Auditory Cortex-Inspired Spectral Attention Modulation for Binaural Sound Localization in HRTF Mismatch

Waradon Phokhinanan, Nicolas Obin, Sylvain Argentieri

AUDITORY CORTEX-INSPIRED SPECTRAL ATTENTION MODULATION FOR BINAURAL SOUND LOCALIZATION IN HRTF MISMATCH

Waradon Phokhinanan[1,2]*, Nicolas Obin*[2] *and Sylvain Argentieri*[1]

[1]Sorbonne Université, CNRS, ISIR, Paris, France
[2]Sorbonne Université, CNRS, STMS lab, IRCAM, Paris, France

ABSTRACT

In applications like noise cancellation and virtual reality, precise sound source localization is crucial. Existing data-driven binaural systems offer high performance in adverse conditions such as noise and reverberation but face limitations with real-time operation and performance degradation in HRTF mismatch scenarios. Our work introduces a compact Vision Transformer tailored to address these issues, with a primary focus on horizontal speech localization. Inspired by the auditory cortex, our model uniquely incorporates spectral attention mechanisms using encoded speech representations. This architecture enhances generalization on the azimuth plane under mismatched HRTFs. Our empirical results show a marked improvement over conventional DNN, CNN-based and Transformer-based models, both in noisy and noise-free environments. Significantly, the proposed model maintains high accuracy in localizing adjacent azimuths, ideal for real-world applications.

***Index Terms*—** sound source localization, binaural audition, HRTF mismatch, attention modulation

1. INTRODUCTION

Binaural sound source localization (BSSL) is an important subsystem that enhances performance in various audio processing domains, such as speech enhancement and speaker identification. For a full review, please see [1]. BSSL is increasingly emphasized, particularly in two-sensor applications like humanoid robots, hearing aids, and virtual reality, where array-based sensors are probably not suitable.

Most BSSL methods rely on the Duplex Theory of binaural cues, such as Interaural Time/Phase Difference (ITD/IPD) and Interaural Level Difference (ILD) [2]. These cues are derived from Head-Related Transfer Functions (HRTFs), which mathematically describe how sound is filtered by an individual's head, ears, and torso. Thus, the notable challenge in BSSL algorithms, aside from noise and reverberation, is that most algorithms perform well only when the same HRTF that was used during simulated training is applied. Recent studies have seldom tackled the issue of HRTF mismatch [3]. Our research aims to address this gap.

2. RELATED WORK

Efforts to enhance targeting accuracy in BSSL primarily focus on improving generalization in noisy and reverberant conditions [4]. These strategies are versatile, designed to function effectively across various environments, whether in full-sphere or on vertical/horizontal planes. Recently, researchers have utilized a diverse set of data-driven techniques, ranging from Deep Feedforward Neural Networks (DNNs) [5] and Convolutional Recurrent Neural Networks (CRNNs) [6], to specialized variants like Transformer-based models [7]. While these models perform well with matched HRTFs, their efficacy in cases of HRTF mismatch remains unexplored.

The generalized parametric models [8] and HRTF template matching techniques [9] could potentially serve as preliminary approaches for addressing HRTF mismatches. However, these models may lack adequate parameter fine-tuning for individual variations, especially in real-world localization of unseen HRTFs. On the other hand, the DNN-based approach proposed by [3] aims to improve HRTF generalization by clustering HRTF databases to identify representative HRTFs. These selected representations are then used to enhance localization accuracy, obviating the need for training on multiple HRTF subjects. Nevertheless, this methodology does not take into account the impact of HRTF mismatches in noisy environments. Similar clustering approaches have also been proposed by [10, 11].

3. PROPOSED ARCHITECTURE

Our objective is to enhance BSSL performance under conditions of HRTF mismatches and noisy environments, while also adhering to computational efficiency constraints to ensure applicability in real-world scenarios.

