Reconstructing sound from auditory spiking trains

Po-Ting Bertram Liu (bert.liu@citi.sinica.edu.tw)
Research Center for Information Technology Innovation, Academia Sinica,
No. 128 Academia Road, Section 2, Nankang, Taipei 115, Taiwan

Abstract: Link to the sound demo page.

Auditory periphery encodes sound stimuli into spike trains and interacts with the higher auditory system; however, decoding activities in the auditory periphery remains under-explored in computational neuroscience. In this study, we aim to reconstruct the acoustic stimuli from the spike trains they elicit. To this end, the decoding models must handle the stochastic responses of auditory nerve fibers (ANFs) and compensate for the adaptations by the highly non-linear and bidirectional interactions in the pathways. We proposed a deep artificial neural network (DANN)-based speech synthesis models to decode the spike trains of ANFs' responses. Our model achieved averaged PESQ and SSIM scores of 4.0969 and 0.9225, respectively. Furthermore, in the generalization test, our model performs well on unseen datasets, including VCTK, MCDC8, and excerpts of single musical instruments. In conclusion, our model reconstructs speech with high fidelity from neuronal spiking activities in human peripheral auditory pathways, and the model effectively compensates for any nonlinear and dynamical Acoustic Reflex (AR) and Medial OlivoCochlear Reflex (MOCR) effects.

Keywords: Auditory periphery, Sound reconstruction, Decoding, Spiking activity

Introduction

Auditory periphery encodes sound stimuli into spike trains and interacts with the higher auditory system through afferent and efferent pathways. Various computational models for the auditory periphery have been constructed (Meddis, Lecluyse, Clark, Jürgens, Tan, Panda, & Brown, 2013), and the output at the level of auditory nerve fiber (ANF) is referred to as the auditory *neurogram*, which consists of the simulated ANFs' spike counts across different channels. ANFs have been shown experimentally to exhibit phase-locking ability to sound stimuli below a few kHz (Meddis & Hewitt, 1991). Thus, the auditory neurograms contain place- and time-coded information. In this study, our goal is to reconstruct the acoustic stimuli from the spike trains of ANFs.

Reconstructing sound waves from peripheral auditory activities was first described in an abstract (Rudnicki, Zuffo, & Hemmert, 2012). Their decoding technique was Multi-Layer Perceptron (MLP). They developed a two-stage algorithm to reconstruct high frequency signals. 1) Spike trains were converted to a spectrogram by 51 MLPs. 2) Spectrogram was transformed to an acoustic signal using an iterative method. Recently, Liu, Stohl, Lopez-Poveda, & Overath(2024) also developed a two-stages decoding model by first using neural networks to convert peri-stimulus time histogram of simulated ANFs to acoustic features, and then recovering sound via WORLD speech synthesizer (Morise, Yokomori, & Ozawa, 2016).

The aforementioned models heavily rely on traditional vocoders which do not take neurograms as their input, and suffer from some other limitations (Paul, Pantazis, Stylianou, 2020). In the present research, we aim to reconstruct the sound stimuli directly from auditory neurograms and believe that this approach may have broad applications such as hearing loss simulation.

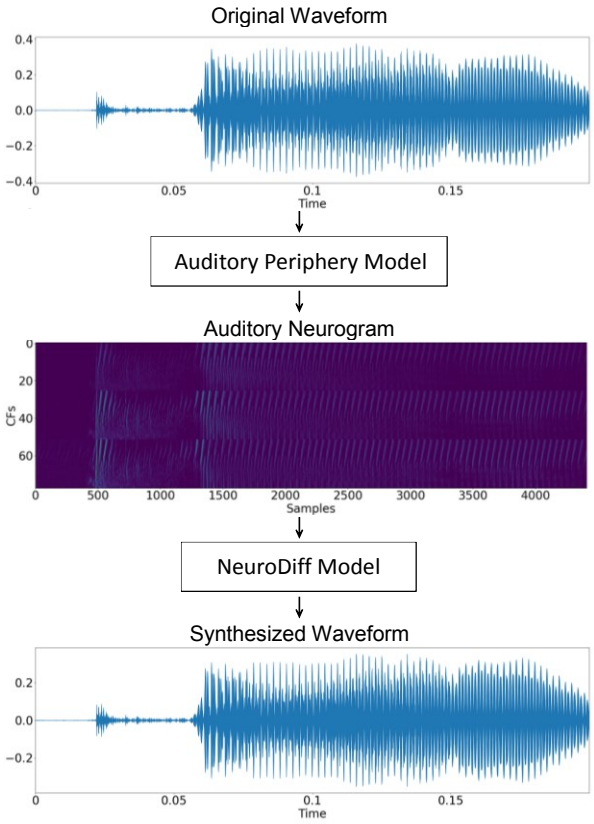

Fig. 1. Overview (Time duration = 200 ms) Characteristic frequencies (CFs).

