LEARNING SPEAKER-LISTENER MUTUAL HEAD ORIENTATION BY LEVERAGING HRTF AND VOICE DIRECTIVITY ON HEADPHONES

Harshvardhan Takawale, Nirupam Roy
University of Maryland College Park

ABSTRACT

Estimation of a speaker's direction and head orientation with binaural recordings can be a critical piece of information in many real-world applications with emerging 'earable' devices, including smart headphones and AR/VR headsets. However, it requires predicting the mutual head orientations of both the speaker and the listener, which is challenging in practice. This paper presents a system for jointly predicting speaker-listener head orientations by leveraging inherent human voice directivity and listener's head-related transfer function (HRTF) as perceived by the ear-mounted microphones on the listener. We propose a convolution neural network model that, given binaural speech recording, can predict the orientation of both speaker and listener with respect to the line joining the two. The system builds on the core observation that the recordings from the left and right ears are differentially affected by the voice directivity as well as the HRTF. We also incorporate the fact that voice is more directional at higher frequencies compared to lower frequencies. Our proposed system achieves 2.5° 90th percentile error in the listener's head orientation and 12.5° 90th percentile error for that of the speaker.

***Index Terms*—** Voice directivity, HRTF, head orientation, voiced sounds, auditory perception

1. INTRODUCTION

It is long known that human voice and hearing are directional in nature [1, 2]. Thus, when humans speak or hear, the sounds contain cues specific to the orientation of the human head. This enables our brains the natural ability to localize sound sources and sense when they are being talked to. Thus, this paper aims to explore whether binaurally recorded human speech contains these cues necessary to estimate the head orientation of not just the listener but also the speaker. The intuition behind the idea is that the sound travels through different channels before reaching each ear. So each of the recorded signals should encode some information regarding both the listener's and the speaker's head orientations.

Our proposed system leverages the directivity of human voice and hearing at different frequencies. We show that these effects are enhanced in the near-field. The differences in interaural recordings are used to infer the orientations. As the human voice directivity pattern (VDP) is dynamic [3], we first identify the segments that have higher directivity based on the harmonic structure [4] present in these regions and mask rest of the signal to suppress noise. We then extract the crucial features like interaural level difference (ILD), interaural time difference (ITD), and differences in the energy of frequency bins[5] and use them to train a deep neural network.

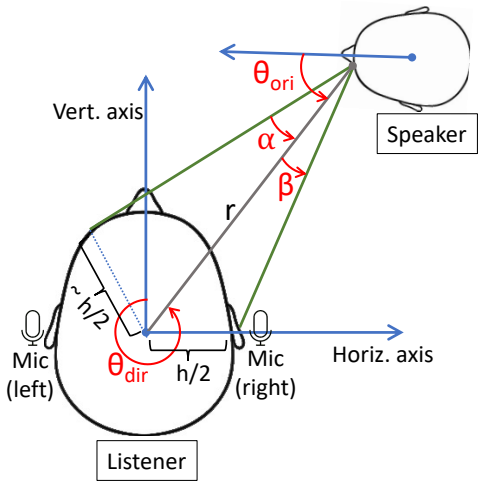

Fig. 1. Near field speaker-listener model overview.

We evaluate our system by creating a dataset that combines a real-world VDP dataset, a real-world HRTF dataset, and a real-world speech dataset. The evaluations show that the system has 2.5° 90th percentile error in detecting listener head orientations and 12.5° 90th percentile error in detecting speaker head orientations. Knowing these head orientations opens up context-aware applications in smart-home environments [6, 7], AR/VR [8], and Human-robot interaction [9].

2. PROBLEM FORMULATION AND INTUITIONS

We aim to find the direction of a speaker as well as her head orientation on the horizontal plane in an egocentric reference frame centered on the listener. The horizontal axis of this reference frame is along the line passing through the listener's ears and the perpendicular axis is along the listener's facing direction, as shown in Figure 1. We now consider a line joining the centers of the listener and the speaker's heads. The angle this line makes with the vertical axis is defined as the speaker's direction θ_{dir}. For the speaker's head orientation, we consider the angle θ_{ori} between the speaker's facing direction and the line joining the listener's center of the head and the speaker's mouth. Note that θ_{dir} and θ_{ori} can vary in-

dependently of each other. The goal is to estimate θ_{dir} and θ_{ori} simultaneously using only the binaural recording of natural speech from a pair of ear-mounted synchronized microphones on the listener.