Inspired by the selective attention mechanisms of the auditory cortex during cocktail parties [12], selective attention provides us with the cognitive flexibility to either focus on a particular sound source in a noisy environment or to assess the entire auditory scene. This capability is enhanced by our top-down regulation of cortical circuits, a phenomenon substantiated by both empirical research and computational models in

neuroscience [13, 14]. Prior BSSL models, inspired by this top-down attention to target sounds in noisy and reverberant environments, utilize masking techniques to filter out interference. These techniques include both shallow integration and layer modulation, as seen in works such as [5, 15]. Nevertheless, such masking approaches may inadvertently discard important information during the localization process [16].

Several studies highlight the importance of spectral tuning for focus in selective attention tasks, using quick changes in spectrotemporal receptive fields [14, 16, 17]. Our model, the Attention Modulation Vision Transformer (AMViT), builds on these findings. It uses a data-driven encoder to learn speech representations as top-down information in both noisy and clear settings. It then dynamically modulates this top-down information in each time-frequency bin, guided by bottom-up binaural cues, thereby moving beyond the sole reliance on masking to access top-down information.

The development of AMViT is an extension of our previous work on Frequency-based Audio Vision-Transformers (FAViT) for BSSL [7]. Compared to other architectures, FAViT features strong time-frequency bin selectivity, enabled by its self-attention mechanism, to localize sound while maintaining a relatively low parameter count. We have therefore refined FAViT to serve as an efficient and effective encoder, optimized for implementing spectral attention modulation.

The key to improving performance in HRTF mismatch scenarios lies in our hypothesis that integrating encoded speech as top-down information with bottom-up binaural cues will enhance the model's accuracy. We propose that the model will learn the interplay between these cues, thereby increasing its adaptability to unfamiliar HRTFs.

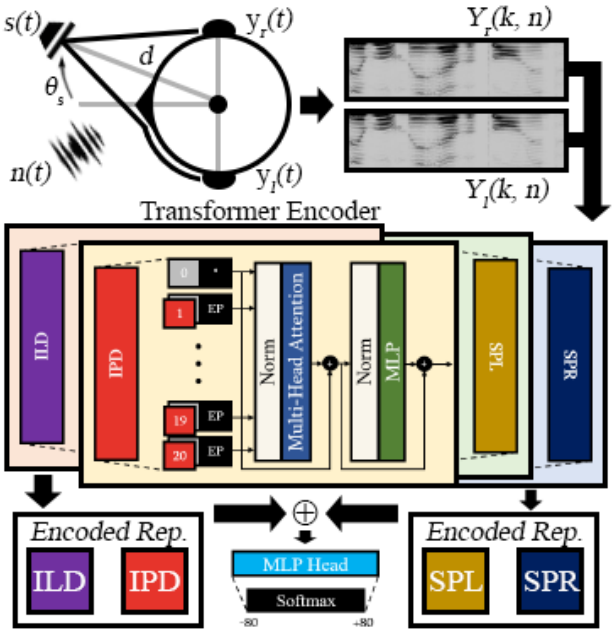


Fig. 1. Adaptation of auditory-inspired attention in AMViT

3.1. Binaural and Spectral Features

Our scheme aims to localize horizontal sound using azimuth angles θ_s, measured in radians. As depicted in Figure 1, both the speaker signal $s(t)$ and ambient noise $n(t)$ are captured relative to a binaural sensor located at coordinates (d, θ_s), which are relative to the center of the head. The signal $s(t)$ is divided into the left $y_l(t)$ and right $y_r(t)$ channels due to the influence of the head. We establish a correlation between these channels and θ_s through their Short-Time Fourier Transform (STFT) spectra $Y_{l/r}$ defined as

$$\begin{aligned} Y_l(k,n) &= \text{STFT}\left(h_l(\theta_s,t) * y_l(t) + n_l(t)\right), \\ Y_r(k,n) &= \text{STFT}\left(h_r(\theta_s,t) * y_r(t) + n_r(t)\right). \end{aligned} \tag{1}$$

The STFT spectra $Y_l(k,n)$ and $Y_r(k,n)$, with frequency index k and time index n, are computed using a Hamming window and a 50 ms overlap from 640 samples and 320 frequency bins at a 16 kHz sampling rate. The Head-Related Impulse Responses $h_l(\theta_s,t)$ and $h_r(\theta_s,t)$ affect the left and right channels, respectively, while $n_l(t)$ and $n_r(t)$ represent noise. The asterisk $*$ signifies convolution.