Datasets and Methods

Auditory neurogram generation

Auditory neurograms were computed by Meddis (2013) Matlab Auditory Periphery (MAP) with the normal hearing condition. The MAP simulated thirty thousand ANFs, which is the number of ANFs in the normal human ear, at high-, medium-, and low-spontaneous firing rates (H-, M-, L- SRs). The ANFs received input from 26 cochlear filters with characteristic frequencies (CFs) between 70 and 8,000 Hz, and the filters were equally spaced on the Equivalent Rectangular Bandwidth (ERB) scale. The MAP also simulated the

Acoustic Reflex (AR) and Medial OlivoCochlear Reflex (MOCR) in the efferent pathways. The stimuli consisted of 90 minutes of speech from the LJSpeech dataset and were presented at the 70-dB sound pressure level when computing the ANFs' responses. The spike trains produced by the ANFs were sampled at 22.05k Hz. The max attenuation levels of AR was 20dB with 10ms latency. For MOCR parameters, MOCR was with 10ms latency and with no tonic attenuation. The spiking rate threshold of MOCR was 65 spikes/s on HSR neurograms. The lowest level of MOCR attenuation was above -35 dB.

Preprocessing and Training Setting

The simulated normal-hearing ANFs' activities with all three SRs were summed into 26 channels per SR to form of the neurograms. Thus, the neurograms had limited spectral and relatively good temporal resolutions and were also affected by AR and MOCR adaptations.

Decoding model: NeuroDiff

Here, we proposed NeuroDiff, a diffusion DANN-based vocoder models (Ho, Jain, & Abbeel, 2020), to reconstruct the sound stimuli from the spike trains of ANFs' responses. The proposed models had neurograms as input and raw waveform as output. Sampling rates and depth bits of this system were at 22,050 Hz and 16-bit, respectively. The loss function of our NeuroDiff model was L1-distance.

Evaluation Metrics

To evaluate the performance of proposed model, the objective metrics were Structural Similarity Index Measure (SSIM), Multi-Resolution Short-Time Fourier Transform (STFT), and Mean Square Error (MSE) between the spectrograms of original and reconstructed signals, and Perceptual Evaluation of Speech Quality (PESQ). Spectral Convergence (SC) was for the accuracy of phase reconstruction and computed between the original and reconstructed signals. STFT was chosen to address the fine-structure of synthesized waveforms. SSIM was chosen to address envelop accuracy. PESQ was computed to measure the overall speech quality.

To address the over-fitting problem, we tested the generalizability of our model by unseen stimulus. Therefore, we prepared test sets, which have four types of stimulus, including unseen British speech of female speaker 's5', in another dataset, VCTK, Mandarin speech (MCDC8 dataset), and excerpts of single musical instruments in Medley-solos-DB, e.g. piano for evaluating acuity of formants structure and violin for testing the reconstruction accuracy of the envelope and vibrato. Twenty samples were selected for testing from all the datasets.

Discussion

We proposed NeuroDiff model, which is the first model that solved the sound reconstruction problem with high fidelity from auditory neurograms. NeuroDiff model can compensate the non-linear effect by adaptation by hair cells, AR and MOCR in the auditory pathways. According to the SSIM and STFT scores in Table 1, our model can generate sound with acuity fine-structure of temporal-spectral information, and compensate the non-linear effect. Therefore, we proposed the first model that solved this long-standing unresolved problem in computational auditory neuroscience. Moreover, according to the generalization test results, NeuroDiff model still has good generalizability on unseen datasets.

Conclusion

Our NeuroDiff model can reconstruct speech with high fidelity from neuronal spiking activities in human peripheral auditory pathways. Despite the limited place- and relatively good time-coded information and dynamic adaptations in spike trains of ANFs, the sounds reconstructed using NeuroDiff models still preserve the speech without noticeable noise or artifacts. Moreover, our NeuroDiff model can effectively compensate for any nonlinear and dynamical AR and MOCR effects.

Table 1: Results of test set and generalization test. ↑ indicates the higher the better, while ↓ indicates the lower the better.

Model	NeuroDiff				
Test-sets	LJSpeech	VCTK	MCDC8	Medley-solos-DB	
	Test-set	British s5	20 samples	piano-test	violin-test
SSIM ↑	0.9225	0.8693	0.9296	0.7687	0.8668
MSE ↓	0.0021	0.0047	0.0052	0.0171	0.0144
STFT ↓	0.9518	1.8579	1.8997	2.3945	1.9829
PESQ ↑	4.0969	3.5433	3.7178	-	-
SC ↓	0.1911	0.4925	0.6563	1.4091	2.1022

Acknowledgments

This work was supported by NSTC grants #111-2423-H002-012.

References

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*.

Liu, J., Stohl, J., Lopez-Poveda, E. A., & Overath, T. (2024). Quantifying the Impact of Auditory Deafferentation on Speech Perception. *Trends in Hearing*.

Meddis, R., & Hewitt, M. J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *The Journal of the Acoustical Society of America*.

Meddis, R., Lecluyse, W., Clark, N. R., Jürgens, T., Tan, C. M., Panda, M. R., & Brown, G. J. (2013). A computer model of the auditory periphery and its application to the study of hearing. *Basic aspects of hearing: physiology and perception*.

Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*.

Paul, D., Pantazis, Y., & Stylianou, Y. (2020). Speaker conditional WaveRNN: Towards universal neural vocoder for unseen speaker and recording conditions. *INTERSPEECH 2020*.

Rudnicki, M., Zuffo, M. K., & Hemmert, W. (2012). Sound decoding from auditory nerve activity. *Front. Comput. Neurosci. Conference Abstract: Bernstein Conference 2012*.