Investigation on the Band Importance of Phase-aware Speech Enhancement

Zhuohuang Zhang[1,2], *Donald S. Williamson*[1], *Yi Shen*[3]

[1]Department of Computer Science, Indiana University, USA
[2]Department of Speech, Language and Hearing Sciences, Indiana University, USA
[3]Department of Speech and Hearing Sciences, University of Washington, USA

zhuozhan@iu.edu, williads@indiana.edu, shenyi@uw.edu

Abstract

Many existing phase-aware speech enhancement algorithms consider the phase at all spectral frequencies to be equally important to perceptual quality and intelligibility. Although improvements are observed according to both objective and subjective measures, as compared to phase-insensitive approaches, it is not clear whether phase information is equally important across the frequency spectrum. In this paper, we investigate the importance of estimating phase across spectral regions, by conducting a pairwise listening study to determine if phase enhancement can be limited to certain frequency bands. Our experimental results suggest that estimating phase at lower-frequency bands is mostly important for speech quality in normal-hearing (NH) listeners. We further propose a hybrid deep-learning framework that adopts two sub-networks for handling phase differently across the spectrum. The proposed hybrid-net significantly improves the model compatibility with low-resource platforms while achieving superior performance to the original phase-aware speech enhancement approaches.
Index Terms: speech enhancement, phase, speech perception

1. Introduction

Environmental noise significantly influences communication between humans and with many applications, such as automatic speech recognition (ASR) and hearing-aids. Numerous speech enhancement algorithms have been proposed to alleviate this speech-in-noise problem [1, 2, 3, 4, 5, 6], where they can generally be divided into phase-insensitive and phase-aware approaches. With phase-aware speech enhancement, the approaches, in general, simultaneously enhance the magnitude and phase responses of noisy speech, after generating a time-frequency (T-F) representation using the short-time Fourier transform (STFT). Phase-insensitive approaches, on the other hand, only enhance the magnitude response.

Multiple studies have shown the perceptual speech quality benefits that phase-aware speech enhancement offers to normal hearing (NH) listeners [7, 8], and its potential benefits to hearing impaired (HI) listeners [9, 10]. Many existing phase-aware speech enhancement approaches allocate the same computational resources to all frequency regions during phase estimation [2, 11, 12], which, in turn, assumes that the estimated phase, at each frequency, is equally important to speech quality for human listeners. The contributions of various frequency regions to speech understanding, i.e., speech intelligibility, has been extensively studied over the past decades [13, 14, 15, 16]. As a result, band importance functions (BIF) have been widely used to characterize the relative importance of different frequency bands to speech intelligibility. It has been reported that the most important spectral regions for speech understanding are at frequencies between 1600 to 2000 Hz [17, 18]. These conclusions, however, are with regards to the magnitude response of speech, and do not consider changes to the phase spectrum. Hence, little is known about how phase from different frequency bands contributes to the the underlying mechanisms of perceived speech quality. From a practical point of view, understanding the impact that different frequency bands of the phase response have on speech quality could help conserve computational resources for more perceptually important regions and may potentially reduce model complexity.

In this work, we hypothesize that phase information from different spectral regions does not uniformly contribute to speech quality. Specifically, we conjecture that pitch perception plays an important role in quality judgement. Pitch of a broadband periodic signal, such as the voiced segments in speech, is coded mainly by temporal fine structure (TFS) cues and spectral cues of resolved harmonics at low frequencies [19, 20, 21]. At high frequencies, when the individual harmonics can no longer be resolved by the peripheral auditory system, temporal envelope cues may also contribute to weak pitch perception [22]. Given the greater importance of low frequencies for pitch perception, it may be also more important to accurately estimate phase information at low frequencies compared to high frequencies. To verify this, we examine the band importance of phase estimation for speech quality judgement by systematically removing the phase information of high-frequency regions. Two versions of enhanced speech are composed from a phase-insensitive and a phase-aware speech enhancement model [23, 24]. In particular, the low-frequency portion of the phase-aware enhanced speech is manually merged with the high-frequency portion of the phase-insensitive enhanced speech to generate the speech stimuli with different phase information across spectral regions. We will refer to these stimuli as the filtered-merged speech below. A listening study is conducted, in which participants compared the perceived speech quality between the full-band phase-aware and the filtered-merged speech. If participants fail to discriminate the two stimuli, then it would indicate that high-frequency phase information does not influence speech quality in a perceptually significant manner.