Our architecture extracts four principal features: two top-down from left/right spectrograms, and two bottom-up from IPD and ILD. SPL (Spectral-feature Left) and SPR (Spectral-feature Right) are derived from $Y_l(k,n)$ and $Y_r(k,n)$ by separating their real and imaginary parts, with

$$\begin{aligned} \text{SPL}(k,n,2) &= (\text{Re}(Y_l(k,n)), \text{Im}(Y_l(k,n)), \\ \text{SPR}(k,n,2) &= (\text{Re}(Y_r(k,n)), \text{Im}(Y_r(k,n)), \\ \text{ILD}(k,n) &= 20\log_{10}\left(\frac{|Y_l(k,n)|}{|Y_r(k,n)|}\right), \\ \text{IPD}(k,n) &= \angle\frac{Y_l(k,n)}{Y_r(k,n)}. \end{aligned} \tag{2}$$

3.2. Patches and Vision Transformer Encoder

The position embedding and attention mechanism remain consistent with FAViT [7]. However, we have modified the architecture: FAViT's single Transformer is now turned into four distinct encoder blocks, each specialized for one of the four features. These encoders accept input shapes (k,n) or $(k,n,2)$ and map them to dimensions $(f,20)$, where f is the number of frequency patches calculated as $f = \frac{k}{n}$ (see Eq. 3). With all these modifications, we anticipate achieving superior performance compared to FAViT in terms of both HRTF and noise mismatch.

In our configuration, $k = 320$ is based on STFT discussed in the previous section, while $n = 16$ frames, corresponding to a total time span of 640 ms. These brief time frames could potentially offer advantages for real-time sound localization. Overall, the extended Transformer blocks in our model collectively comprise 0.25 million parameters. Each encoder block employs a 4-layer Multi-Head Self-Attention (MSA)

followed by a Multilayer Perceptron (MLP) as a self-attention mechanism, iterated eight times (iter $= 8$). The MLP layers use a 0.005 learning rate, 0.0001 weight decay, batch size of 16, and the Adam optimizer. This turns each patch into a 20-dimensional vector. The dot products of encoded ILD and IPD features form a bottom-up representation (RP_{BU}). Similarly, the dot products of encoded spectrograms for the left and right audio channels generate a top-down representation (RP_{TD}). We chose dot products based on evidence showing they are robust and do not improve accuracy with alternative methods in our experiments. In the end, the parameters used in this paper are defined along

$$\begin{aligned} \text{RP}_{\text{feature}}(f, 20) &= \text{Encoder}^{\text{iter}=8}(\text{Feature}(shape)), \\ \text{RP}_{\text{BU}}(f, 20) &= \text{RP}_{\text{ILD}}(f, 20) \cdot \text{RP}_{\text{IPD}}(f, 20), \\ \text{RP}_{\text{TD}}(f, 20) &= \text{RP}_{\text{SPL}}(f, 20) \cdot \text{RP}_{\text{SPR}}(f, 20). \end{aligned} \tag{3}$$

3.3. Modulation and Classification

The next step focuses on modulating RP_{BU} and RP_{TD}, both of which serve as frequency-distributed representations. Conventional top-down attention mechanisms within the auditory cortex are more closely associated with nonlinear spike activities than with the actual auditory stimuli. While the modulation of these spike activities is still an active area of research [16, 14], our approach takes cues from these neural behaviors to enhance our time-frequency representation. Various mathematical techniques, such as addition and subtraction, have been explored for the purpose of modulating binaural cues [18]. Furthermore, numerous neuroscientific studies advocate for the inclusion of an additional layer to unify these diverse representations [16, 14].