Intuition: Humans have an excellent capability to perceptually find sounds' direction of arrival which is largely attributed to unique directional filtering by the listener's head, outer ears, and torso leading to differential binaural reception in the left and right ears. This effect is known as Head-Related Transfer Function (HRTF) which leads to spatial hearing cues. A large body of works leveraged HRTF to computationally estimate the sound's direction [10, 11]. However, as a function of a person's individual physiological factors, the HRTFs do not generalize across users and it limits the achievable accuracy in direction estimation. Binaural localization methods with body-worn microphones often assume prior knowledge of the listener's personal HRTF data to improve estimation accuracy [12]. Moreover, existing work assumes HRTF is the only factor that alters the received sound ignoring the impacts from the orientation of the speaker's head. Existing measurements show that a human speaker does not radiate voice sounds uniformly in all directions, rather the sound shows a frequency-dependent radiation pattern due to the diffraction with the head. This leads to the VDP that shows higher frequencies are attenuated more than lower frequency sounds toward the back of the head, which adds a cue for speaker's head orientation with respect to the listener. Therefore, the binaural recording in the listener's ear-worn microphones manifests a combined effect of the speaker's VDP and listener's HRTF leading to a unique opportunity to estimate both the direction and orientation of the speaker (θ_{dir} and θ_{ori}).

Several recent works leveraged VDP to estimate the speaker's head orientation using standalone microphone arrays[13]. However, due to the lateral symmetry, the VDP-based head orientation estimation performs poorly with several tens of degrees of error [14, 15]. In this paper, we hypothesize that (a) it is possible to jointly estimate speakers' direction (θ_{dir}) and orientation (θ_{ori}) from a single binaural recording and (b) this joint parameter estimation can enhance the accuracy of both of these angles. When we verify the diversity of HRTF and VDP across different angles separately using the correlation matrix shown in Figure 2(a) and 2(b), the VDP shows significant confusion. However, a correlation matrix with combined features, as shown in Figure 2(c), improves overall diversity indicating a possibility of joint estimation.

Challenges: We consider social conversational distance to be the physical scale where the speaker-listener orientation estimation can be applied. Therefore we consider a near field region of sound for our model, which is between 0.5 to 1.5 meters in range. However, a majority of the available HRTF datasets are collected at far-field. Such datasets do not capture the substantial difference of interaural distances from the source in the near-field and also the diffraction and shadowing effects on nearby sound sources. Moreover, the distance and angle between the centers of the speaker/listener heads differ significantly from those of ears in near-field. This requires us to transform existing HRTF datasets for our near-field scenarios using a model elaborated next.

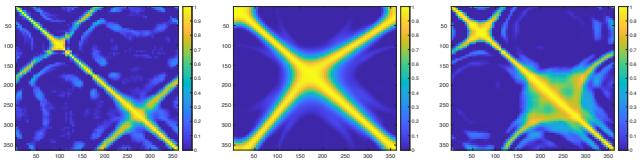

Fig. 2. Correlation matrix of (a) HRTF, (b) VDP, and (c) combined HRTF and VDP.

2.1. Near-field speaker-listener model:

When the speech signal originates from the mouth of the speaker to the left and right ears of the listener, the signal undergoes changes owing to the combination of head-related transfer function (HRTF) of the listener, $H(\theta_{dir})$, and voice directivity pattern (VDP) of the speaker, $V(\theta_{ori})$. The left and right ears are separated by a distance of h which is the width of the head. We define r as the distance between the speaker and the listener. Thus, for the left and right ears, there is an additional offset in angle defined by α and β respectively. When the right ear of the listener is ipsilateral to the speaker, these angles can be derived in terms of h, r, and θ_{dir} as follows.

$$\alpha = \arcsin(\frac{\frac{h}{2}}{r}), \;\; \beta = \arctan(\frac{\frac{h*cos\theta_{dir}}{2}}{r - \frac{h*sin\theta_{dir}}{2}}) \quad (1)$$

The resultant signal at the left and right ear in the frequency domain, $Y_l(f)$ and $Y_r(f)$ can thus be defined as follows.

$$\begin{aligned} Y_l(f) &= X(f)H_{ln}V(\theta_{ori} - \alpha) \\ Y_r(f) &= X(f)H_{rn}V(\theta_{ori} + \beta) \end{aligned} \quad (2)$$

where $X(f)$ is the source speech signal in frequency domain and H_{ln} and H_{rn} are left and right HRTFs compensated for near field.