Furthermore, a novel hybrid speech enhancement framework is proposed driven by the findings of band importance of phase estimation. Specifically, the proposed hybrid-net adopts different strategies dealing with phase estimation in different frequency regions. A phase-aware deep-learning sub-network is adopted for low-frequency bands and another phase-insensitive sub-network is applied for high-frequency bands, based on our findings that low-frequency phase estimation contributes mostly to speech quality. The approaches are evaluated on a simulated speech corpus using a human listening study. Network statistics including network size and computational cost are also reported.

The rest of this paper is organized as follows. Section 2 describes the method to investigate the band importance of phase estimation. Section 3 introduces the band importance-driven hybrid-net and reports the performance from a listening study. Finally, we discuss the results and findings in section 4.

2. Band importance for phase estimation

To investigate the importance of estimating phase at different frequency bands, we first generate speech samples by merging the enhanced speech produced by two independent speech enhancement algorithms (i.e., phase-aware and phase-insensitive). In particular, the low-frequency components of phase-aware enhanced speech are combined with the high-frequency components of phase-insensitive enhanced speech, and the crossover frequency is systematically varied. If listeners cannot tell a difference after replacing high-frequency component with phase-insensitive speech above a certain crossover frequency, then we may infer that the phase information above the crossover frequency does not significantly contribute to speech quality. Accordingly, we conduct a pairwise comparison listening study between filtered-merged and full band phase-aware enhanced speech to find the crossover frequency where the benefits from phase-aware speech enhancement start to diminish.

2.1. Speech materials and system configurations

A total of 1440 clean speech utterances (i.e., 720 utterances for each gender) from IEEE corpus [25] are used. In the training set, 80% of them are mixed with eight different noises, including multi-talker babble, factory, cafeteria, thunderstorm, washing machine, vacuum, train and engine noises from AzBio [26], NOISEX-92 [27] and ESC-50 corpora [28]. The signal-to-noise ratios (SNRs) in the training set range from -6 to 0 dB with a step size of 1 dB, resulting in 64512 mixtures. 10% of the speech signals are used to generate the development and testing sets. Similar to the setup described in [24], we mix the speech with factory noise at a -5 dB SNR in the development set. In the testing set, the remaining 10% of the speech utterances are mixed with babble and cafeteria noises at -5, 0, and 5 dB SNRs (864 mixtures in total). All signals are resampled to 16 kHz before further processing. We use a 320-point FFT together with a 20 ms hamming window (10 ms hop size) for the STFT.

We adopt two state-of-the-art speech enhancement models, based on the convolutional recurrent network (CRN) [23, 24], for phase-insensitive and phase-aware speech enhancement, respectively. The CRNs feature an encoder-decoder-like architecture, with a recurrent block between the encoder and decoder to help capture the temporal correlations. Similar to the original works [23, 24], there are five convolution/deconvolution blocks in the encoder/decoder, with $Time \times Frequency = 1 \times 3$ kernels. The output channels are set to $(16, 32, 64, 128, 256)$ in the encoder, and $(128, 64, 32, 16, 1)$ in the decoder. Batch normalization [29] and exponential linear units (ELU) [30] are applied after each convolution/deconvolution layer except in the output layer. Skip connections are also used between the encoder and decoder. The recurrent block is based on long short-term memory (LSTM) cells with two layers and 1024-units each. To merge the enhanced speech signals from the two models, the output of the phase-insensitive model is high-pass filtered while the output of the phase-aware model is low-pass filtered using Butterworth filters with a 70-dB attenuation in the stop band. The crossover frequencies between the high- and low-frequency components are 250, 1000, 2000, 4000 Hz in separate conditions. For each crossover frequency, there are 20 enhanced speech utterances randomly selected for the testing set. Half of them with original SNR at -5 dB and the other half at 5 dB.

2.2. Procedure

A total of 20 participants were recruited (12 males and 8 females), ranging from 23 to 58 years of age (avg. 35.6 years) using Amazon Mechanical Turk. All participants were native speakers of American English that self-reported to have normal hearing (NH). This study was approved by the Institutional Review Board (IRB) at Indiana University and the University of Washington. Informed consent was obtained from all participants before data collection began.