Building on the methods described above, our study evaluates five modulation methods denoted as $\oplus$ in the following: addition, subtraction, element-wise multiplication, dot product, and a 125-neurons MLP layer. Our objective is then to identify the optimal modulation technique for our representations. The classifier features a 3-layer MLP head with dimensions $[512, 256, 100]$. An L2 regularizer is applied to the MLP weights for regularization. The output layer uses a Softmax function to normalize the final representation, which is denoted as Mod_{type} and defined according to

$$\begin{aligned} \text{Mod}_{\text{type}} &= \begin{cases} \text{RP}_{\text{BU}} \oplus \text{RP}_{\text{TD}}, & \oplus \in \{+, -, \times, \cdot\} \\ \text{MLP}(\text{RP}_{\text{BU}}, \text{RP}_{\text{TD}}) \end{cases}, \\ \text{CL}(\theta_s) &= \text{Softmax}(\text{MLP}_{\text{Head}}(\text{Mod}_{\text{type}})), \end{aligned} \tag{4}$$

where $\text{CL}(\theta_s)$ are class probabilities computed from Mod_{type} for each localization azimuth. A Cross-Entropy Loss function is used for optimization. The addition of this classification layer increases the model's total parameter count by approximately one million.

4. EXPERIMENTS

4.1. Experimental Design and Evaluation

The experiment localizes a single simulated sound source to one of 25 azimuth angles on a horizontal plane. These angles are specified at intervals that include $-80°, -65°, -55°$, as well as $-45°$ to $45°$ in $5°$ increments, and $55°, 65°, 80°$.

We evaluate the performance of our proposed model using five different modulation methods, as described in Section 3.3. These methods are denoted as AMViT1 for addition, AMViT2 for subtraction, AMViT3 for multiplication, AMViT4 for dot product, and AMViT5 for an additional MLP layer. These evaluations are carried out under varying conditions of noise, speakers, and both the same and different HRTFs. We also compare our model's performance against benchmark models, including a replicated DNN model [3], CNN model [15], and the original FAViT model [7]. The DNN model was chosen as it was used to address HRTF mismatch, while the CNN model was selected to represent recent top-down modulation techniques using masking. All spatialized training and testing data were prepared once and kept consistent across all experiments.

The proposed metrics include classification accuracy, and both a tolerant accuracy allowing ± 1 class deviation and RMSE in degrees, defined as

$$\begin{aligned} \text{Tolerant Accuracy} &= \frac{1}{N} \sum_{s=1}^{N} \mathbb{1} \left(|C_s - \hat{C}_s| \leq 1 \right), \\ \text{RMSE} &= \sqrt{\frac{1}{N} \sum_{s=1}^{N} \left(g(C_s) - g(\hat{C}_s) \right)^2}, \end{aligned} \tag{5}$$

where g is a function defined to map classes to azimuth angles for RMSE calculations, N is the sample count, C_s is the true class, and $\hat{C}_s = \text{argmax}(\text{CL}(\theta_s))$ is the predicted class.

4.2. Data and Acoustic Elements

We utilized the TIMIT database [19] for speech data, choosing 40 audio signals for training and a distinct set of 16 for testing, with an equal gender distribution. These sets were spatialized based on Eq. 1, using 45 HRTF subjects from the CIPIC database [20] for training and 5 from the REIC database [21] for testing. These HRTFs simulated sound source locations at 25 angular positions, as outlined in Section 4.1. Audio was resampled to 44.1 kHz for convolution with HRIR, then downsampled to 16 kHz.

For additive noise, we selected four types each from the spatially uncorrelated Noisex92 [22] and non-stationary noise databases [23] for training and testing, respectively, in line with [7]. The model was trained and tested at various SNRs: $[-5, 5, 15, 25]$ dB for training and $[0, 10, 20]$ dB for testing, with the aim of enhancing generalization performance across different SNRs, noise types, and HRTF conditions.

Models	Parameters	Seen HRTFs		Unseen HRTFs					
		Unseen Noises		Seen Noises		Unseen Noises			
		Acc.	RMSE	Acc.	RMSE	Acc.	RMSE	Tolerant Acc.	RMSE
DNN (Wang et al.)	2.0 M	81.6	2.5	59.1	8.7	43.3	11.5	70.9	5.8
CNN (Hu et al.)	0.8 M	89.4	2.3	58.6	8.8	47.6	11.2	72.4	5.5
FAViT	0.6 M	89.2	2.4	59.4	8.5	49.3	11.3	73.8	5.4
AMViT1 (+)	1.3 M	88.5	2.4	44.1	12.5	33.1	16.0	41.4	9.2
AMViT2 (-)	1.3 M	87.6	2.5	45.6	12.2	28.4	17.0	40.7	9.5
AMViT3 ($\times$)	1.3 M	**93.5**	**2.0**	**63.4**	**7.5**	**53.5**	**11.0**	**89.6**	4.5
AMViT4 ($\cdot$)	1.3 M	82.6	2.6	40.5	13.0	26.4	17.5	33.5	10.0
AMViT5 (MLP)	1.4 M	93.4	2.1	61.2	8.0	51.2	11.8	89.5	**4.4**

