A Differentiable Model of Auditory Processing

Leslie Li
Department of Computer Science,
Program in Neuroscience and Cognitive Science, & Department of Linguistics
University of Maryland

While models in audio and speech processing are becoming deeper and more end-to-end, they suffer from expensive training and a lack of robustness. In this project, we draw from a m model of hearing and present a differentiable auditory processing model, combining traditional signal processing approaches with deep frameworks. We showcase applications of this model in spectrogram inversion, unsupervised learning and speech enhancement. Results showed promising speed advantage and audio quality, even with less than an hour of training data. We also discuss the potential application of this model on audio personalization and hearing disorders.

Introduction

In the last few years, audio and speech technology has moved towards deep and end-to-end methods. While deep learning achieved great advances in automatic speech recognition, source separation, speech enhancement and more, the deep and data-driven methods are often constrained by available data and overfitting of the models. Specifically, training large models is expensive in terms of both hardware and datasets. For example, to pretrain a speech model for downstream tasks, hundreds of hours of speech are needed (see review in Mohamed et al., 2022). Compared with this, the human auditory system can perform many tasks such as source separation and emotion detection effortlessly. Additionally, large, end-to-end audio models are vulnerable to adversarial attacks, with performance dropping drastically when small modifications that are imperceptible to humans are made to the audio clips (Wu et al., 2022).

Looking back at the history of audio and speech processing, many model-based methods have been developed where human perception is relevant. The spectrogram, as a famous example, is computationally similar to the human cochlea, where sound in time domain is transformed into energy along various frequency bins as well as time. While human ears perform such decomposition through physical means, the spectrogram is commonly computed using fast Fourier transform. As another example, the mel frequency cepstral coefficient (MFCC), after its development for engineering purposes, was found to be conceptually similar to the cortical analysis of sound (see discussion in e.g. Meyer & Kollmeier, 2011). While MFCCs are obtained by applying another Fourier transform on the spectrogram and finding its principal components, auditory processes in the brain also represent sound by its spectral modulation (i.e., scale), which can be modeled by a second Fourier transform along the frequency axis. These connections between the engineering tools and neuroscience motivates us to explore more engineering applications of models in auditory neuroscience.

Although methods rooted in signal processing are falling out of fashion in lieu of deep learning, integrating them into deep learning models may improve the data efficiency and robustness. The model of auditory cortical processing, for exam-

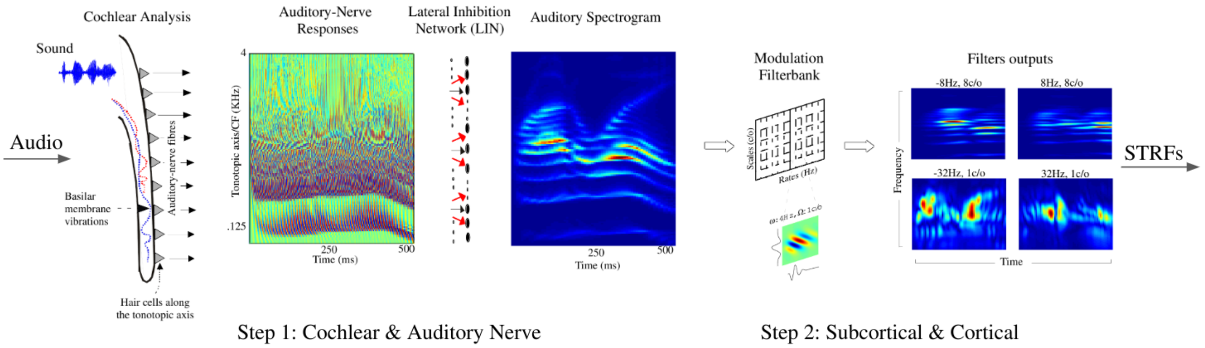

Figure 1. The differentiable auditory processing model in two stages. In the first stage, audio signal is transformed into auditory spectrogram. In the second stage, auditory spectrogram is transformed into STRFs. Image materials are taken from Elhilali (2004).