All participants self-reported that they were seated in a quiet environment during the listening study. A behavioral headphone check procedure, based on discrimination of binaural pitch [31, 32], was first included to ensure headphones were worn by all participants. The headphone check would fail if loudspeakers were used or the headphone was only worn on one ear.

Following the headphone check, five practice trials were conducted to familiarize the subjects with the experimental task and to allow them to adjust the volume to a comfortable level. In each practice trial, the participant compared the quality between a pair of full band phase-insensitive and phase-aware enhanced speech. The participant had the opportunity to replay each of the stimuli an unlimited number of times.

In the main experimental task, the participants compared pairs of full band phase-sensitive enhanced speech and filtered-merged speech. As in the practice trials, the participant was able to replay each of the stimuli multiple times, before identifying the one with a higher perceived quality. The order of the two stimuli was permuted across trials. The crossover frequencies for the filtered-merged speech were 0 Hz (i.e., using original noisy phase for reconstruction, the phase-insensitive model, denoted as 'Mag.'), 250 Hz, 500 Hz, 1000 Hz, 2000 Hz and 4000 Hz. There were a total of 120 trials with 120 different test sentences (i.e., 10 repetitions $\times$ 2 SNRs $\times$ 6 crossover frequencies), tested in random order.

2.3. Results

For each crossover frequency, the discriminability between the full-band phase-aware and filtered-merged speech is summarized using the d' (d-prime) metric, which is defined as [33]:

$$d' = z(H) - z(F), \tag{1}$$

where z denotes the z-transform (i.e., inverse Gaussian distribution), H and F represent the hit rate and false alarm rate, respectively. A hit occurs when the full-band phase-aware enhanced speech is chosen as the preferred stimulus; a false alarm occurs when the filtered-merged speech is chosen as the preferred one. Note that a positive d' indicates a preference towards full-band phase-aware enhanced speech. A d' close to zero indicates that the listeners cannot reliably differentiate the difference between the two stimuli. The estimated d' will be used as the dependent variable to study the effects of crossover frequency and SNR.

Figure 1 provides the d' values for each condition ('Mag.' indicates the phase-insensitive model). Note that all conditions are compared against the stimuli generated by the full-band phase-aware speech enhancement model. It is observed that the d' value decreases as the crossover frequency increases.

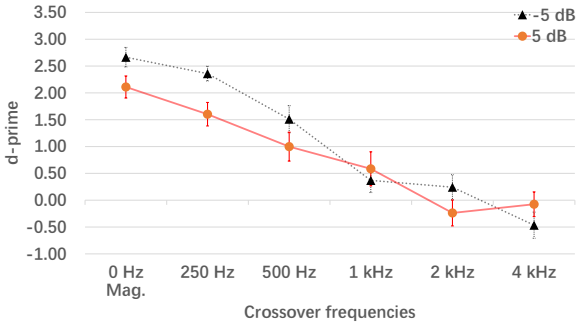

Figure 1: *d' values as a function of crossover frequency and mixture SNR. Error bars indicate the $\pm$ standard errors.*

For low crossover frequencies, there is a strong preference for the full-band phase-aware enhanced speech over the filtered-merged speech, reflected as large positive d' values. On the other hand, listeners cannot reliably differentiate the two enhancement models with high crossover frequencies ($\geq$ 2 kHz). This means that estimating phase information at frequencies higher than 2 kHz does not seem to improve speech quality.

A repeated measures analysis of variance (ANOVA) is conducted, and it shows significant main effects of crossover frequency [$F(3.06, 58.10) = 48.58$, $p < .001$, $\eta_p^2 = .719$, Greenhouse-Geisser corrected], and background SNR [$F(1, 19) = 4.68$, $p = .043$, $\eta_p^2 = .198$]. We also found significant interactions between crossover frequency and background SNR [$F(5, 95) = 3.73$, $p = .004$, $\eta_p^2 = .164$], where the effect of SNR is stronger at lower crossover frequencies (as reflected in Figure 1). This suggests that estimating low-frequency phase information is more important at lower SNRs.

3. Band Importance-driven hybrid-net for phase-aware enhancement

Inspired by the finding from the listening experiment described above, we introduce a novel hybrid speech enhancement framework in this section. The proposed framework consists of a phase-aware CRN that estimates low-frequency speech and then merges them with results from another phase-insensitive CRN that processes the high-frequency components.