Table 1. The table shows localization accuracy (%), parameters, and RMSE for seen/unseen HRTFs and noise conditions.

5. RESULTS AND DISCUSSION

5.1. Modulation and Efficiency Performance

As illustrated in Table 1, our newly proposed model AMViT3, which employs multiplication-based modulation, performs the best in all scenarios and marginally outperforms its MLP-based equivalent AMViT5. It achieves an accuracy of 53.5% compared to AMViT5's 51.2%, particularly excelling in the most challenging tasks, unfamiliar noise and HRTFs. Alternative modulation methods like addition and dot products are less effective, suggesting they could dilute key information. Nonetheless, it is worth considering that fine-tuning the hyperparameters of the MLP layers or exploring other layer options for modulation may yield further improvements.

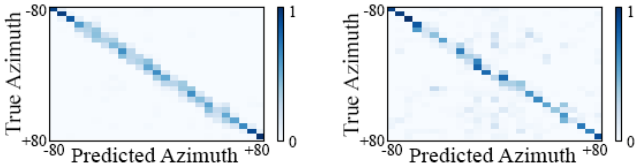
Our AMViT3 model approximately doubles the parameter count of the original FAViT due to the inclusion of three new encoders. Despite this, it remains more parameter-efficient than other large deep-learning models (over 20 M.). AMViT3 demonstrates robust performance in both familiar and unfamiliar HRTF conditions, peaking at 93.5% accuracy in scenarios with unfamiliar noise but known HRTFs. These results indicate that our top-down modulation of speech representation enhances localization capabilities in noisy environments, aligning well with our human-inspired objectives.

5.2. HRTF Generalization

Addressing HRTF mismatch is a significant challenge due to individual variability. This is evident as all models perform similarly under the same or different HRTF conditions in both noisy and noise-free environments. Although our best model, AMViT3, achieves a higher accuracy rate of 65.4% in identical noise conditions, it underscores the persistent challenges in the field. Although our model fell short of the 80% accuracy goal, it did substantiate our hypothesis: combining auditory traits and binaural cues improves resilience. It outperformed existing benchmarks and excelled in HRTF generalization in both seen and unseen noise conditions. With more HRTF data, we expect the model could improve further.

AMViT3 and FAViT exhibit relatively similar accuracy in determining the exact locations of unseen HRTFs. However, as Figure 2 clearly demonstrates, the distribution patterns of their confusion matrices differ significantly. This distinction enables AMViT3 to outperform FAViT in terms of tolerance with an accuracy rate of $89.6\% \pm 1$ within a one-class range, making it suitable for real-world applications. This suggests that AMViT3's top-down modulation enhances performance in HRTF mismatches. Please note that our AMViT3 and benchmark models could perform better with more speech data. However, limited computational resources constrain us, as we trained on 45 HRTF subjects. Despite this, AMViT3 shows strongest localization accuracy in the same environment.

To boost HRTF generalization, future work could explore better time-frequency speech features and consider dimensionality reduction like PCA. Extensions to reverberant conditions and elevation mismatches are also worth investigating.



(a) AMViT3: Acc. 53.5% (b) FAViT: Acc. 49.3%

Fig. 2. Confusion matrices: AMViT3 (diagonal-dominant) vs. FAViT (scattered) under unseen noises and HRTFs

6. CONCLUSION

Our experiment confirms our hypotheses: top-down spectral modulation enhances sound localization amid unfamiliar noises and HRTF mismatches. Element-wise multiplication and an extra MLP layer excel in our binaural and spectral setups. While our AMViT model is not perfect at pinpointing exact azimuths, its strong performance in adjacent positions suggests real-world utility.