ple, has been directly applied for supervised tasks such as speech detection (Mesgarani, Slaney, & Shamma, 2006) and sound segregation (Elhilali & Shamma, 2008). The concept of modulation-based feature detection was also utilized to build Gabor-based filters to detect spectral and temporal modulation in the audio signal. Models using such Gabor-based features outperformed the MFCC counterparts and showed more robustness to additive noise as well as speaking styles (Meyer & Kollmeier, 2011). Recently, the Gabor-based features have also been combined with differentiable approaches to increase performance and robustness in voice type discrimination (Vuong, Xia, & Stern, 2020), speech enhancement (Vuong, Xia, & Stern, 2021), and music tagging (Ma & Stern, 2022). In these studies, The Gabor-based features were seen as one convolutional layer of a neural network, with the convolution filters constrained to only extract Gabor-like spectral and temporal modulations.

In this project, we aim to combine the auditory signal processing methods and deep learning more closely towards a differentiable model of auditory processing. Our model is composed of two stages, an earlier stage covering the ear (cochlea) and early auditory nerve components, and a later stage covering subcortical and cortical auditory processing. The code and audio demonstrations are available at `https://github.com/smiledra/diffaud`.

The rest of the paper is organized as the following. First, we give a more detailed introduction of our differentiable auditory processing model. Then, in the next three sections, we introduce three applications of the differentiable model of auditory processing. First, we use the forward model of auditory spectrogram to solve the inversion problem — to get the audio waveform given an auditory spectrogram. Secondly, we build an autoencoder that combines the differentiable model with a neural network, and reconstructed auditory spectrogram and even audio from STRF representations. Lastly, we demonstrate some preliminary results from applying this model to speech enhancement and towards source separation. We then discuss the broader contributions and some future directions.

The differentiable auditory processing model

In this section, we give a brief introduction of the forward model of auditory processing. We refer readers to e.g. Chi, Ru, and Shamma (2005) and Elhilali (2004) for more details on this model.

Step I. Auditory Spectrogram. The first stage of our model converts audio waveforms into auditory spectrograms. Spectrograms are representations of audio that display the energy in sound along different frequency and time bins. Despite the popularity of end-to-end methods, spectrograms still remain common as the first step of processing for many audio and speech models.

The extraction of a spectrogram typically involves short-time Fourier transform (STFT), from which the magnitude is kept and phase discarded. This poses a nontrivial problem if audio were to be reconstructed from the spectrogram. Since phase information is lost in the spectrogram, an infinite amount of audio patterns only differing in phase could all generate the same spectrogram.

The auditory spectrogram is conceptually similar to the spectrogram, but the signal is extracted differently, with several potential advantages. While a traditional spectrogram performs STFT on windowed audio, the extraction of the auditory spectrogram follows a few stages that are based on early auditory processing in the human ear. As shown in Figure 1, the audio signal is passed through a bank of constant-Q filters, which corresponds to the biological process in the basilar membrane. Then, the signal is passed through half-wave rectifying, first difference, and leaky integration that mimics transduction (changing the signal from physical vibration to neural signal) and downstream processes that take place in the auditory nerve. Similar to the traditional spectrogram, the auditory spectrogram also displays energy along the frequency and time axes. However, the auditory spectrogram represents information in a way that more closely matches human perception and auditory processing.

Step II. Spectro-Temporal Receptive Fields (STRFs). After a spectrogram is obtained, the brain continues to extract relevant features from the spectrogram for higher-level information, which eventually leads to complex tasks such as comprehension of words and appreciation of music. In neuroscience, it has been proposed that humans and other animals alike analyze audio in terms of its spectral and temporal modulations. This is conceptually equivalent to applying 2D band-pass filtering to the auditory spectrogram. Such a band-pass filter can be described by two parameters: the scale (Ω, or spectral modulation), which is the pass band along the frequency axis, and the rate (ω, or temporal modulation), the pass band along the temporal axis. For one single neuron with a tuned frequency f at a given time t, its activation $r_f(t)$ at a certain time can be then characterized as a function of frequency, scale, and rate:

$$r_f(t) = \mathrm{STRF}(f, t; \Omega, \omega) \tag{1}$$

where STRF(.) stands for the 2D band-pass filter. With a bank of such filters, the cortical analysis of a sound can then be represented as a four-dimensional cube along frequency, time, scale, and rate. The linearity of the STRF model makes it easily interpretable: a high scale (approx. 4–6 cycles/octaves) corresponds to pitch-related information, as the fine structure along the frequency axis often corresponds to multiples of the fundamental frequency; a low scale (approx. 0–3 cycles/octaves) often corresponds to information such as vowel identity and musical timbre, as the coarse structure along the frequency axis corresponds to the transfer function of the vocal tract or musical instrument (see Elliott & Theunissen, 2009 for in-depth discussions). In temporal modulation, a slower rate (e.g., around 4 Hz) may correspond more to syllable-level information, and faster rates (e.g. 20–30 Hz) corresponds more to the temporal fine structure.

Differentiability. The pipeline from waveform to cortical representations using STRF features was made fully differentiable under JAX (Bradbury et al., 2018). For this project, we use back-propagation to update parameters including the scale and rate in the cortical filterbank. In Case I, we also use differentiability to perform iterated updates for model inversion.

Case I. Inversion of Auditory Spectrogram

The problem of spectrogram inversion has long been of interest. Recently, neural networks have been trained to solve this problem in a black-box manner. For example, Kumar et al. (2019) trained a general adversarial network with more than 4 million parameters on 20+ hours of speech. Additionally, training needs to be done separately for different types of tasks (e.g. speech synthesis vs. music generation). On the other hand, traditional methods for spectrogram inversion that do not require training also exist. For example, the Griffin-Lin algorithm (Griffin & Lim, 1984) was proposed to obtain the audio through iterated update. Similarly, a iterative algorithm has been proposed for the auditory spectrogram (Chi et al., 2005), and was applied to audio resynthesis from its cortical representations (Zotkin, Chi, Shamma, & Duraiswami, 2005). Furthermore, a comparison between iterative methods on the two types of spectrograms was made in Decorsière, Søndergaard, MacDonald, and Dau (2014), arguing that not only is auditory spectrograms easier to invert, but also that bringing human perception into the loop of iterative methods can even increase performance of traditional spectrogram inversion. In this section, we use the first half of our differentiable model to perform spectrogram inversion.

Method. We implemented the forward model in JAX, and used automatic differentiation to obtain the gradients for iterative updates. Our implementation conceptually adheres to the one described in (Chi et al., 2005), with modifications in parallelization and vectorization to utilize the GPU speedup available ot JAX. Due to the nature of the algorithm, no training is necessary for spectrogram inversion. Testing is done using clips in the Common Voice corpus (Ardila et al., 2019). All audio clips are sampled at 16 kHz.

The auditory spectrogram can be extracted with different resolution, compression parameters, among other customizations. In this project, we focus on parameters that is commonly used for modeling in auditory neuroscience. We used 128 frequency channels, and no compression (i.e. identity) at the outer hair cell stage.

For the iterative update algorithm, we initialized the audio at zero and used the forward model to obtain the gradient of the loss with respect to the audio. The audio is then updated using gradient descent with learning rate 0.025. After every ten iterations, we plot the spectral convergence between the reconstructed audio $\hat{x}$ and the ground truth x:

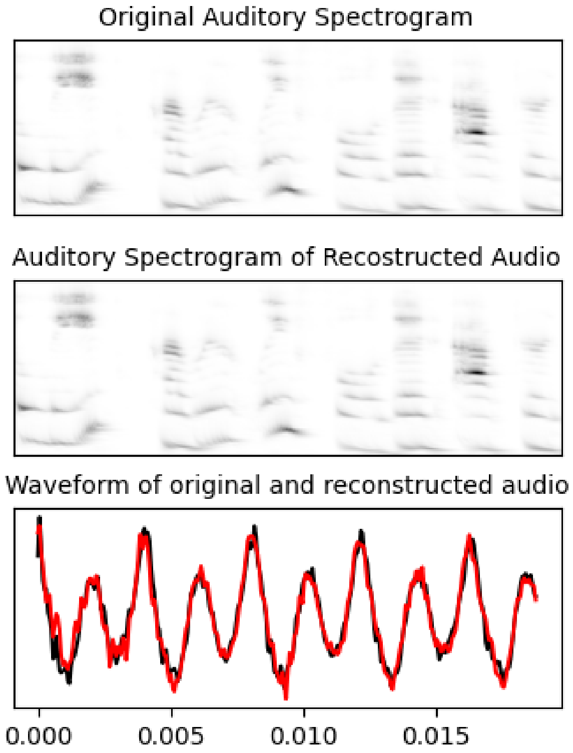

Figure 2. Top: the auditory spectrogram of the ground-truth audio. Middle: the auditory spectrogram of the reconstructed audio. Bottom: a close look on the audio of a fraction (0.188 s, 300 frames) of a vowel; the black line represents the original audio, and the red line represents the reconstructed audio.

$$C = \frac{||AS(\hat{x}) - AS(x)||_F}{||AS(x)||_F} \tag{2}$$

where $AS(.)$ stands for the auditory spectrogram transform and F stands for the Frobenius norm.

Results. The reconstructed audio generally has high quality, although some parts of the audio sounded artificial. When the auditory spectrogram of the original and reconstructed audio were plotted, they look almost identical (see Figure 2). Additionally, as the auditory spectrogram preserves some low-frequency phase information, the phase of such low frequencies was restored in the reconstructed audio. As shown in the bottom panel

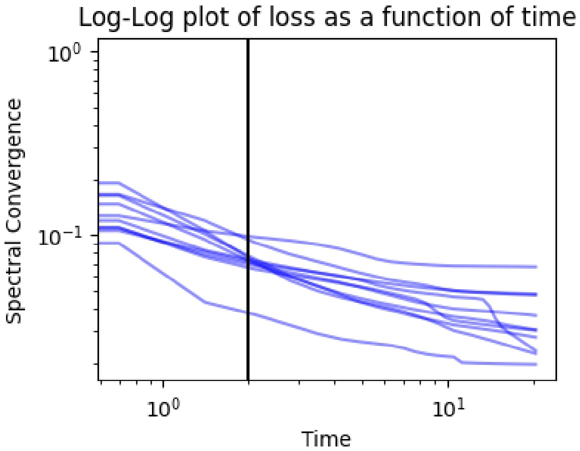

Figure 3. The loss (spectral convergence) of the reconstructed audio as a function of time, plotted in log-log space. Each blue line represents one of ten 2-second audio tokens. The black vertical line marks the two-second threshold, which is the bound for real-time conversion.

of Figure 2, the slower, dominating frequency of the reconstructed audio matches the original almost exactly, and the higher frequencies deviated locally. This showcased one advantage of the auditory spectrogram, that some phase information (which might be the ones that are important to human perception) are preserved.

In terms of speed, the spectrogram inversion algorithm was able to reconstruct intelligible audio in real time (approx. 25 iterations). If more iterations are run, the reconstructed audio can approach the quality of the original audio. In the following sections, we set the number of iterations at 300 iterations.

Discussion. Using the auditory spectrogram for spectrogram conversion could have the following advantages compared with the spectrogram. Specifically, the phase delays in the constant-Q filterbank creates different delays in different channels, which allows the auditory spectrogram to retain some phase information. While this is known to happen in early human auditory processing as well, the phase information at this resolution is completely lost in the spectrogram. Therefore, the auditory spectrogram allows for extra information to be preserved for inversion. Additionally, as phase information for higher frequencies are lost in human auditory processing early on, it is possible that disturbing the phase information in the higher frequencies do not lead to much difference in human perception. Therefore, inverting the auditory spectrogram could be more faithful to human perception and therefore more robust.