3.1. Network architecture

The architecture of the proposed approach (denoted as hybrid-net) is shown in Figure 2. The noisy speech is first transformed into the T-F domain using the STFT, where the STFT is further split into separate halves, using a 4 kHz cutoff frequency. The real and imaginary parts of the lower frequency (i.e., 0 to 4 kHz) spectrogram are fed into a sub-network based on a phase-aware CRN. On the other hand, the high frequency (i.e., 4 to 8 kHz) magnitude spectrogram is fed into another sub-network based on a phase-insensitive CRN. Next, the two sub-networks encode the input features separately with additional residual connections (i.e., additions) to enable information sharing across sub-networks within the encoders, recurrent layers and decoders. The phase-aware CRN then estimates the real and imaginary parts for the lower-frequency half of the spectrogram, while the phase-insensitive CRN predicts the magnitude spectrogram for the higher-frequency half. Lastly, the estimated spectrograms (noisy phase used for higher-frequency portion) are resynthesized to time-domain enhanced speech using the inverse STFT.

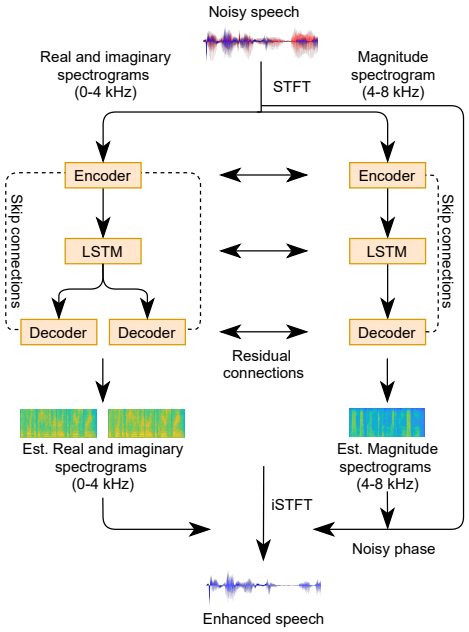

Figure 2: *Network architecture of the proposed hybrid-net.*

The configurations of the two sub-networks are identical to the ones described before, except that we use two 512-unit recurrent layers in both networks. The same speech materials are used. The hybrid-net is trained with 60 epochs (or until convergence). ADAM optimization [34] is used with a learning rate of $1e^{-3}$. The mini-batch size is set to 24.

3.2. Listening experiment

An additional 20 subjects (12 males and 8 females) were recruited from Amazon Mechanical Turk, with ages from 23 to 46 years old (avg. 38.5 years), to compare the perceived quality of the stimuli generated by the proposed hybrid-net against those generated by three other speech enhancement systems. These approaches include, (1) the original phase-aware CRN [24], (2) phase-insensitive CRN [23] and (3) the filtered-merged speech with a 4-kHz crossover frequency (e.g. the manual approach from section 2). We follow the same experimental procedure as described before, unless stated otherwise. After the practice stage, there are 60 trials (i.e., 10 repetitions $\times$ 2 SNRs $\times$ 3 conditions) randomly ordered in the main experimental stage and the entire online experiment took less than 30 minutes for each participant to complete. All participants self-reported to have normal hearing, are native speakers of American English, and were in a quiet environment during the listening experiment.

3.3. Results

3.3.1. Subjective results

Figure 3 shows the d' values at each crossover frequency, for each pair of comparisons (i.e., phase-insensitive model, denoted as 'Mag.'; filtered-merged speech at 4000 Hz, denoted as '4 kHz'; and phase-aware model). Note that all models are compared against the stimuli generated by the proposed hybrid-net and a higher d' value (above 0) indicates that stimuli gener-

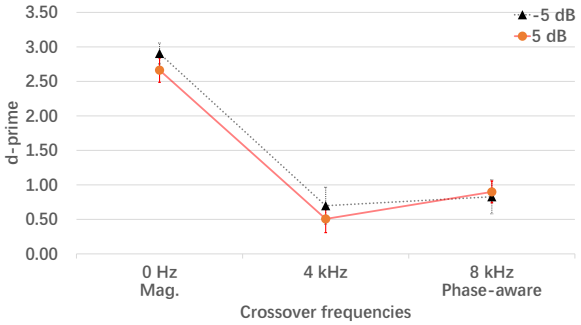

Figure 3: *d' values across different models. Error bars indicate the ± standard errors. Two background SNRs (before enhancement) are included.*

