J. Meyer, "The directivity and auditory impressions of singers," *Acustica*, vol. 58, pp. 130–140, 1985.

[5] T. Halkosaari, M. Vaalgamaa, and M. Karjalainen, "Directivity of artificial and human speech," *J. Audio Eng. Soc.*, vol. 53, no. 7/8, pp. 620–631, 2005.

[6] D. Cabrera, P. J. Davis, and A. Connolly, "Long-term horizontal vocal directivity of opera singers: Effects of singing projection and acoustic environment," *J. Voice*, vol. 25, no. 6, pp. 291–303, 2011.

[7] F. Bozzoli, M. Viktorovitch, and A. Farina, "Balloons of directivity of real and artificial mouth used in determining speech transmission index," in *Audio Engineering Society 118*, Barcelona, Spain, 2005.

[8] S. D. Bellows and T. W. Leishman, "Optimal microphone placement for single-channel sound-power spectrum estimation and reverberation effects," *J. Audio Eng. Soc*, vol. 71, no. 1/2, pp. 20–33, 2023. DOI: https://doi.org/10.17743/jaes.2022.0052.

[9] W. T. Chu and A. C. Warnock, "Detailed directivity of sound fields around the human head during speech," *National Research Council Canada*, vol. IRC-RR-104, pp. 1–47, Jan. 2002. DOI: https://doi.org/10.4224/20378930.

[10] T. W. Leishman, S. D. Bellows, C. M. Pincock, and J. K. Whiting, "High-resolution spherical directivity of live speech from a multiple-capture transfer function method," *J. Acoust. Soc. Am.*, vol. 149, no. 3, pp. 1507–1523, Mar. 2021. DOI: 10.1121/10.0003363.

[11] B. F. Katz and C. D'Alessandro, "Directivity measurements of the singing voice," in *Proceedings of the 19th International Congress on Acoustics*, Madrid, 2007.

[12] B. B. Monson, E. J. Hunter, and B. H. Story, "Horizontal directivity of low- and high-frequency energy in speech and singing," *J. Acoust. Soc. Am.*, vol. 132, pp. 433–441, 2012.

[13] C. Pörschmann and J. M. Arend, "Investigating phoneme-dependencies of spherical voice directivity patterns," *J. Acoust. Soc. Am.*, vol. 149, no. 6, pp. 4553–4564, 2021. DOI: 10.1121/10.0005401. eprint: https://doi.org/10.1121/10.0005401. [Online]. Available: https://doi.org/10.1121/10.0005401.

[14] S. Bellows and T. W. Leishman, "Effect of Head Orientation on Speech Directivity," in *Proc. Interspeech 2022*, Sep. 2022, pp. 246–250. DOI: 10.21437/Interspeech.2022-553.

[15] J. L. Flanagan, "Analog measurements of sound radiation from the mouth," *J. Acoust. Soc. Am.*, vol. 32, no. 12, pp. 1613–1620, Dec. 1960. DOI: 10.1121/1.1907972.

[16] K. Sugiyama and H. Irii, "Comparison of the sound pressure radiation from a prolate spheroid and the human mouth," *Acustica*, vol. 73, pp. 271–276, 1991.

[17] M. Brandner, R. Blandin, M. Frank, and A. Sontacchi, "A pilot study on the influence of mouth configuration and torso on singing voice directivity," *J. Acoust. Soc. Am.*, vol. 148, no. 3, pp. 1169–1180, Sep. 2020. DOI: 10.1121/10.0001736. eprint: https://doi.org/10.1121/10.0001736. [Online]. Available: https://doi.org/10.1121/10.0001736.

[18] S. D. Bellows, B. F. G. Katz, and T. W. Leishman, "Impact of leg diffraction and scattering on voice directivity patterns," in *Proceedings of the 11th Convention of the European Acoustics Association Forum Acusticum/EuroNoise 2025*, Malaga, Spain, 2025, pp. 4473–4480. DOI: 10.61782/fa.2025.0448.

[19] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time facial surface geometry from monocular video on mobile gpus," in *CVPR Workshop on Computer Vision for Augmented and Virtual Reality 2019*, Long Beach, CA, 2019. [Online]. Available: https://arxiv.org/abs/1907.06724.

[20] Y. Lee and W. Lim, "Shoelace formula: Connecting the area of a polygon with vector cross product," *The Mathematics Teacher*, vol. 110, no. 8, pp. 631–636, 2017, ISSN: 00255769. Accessed: Mar. 27, 2026. [Online]. Available: http://www.jstor.org/stable/10.5951/mathteacher.110.8.0631.

[21] S. D. Bellows and B. F. G. Katz, "Combining multiple sparse measurements with reference data regularization to create spherical directivities applied to voice data," *Acta Acustica*, vol. 8, no. 14, pp. 1–12, Mar. 2024. DOI: 10.1051/aacus/2024006.

[22] P. M. Morse and K. U. Ingard, *Theoretical Acoustics*. Princeton: Princeton University Press, 1968.

[23] S. D. Bellows and T. W. Leishman, "A spherical-harmonic-based framework for spatial sampling considerations of musical instrument and voice directivity measurements," in *Proceedings of Forum Acusticum*, Turin, Italy, Sep. 2023, pp. 4747–4754. DOI: https://doi.org/10.61782/fa.2023.0427.

[24] W. Kang *et al.*, "Delay-penalized transducer for low-latency streaming ASR," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[25] H. Chen, R. Mira, S. Petridis, and M. Pantic, "RT-LA-VocE: Real-time low-SNR audio-visual speech enhancement," in *Proceedings of Interspeech 2024*, 2024, pp. 3230–3234.