3. SYSTEM DESIGN

In this section, we elaborate on the inner workings of the system. The system can majorly be divided into 4 sections - 1) The dataset generation pipeline 2) the pre-processing pipeline 3) the Feature extraction pipeline and 4) the 1-D CNN-based regression model as shown in Figure 3.

3.1. Binaural feature extraction:

We extract the interaural level difference (ILD) and interaural time difference (ITD) to be used as core features. We define ILD and ITD as follows.

$$\begin{aligned} ILD(\theta_{dir}, \theta_{ori}, f, t) &= 20\log_{10}(\frac{|Y_l(f,t)|}{|Y_r(f,t)|}) \\ ITD(\theta_{dir}, \theta_{ori}, f, t) &= \frac{1}{2\pi f}\angle(\frac{Y_l(f,t)}{Y_r(f,t)}) \end{aligned} \quad (3)$$

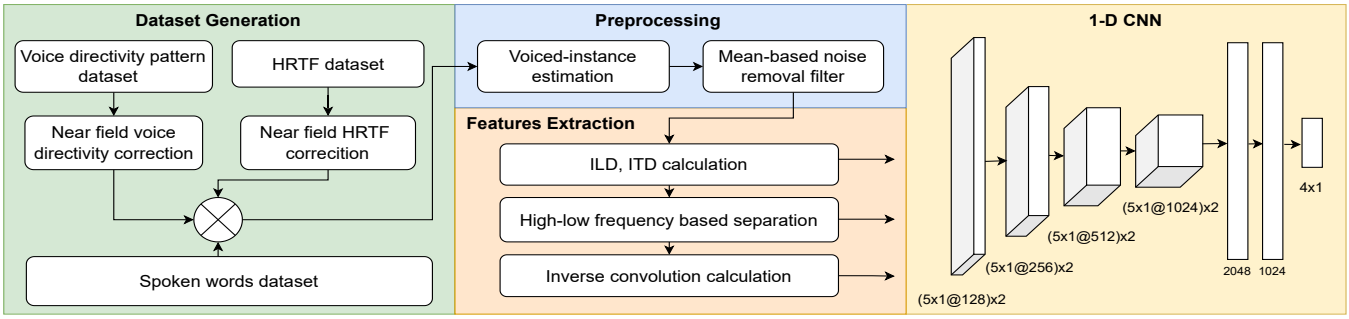

Fig. 3. System overview

3.2. Pre-processing for speech signal:

Voiced vs Unvoiced speech: Voiced speech has a higher directivity index. Thus, during preprocessing, we extract segments that contain voiced speech and mask the rest of the segments as shown in Fig 4. We use the VOICEBOX toolbox on Matlab for this purpose. The toolbox uses PEFAC [16] to first estimate the fundamental pitch and use it to find the probability of the segment being voiced. 1) We identify that the voiced section is sufficient to identify orientation as it has higher directivity. Thus we remove unvoiced sections that could otherwise introduce unnecessary noise. This is done by identifying the harmonic structure of the human speech.

Missing frequencies in Human speech: As seen in Fig 4, human speech has a harmonic structure. Thus, to reduce the noise in preprocessing, we zero force all the values below a threshold. The threshold is decided for each frequency bin based on the mean energy per bin.

3.3. Near-field correction

Most of the large real-world HRTF datasets are collected in the far field (e.g. RIEC dataset is collected at a distance of 1.5m). When trying to convert the far-field HRTF data, there are various near-field effects that need to be taken into account. As the source is nearby, the sound reaches the listener as a spherical wave instead of a planar wave and makes different angles with the ears. Also, the human head acts as a low-pass filter for the contralateral ear. To model these behaviors, [14] apply distance variation functions in the spherical harmonics domain. We follow the same method from the SUpDEq library. We also consider the parallax effect due to the directional nature of the speaker. We consider a simple spherical head model and speaker as a point source to calculate and adjust the VDP.