Table 1: *Network statistics of speech enhancement systems.*

	Network Statistics		
System	# of Param. (M)	MACs (G)	Inference Speed (ms)
Phase-insensitive CRN [23]	17.19	50.87	**8.7**
Phase-aware CRN [24]	17.45	59.22	9.0
Hybrid-net	**9.46**	**36.28**	14.2

ated by the proposed hybrid-net are more preferred. The average d' value is above zero for all comparison models, suggesting that the proposed hybrid-net produces enhanced speech with better perceived quality. This is confirmed by two-tailed t-tests showing significant differences between the d' values and zero [$t(19) = 18.80$, $p < .001$ for hybrid-net versus phase-insensitive enhanced speech; $t(19) = 3.04$, $p = .007$ for hybrid-net versus filtered-merged speech at 4000 Hz; and $t(19) = 5.41$, $p < .001$ for hybrid-net versus full-band phase-sensitive enhanced speech]. Note that listeners prefer the hybrid-net enhanced speech even compared to the full-band phase-aware CRN.

A repeated measures ANOVA was conducted on d' value scores and the results show a significant main effect of different comparison models [$F(2, 38) = 101.67$, $p < .001$, $\eta_p^2 = .843$]. There are no significant effects of background SNR [$F(1, 19) = .691$, $p = .416$, $\eta_p^2 = .035$], neither are there any significant interactions between comparison models and background SNR [$F(2, 38) = .61$, $p = .549$, $\eta_p^2 = .031$].

3.3.2. *Computational efficiency*

We further determine whether the hybrid enhancement framework reduces the computational resources (e.g., model size, computations involved) compared to the original full-band phase-aware CRN. We present the model size (i.e., number of parameters), multiply-accumulate (MAC) operations [1] and inference speed (i.e., average running time for processing 1s of audio input) as metrics for model complexity in Table 1. The inference speed is measured using a single Nvidia Tesla V100 GPU, where we set the batch size to 1.

The proposed hybrid-net is approximately half the size (54.2%) of the original phase-aware CRN (i.e., 9.46 M vs. 17.45 M) and achieves a relative 38.7% reduction in MACs (i.e., 36.28 G vs. 59.22 G). The results here suggest that the proposed hybrid-net has better compatibility for low-resource devices, such as digital hearing-aids. However, the inference speed for the proposed hybrid-net is slower than other speech enhancement systems. This is likely caused by the two computation flows and their interactions for the low-frequency and high-frequency processing. This gap could be potentially alleviated by optimizing the parallel process between the two computation flows.

[1] THOP, https://github.com/Lyken17/pytorch-OpCounter

4. Discussion

Experimental results demonstrate that estimating phase is mostly important at low frequency regions for human perception of speech quality for NH listeners. One possible explanation is that speech-quality judgement is, at least partially, related to pitch perception in human listeners. The spectral and TFS cues at low frequencies are much stronger pitch cues than the temporal envelope cues at high frequencies, and they are easily degraded by the presence of low-frequency phase distortions. Therefore, phase distortions at low frequencies may be associated with more noticeably poorer pitch salience than at high frequencies.

We have also noticed that the proposed hybrid-net achieves even better performance than the full-band phase-aware model (i.e., as illustrated in Figure 3). We postulate that this could be caused by the more accurate estimation of phase component at the lower-frequency region with hybrid-net compared to the full-band phase-aware model. As the proposed hybrid-net has a specific sub-network handling phase estimation at low-frequency regions, compared to the full-band phase-aware CRN that estimates the full-band phase components across entire spectral regions. It is possible that low-frequency components are weighted higher for speech quality judgement, therefore leading to better perceived quality for hybrid-net.

5. References

[1] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[2] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *ICASSP*. IEEE, 2016, pp. 5220–5224.

[3] A. Li, M. Yuan, C. Zheng, and X. Li, "Speech enhancement using progressive learning-based convolutional recurrent neural network," *Applied Acoustics*, vol. 166, p. 107347, 2020.

[4] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust asr," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1778–1787, 2020.

[5] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, D. S. Williamson, and D. Yu, "Multi-channel multi-frame ADL-MVDR for target speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3526–3540, 2021.