7. REFERENCES

[1] Dhwani Desai and Ninad Mehendale, "A review on sound source localization systems," *Archives of Computational Methods in Engineering*, vol. 29, no. 7, pp. 4631–4642, 2022.

[2] Joseph Carl Robnett Licklider, "A duplex theory of pitch perception," *The Journal of the Acoustical Society of America*, vol. 23, no. 1_Supplement, pp. 147–147, 1951.

[3] 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.

[4] Pierre-Amaury Grumiaux, Sran Kitić, Laurent Girin, and Alexandre Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, 2022.

[5] Ning Ma, Jose A Gonzalez, and Guy J Brown, "Robust binaural localization of a target sound source by combining spectral source models and deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2122–2131, 2018.

[6] Paolo Vecchiotti, Ning Ma, Stefano Squartini, and Guy J Brown, "End-to-end binaural sound localisation from the raw waveform," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 451–455.

[7] Waradon Phokhinanan, Nicolas Obin, and Sylvain Argentieri, "Binaural Sound Localization in Noisy Environments Using Frequency-Based Audio Vision Transformer (FAViT)," in *Proc. INTERSPEECH 2023*, 2023, pp. 3704–3708.

[8] Martin Raspaud, Harald Viste, and Gianpaolo Evangelista, "Binaural source localization by joint estimation of ild and itd," *Ieee transactions on audio, speech, and language processing*, vol. 18, no. 1, pp. 68–77, 2009.

[9] Fakheredine Keyrouz, Youssef Naous, and Klaus Diepold, "A new method for binaural 3-d localization based on hrtfs," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. IEEE, 2006, vol. 5, pp. V–V.

[10] Hugh O'Dwyer and Francis Boland, "Hrtf clustering for robust training of a dnn for sound source localization," *Journal of the Audio Engineering Society*, vol. 70, no. 12, pp. 1015–1026, 2022.

[11] Kai Qian, Jing Wang, Wenjing Yang, and Miao Liu, "Binaural sound source localization based on neural networks in mismatched hrtf condition," in *Proceedings of the 8th International Conference on Computing and Artificial Intelligence*, 2022, pp. 62–67.

[12] Josh H McDermott, "The cocktail party problem," *Current Biology*, vol. 19, no. 22, pp. R1024–R1027, 2009.

[13] Xiao-Jing Wang and Guangyu Robert Yang, "A disinhibitory circuit motif and flexible information routing in the brain," *Current opinion in neurobiology*, vol. 49, pp. 75–83, 2018.

[14] Kenny F Chou and Kamal Sen, "Aim: A network model of attention in auditory cortex," *PLoS Computational Biology*, vol. 17, no. 8, pp. e1009356, 2021.

[15] Qi Hu, Ning Ma, and Guy J Brown, "Robust binaural sound localisation with temporal attention," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] Jordan D Chambers, Diego Elgueda, Jonathan B Fritz, Shihab A Shamma, Anthony N Burkitt, and David B Grayden, "Computational neural modeling of auditory cortical receptive fields," *Frontiers in computational neuroscience*, vol. 13, pp. 28, 2019.

[17] Jennifer K Bizley and Yale E Cohen, "The what, where and how of auditory-object perception," *Nature Reviews Neuroscience*, vol. 14, no. 10, pp. 693–707, 2013.

[18] Kiki van der Heijden and Siamak Mehrkanoon, "Goal-driven, neurobiological-inspired convolutional neural network models of human spatial hearing," *Neurocomputing*, vol. 470, pp. 432–442, 2022.

[19] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, pp. 27403, 1993.

[20] V Ralph Algazi, Richard O Duda, Dennis M Thompson, and Carlos Avendano, "The cipic hrtf database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*. IEEE, 2001, pp. 99–102.

[21] Kanji Watanabe, Yukio Iwaya, Yôiti 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.

[22] Andrew Varga and Herman JM Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[23] David Dean, Sridha Sridharan, Robert Vogt, and Michael Mason, "The qut-noise-timit corpus for evaluation of voice activity detection algorithms," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association*. International Speech Communication Association, 2010, pp. 3110–3113.