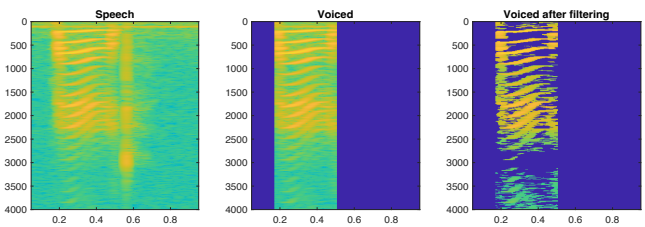

Fig. 4. Signal before and after voiced-unvoiced filtering

3.4. Separating facing configurations in ILD-ITD plane

To infer θ_{dir}, we mainly rely on the ITD values. When the inter-time-difference value is less, it indicates that the listener is facing (towards or away from) the sound source. Having said that, the ITD values at higher frequencies undergo aliasing. The aliasing frequency depends on the size of listener's head h. if we consider h = 18cm, the aliasing frequency would be 952 Hz based on the spatial aliasing limit. Thus we divide ITD values into two halves ITD_{low} and ITD_{high} before sending it to the machine learning model. To infer θ_{ori}, we rely on the ILD values. As VDP has more directivity at higher frequencies, the ILD values at higher frequencies, ILD_{high}, are affected more by the speaker-facing direction than the ILD values at lower frequencies ILD_{low}.

As ILD_{low}, ILD_{high}, ITD_{low}, and ITD_{high} have their unique characteristics, they are fed to the CNN as separate channels so that they can be treated as independent features.

3.5. Enhancing VDF features with inverse convolution

As discussed earlier, voice radiation has different patterns for higher and lower frequencies. The higher frequencies have greater variation with direction [1]. On the other hand, lower frequencies are more omnidirectional. To emphasize this fact, we define a new feature - the ratio of energies of high and low frequencies - that can be useful to determine θ_{ori}. We convolve ILD_{high} with $1/ILD_{low}$ and use the resultant values as the fifth channel of input.

3.6. DNN model design

Architecture: We use a 1-D convolution neural network(CNN) with 8 convolution layers and 3 fully connected layers. We use dropout layer after the convolution layers for regularization. Each layer uses ReLU activation. The input contains 5 channels. We perform FFT on the 1-second recording of the dataset we created earlier and then pass it through the preprocessing pipeline. We use dropout regularization before the fully connected layers. The model is trained on NVIDIA Ampere A100 GPUs using ADAM optimizer with batch size 50 and learning rate 5×10^{-4}.

4. EVALUATION

For generating data to determine listener and speaker head orientations, we use the following three datasets. **(a) HRTF**

dataset: We use the RIEC database[17]. The RIEC database contains HRTFs from 105 subjects (210 different ears). The database is measured in 5-degree intervals in azimuth and 10° interval in elevation. For our study, we are only considering the azimuthal plane. **(b) VDP dataset:** We use the data from [18] [19] and [20] for the VDP. The database contains 288 VDP data. We only use the reference measurements of vowel utterances and ignore the measurements for other scenarios such as the subject holding a hand in front of the mouth or cupping the hands around the mouth. **(c) Spoken words dataset:** We use the Speech Commands database[21] for the spoken words. The database contains 105,829 utterances of 35 words from 2,618 subjects. We generate the audio for our study we randomly pick an HRTF, a VDP, and a spoken word along with a random θ_{dir} value and θ_{ori} value. This gives us a diverse set of listeners, speakers, listener-facing direction, speaker-facing direction, and spoken words. Overall, we generated 560,000 recordings.

4.1. Orientation Estimation

In this section, we discuss the accuracy of the model in predicting θ_{dir} and θ_{ori}. We convert the ground-truth θ_{dir} and θ_{ori} into $sin(\theta_{dir})$, $sin(\theta_{ori})$, $cos(\theta_{dir})$ and $cos(\theta_{ori})$ and use mean-square error loss for training this model. In Fig 5 (a) the blue curve shows the cumulative distribution function (CDF) plot for the error in predicted listener head orientation. We reach the 90^{th} percentile for around 2.5° error. In Fig 5 (b) the blue curve shows the CDF plot for error in predicted speaker head orientation with 90^{th} percentile error 12.5°.