[6] J. Li *et al.*, "Densely connected multi-stage model with channel wise subband feature for real-time speech enhancement," in *ICASSP*. IEEE, 2021, pp. 6638–6642.

[7] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[8] M. Krawczyk-Becker and T. Gerkmann, "An evaluation of the perceptual quality of phase-aware single-channel speech enhancement," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. EL364–EL369, 2016.

[9] Z. Zhang, D. S. Williamson, and Y. Shen, "Investigation of phase distortion on perceived speech quality for hearing-impaired listeners," in *INTERSPEECH*, 2020, pp. 2512–2516.

[10] E. W. Healy, K. Tan, E. M. Johnson, and D. Wang, "An effectively causal deep learning algorithm to increase intelligibility in untrained noises for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 149, no. 6, pp. 3943–3953, 2021.

[11] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement," in *INTERSPEECH*, 2020, pp. 2472–2476.

[12] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1829–1843, 2021.

[13] C. V. Pavlovic, "Derivation of primary parameters and procedures for use in speech intelligibility predictions," *The Journal of the Acoustical Society of America*, vol. 82, no. 2, pp. 413–422, 1987.

[14] ANSI, "Methods for calculation of the speech intelligibility index," *American National Standard Institute*, 1997.

[15] E. W. Healy, S. E. Yoho, and F. Apoux, "Band importance for sentences and words reexamined," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 463–473, 2013.

[16] Y. Shen, D. Yun, and Y. Liu, "Individualized estimation of the speech intelligibility index for short sentences: Test-retest reliability," *The Journal of the Acoustical Society of America*, vol. 148, no. 3, pp. 1647–1661, 2020.

[17] R. A. DePaolis, C. P. Janota, and T. Frank, "Frequency importance functions for words, sentences, and continuous discourse," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 4, pp. 714–723, 1996.

[18] L. L. Wong, A. H. Ho, E. W. Chua, and S. D. Soli, "Development of the cantonese speech intelligibility index," *The Journal of the acoustical society of America*, vol. 121, no. 4, pp. 2350–2361, 2007.

[19] C. M. McKay, H. J. McDermott, and R. P. Carlyon, "Place and temporal cues in pitch perception: Are they truly independent?" *Acoustics Research Letters Online*, vol. 1, no. 1, pp. 25–30, 2000.

[20] T. Green, A. Faulkner, and S. Rosen, "Spectral and temporal cues to pitch in noise-excited vocoder simulations of continuous-interleaved-sampling cochlear implants," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2155–2164, 2002.

[21] M. S. Osmanski, X. Song, and X. Wang, "The role of harmonic resolvability in pitch perception in a vocal nonhuman primate, the common marmoset (callithrix jacchus)," *Journal of Neuroscience*, vol. 33, no. 21, pp. 9161–9168, 2013.

[22] R. Meddis and M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i: Pitch identification," *The Journal of the Acoustical Society of America*, vol. 89, no. 6, pp. 2866–2882, 1991.

[23] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in *Interspeech*, 2018, pp. 3229–3233.

[24] ——, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2019.

[25] E. Rothauser, "IEEE recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[26] A. J. Spahr, M. F. Dorman, L. M. Litvak, S. Van Wie, R. H. Gifford, P. C. Loizou, L. M. Loiselle, T. Oakes, and S. Cook, "Development and validation of the AzBio sentence lists," *Ear and hearing*, vol. 33, no. 1, p. 112, 2012.

[27] A. Varga and H. J. 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.

[28] K. J. Piczak, "ESC: dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[30] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *ICLR*, 2016.

[31] M. Chait, D. Poeppel, and J. Z. Simon, "Neural response correlates of detection of monaurally and binaurally created pitches in humans," *Cerebral cortex*, vol. 16, no. 6, pp. 835–848, 2006.

[32] A. E. Milne, R. Bianco, K. C. Poole, S. Zhao, A. J. Oxenham, A. J. Billig, and M. Chait, "An online headphone screening test based on dichotic pitch," *Behavior Research Methods*, vol. 53, no. 4, pp. 1551–1562, 2021.

[33] N. A. Macmillan and C. D. Creelman, *Detection theory: A user's guide*. Psychology press, 2004.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ICLR*, 2015.