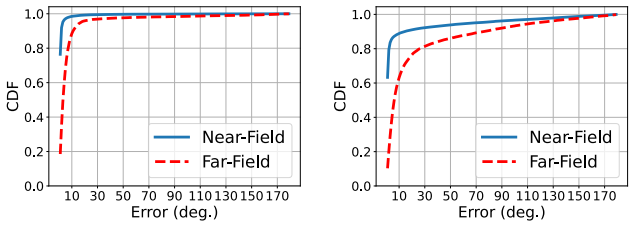

Fig. 5. CDF of error in (a) θ_{dir} and (b) θ_{ori} for near-/far-field.

To verify which azimuthal sector incurs a higher loss, we plot the error for each angular sector. Fig 6 (a) shows the error in θ_{dir} for every <10°. The mean error in each sector is less than <1° for all sectors except one. Fig 6 (b) shows the error in θ_{ori} for every <10°. The mean error in each sector is less than <8° for the majority of the sectors. Note that the training stage assumes the VDP is recorded without unnatural scatterers, such as a hand or facemask covering the mouth. Such scenarios will require the consideration of additional datasets outside the scope of this work.

Impact of near-field correction: In this section, we evaluate how near-field aids in the estimation of θ_{dir} and θ_{ori}. Fig 5 compares the CDF plot for error in θ_{dir} and θ_{ori}. In both cases the error significantly reduces. If we compare the 80th percentile error, θ_{dir} improves from 8° to 1.3° and phi improves from 25° to 2°. The reason behind this improvement is the more pronounced diversity in the acoustic channels between the ears and the mouth even for a smaller change in θ_{dir} and θ_{ori}

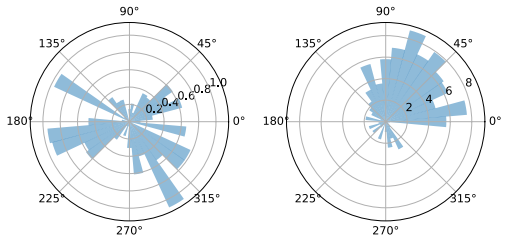

Fig. 6. Polar error plot of (a) θ_{dir} and (b) θ_{ori}

Performance with personalized training: In this section, we compare the results between two cases - 1) when the user's HRTF is known and 2) when it is not known. When the HRTF is known, we re-train our model on that particular HRTF for multiple speakers. Fig 7 (a) and 7 (b) shows the comparison between CDF plots for both cases for θ_{dir} and θ_{ori} respectively. For both θ_{dir} and θ_{ori} the results improve significantly.

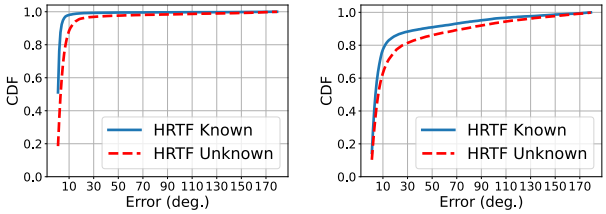

Fig. 7. Error in (a) θ_{dir} and (b) θ_{ori} for know/unknown HRTF.

4.2. Classification performance

We first evaluate our model for accuracy in an application scenario requiring classification of states to find if the speaker and listener are facing each other. One person is facing if the other is within 25° sector in the front. We considered four classes: (a) "facing & facing", (b) "facing & non-facing", (c) "non-facing & facing", and (d) "non-facing & non-facing". The classifier shows accuracy of $> 90\%$ for all configurations except for the "non-facing & non-facing" configuration where the accuracy drops to 89%.

5. CONCLUSION

We presented a system for estimating speaker's direction and head orientation using binaural recording on the listener's ear-mounted microphones in the near-field. The system leverages the speaker's VDP and listener's HRTF jointly to achieve 90^{th} percentile errors of 2.5° and 12.5° in direction and orientation respectively.

Acknowledgments: The authors would like to thank the anonymous reviewers for their helpful comments. This work was partially supported by NSF CAREER Award #2238433 and Meta Research Awards. We also thank the various companies sponsoring the iCoSMoS laboratory at UMD.

6. REFERENCES

[1] HK Dunn and DW Farnsworth, "Exploration of pressure field around the human head during speech," *The Journal of the Acoustical Society of America*, vol. 10, no. 3, pp. 184–199, 1939.

[2] H Steven Colburn, Barbara Shinn-Cunningham, Gerald Kidd, Jr, and Nat Durlach, "The perceptual consequences of binaural hearing: Las consecuencias perceptuales de la audición binaural," *International Journal of Audiology*, vol. 45, no. sup1, pp. 34–44, 2006.

[3] Camille Noufi, Dejan Markovic, and Peter Dodds, "Reconstructing the dynamic directivity of unconstrained speech," *arXiv preprint arXiv:2209.04473*, 2022.

[4] Irtaza Shahid, Yang Bai, Nakul Garg, and Nirupam Roy, "Voicefind: Noise-resilient speech recovery in commodity headphones," in *Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications*, New York, NY, USA, 2022, IASA '22, p. 13–18, Association for Computing Machinery.

[5] Björn Lindblom and Johan Sundberg, "The human voice in speech and singing," *Springer handbook of acoustics*, pp. 703–746, 2014.

[6] Yang Bai, Nakul Garg, Harshvardhan Takawale, Anupam Das, and Nirupam Roy, "Natural voice interface for the next generation of smart spaces," in *Proceedings of the 24th International Workshop on Mobile Computing Systems and Applications*, 2023, pp. 140–140.

[7] Sheng Shen, Daguan Chen, Yu-Lin Wei, Zhijian Yang, and Romit Roy Choudhury, "Voice localization using nearby wall reflections," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–14.

[8] Marco Binelli, Daniel Pinardi, Tiziano Nili, and Angelo Farina, "Individualized hrtf for playing vr videos with ambisonics spatial audio on hmds," in *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2018.

[9] Antoine Deleforge and Radu Horaud, "The cocktail party robot: Sound source separation and localisation with an active binaural head," in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, 2012, HRI '12, p. 431–438, Association for Computing Machinery.

[10] Jing Wang, Jin Wang, Kai Qian, Xiang Xie, and Jingming Kuang, "Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched hrtf condition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, pp. 1–16, 2020.

[11] Catarina Mendonça, Guilherme Campos, Paulo Dias, José Vieira, João P Ferreira, and Jorge A Santos, "On the improvement of localization accuracy with non-individualized hrtf-based sounds," *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 821–830, 2012.

[12] Josefa Oberem, Jan-Gerrit Richter, Dorothea Setzer, Julia Seibold, Iring Koch, and Janina Fels, "Experiments on localization accuracy with non-individual and individual hrtfs comparing static and dynamic reproduction methods," .

[13] Yu-Lin Wei, Rui Li, Abhinav Mehrotra, Romit Roy Choudhury, and Nic Lane, "Inferring facing direction from voice signals," *arXiv preprint arXiv:2109.13094*, 2021.

[14] Johannes M Arend and Christoph Pörschmann, "Synthesis of near-field hrtfs by directional equalization of far-field datasets," *Proc. of the 45th DAGA, Rostock, Germany*, pp. 1454–1457, 2019.

[15] Qiang Yang and Yuanqing Zheng, "Model-based head orientation estimation for smart devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 3, pp. 1–24, 2021.

[16] Sira Gonzalez and Mike Brookes, "Pefac - a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 518–530, 2014.

[17] Kanji Watanabe, Yukio Iwaya, Yo iti Suzuki, Shouichi Takane, and Sojun Sato, "Dataset of head-related transfer functions measured with a circular loudspeaker array," *Acoustical Science and Technology*, vol. 35, no. 3, pp. 159–165, 2014.

[18] Christoph Pörschmann and Johannes M Arend, "Effects of hand postures on voice directivity," *JASA Express Letters*, vol. 2, no. 3, 2022.

[19] Christoph Pörschmann and Johannes M Arend, "Investigating phoneme-dependencies of spherical voice directivity patterns," *The Journal of the Acoustical Society of America*, vol. 149, no. 6, pp. 4553–4564, 2021.

[20] Christoph Pörschmann and Johannes M Arend, "Investigating phoneme-dependencies of spherical voice directivity patterns ii: Various groups of phonemes," *The Journal of the Acoustical Society of America*, vol. 153, no. 1, pp. 179–190, 2023.

[21] Pete Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.