Case II. Speech Autoencoders

In the previous section, we focused on only the cochlear part of the model. Now, we extend to the cortical part the model in order to explore the complex cortical feature space for audio processing. Here, as a first step, we present autoencoders trained using STRFs. Specifically, a small convolutional neural network (CNN) was trained to reconstruct the auditory spectrogram from STRF representations, and training occured jointly for both the STRF step and CNN reconstruction. This autoencoder architecture has two motivations. Firstly, it has been of interest to reconstruct sound from cortical representations. In neuroscience, speech can be reconstructed from single-neuron recordings from ferrets, using a reverse filtering approach based on the STRF model (Mesgarani, David, Fritz, & Shamma, 2009). Such reconstruction can also have engineering applications. For example, sounds can be manipulated in the STRF space and then resynthesized into waveform. As a result, manipulations like changing pitch, timber, or morphing two sounds together are possible (Zotkin et al., 2005). With respect to the current project, a differentiable model and related deep approaches give the potential to perform higher-quality synthesis and manipulations. Secondly, using a convolutional architecture with Gabor feature detectors, Vuong et al. (2021) has designed a linear autoencoder on speech. In their work, the autoencoder was run on log mel-spectrograms instead of the auditory spectrogram, and the model architecture was computationally similar, but implemented spectrotemporal filters

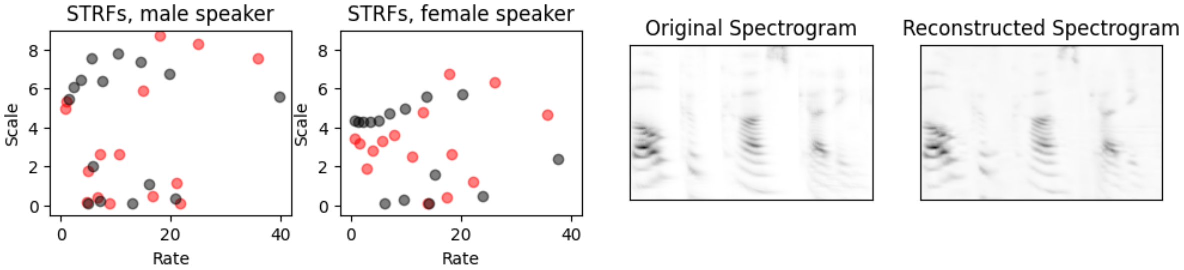

Figure 4. Results of the STRF-based autoencoder. Left: the learned STRF parameters. Each dot represents one STRF parameter, with its spectral modulation plotted along the y-axis, and temporal modulation along the x-axis. Red dots indicate downward-shifting STRFs (i.e., frequency modulation moves downwards along time), and black dots indicates upward-shifting STRFs. STRFs parameters are plotted separately for the two models trained on different speakers. Right: one example of the auditory spectrogram reconstructed by the autoencoder, as well as the original. The male speaker model was used.

using cross-correlation with Gabor features instead of bandpass filtering as in the neuroscience studies. The autoencoder results indicate that the trained autoencoders were able to perform near-perfect reconstruction of the log mel-spectrogram. Additionally, the authors found that successful models learned spectrotemporal filters that are relatively low in both spectral and temporal modulation, similar to what has been observed in humans. In the current study, it is worth replicating these results in our model, which is implemented more closely adhering to neuroscience modeling. Additionally, with the results from the previous section on auditory spectrogram inversion, it is possible to reconstruct STRF representations all the way to audio.

Method. In our model, we used the cortical part of the differentiable auditory processing model to extract STRF representations from auditory spectrograms, and a small CNN to reconstruct the auditory spectrogram. Following Vuong et al. (2021), we used 30 STRFs. Half of the STRFs are upward-tilting and half are downward-tilting. The STRFs are initialized uniformly random between scale $[0, 8]$ and rate $[0, 30]$. The CNN has three layers, with respectively 10, 2, and 1 channels, and 3×3 filters within each layer, and gelu activation function at the end of each layer.

In this paper, we report two models that are respectively trained on one male speaker and one female speaker. The speech data is obtained from the Wall Street Journal corpus (Paul & Baker, 1992), and each speaker contained around 40 minutes of speech material. Each model is trained for 200k steps with the adam optimizer, which jointly optimized the STRF and CNN parameters. The learning rates were initialized at 0.001. For each step, we use a minibatch of 4 samples of one second (200 samples in the auditory spectrogram) each, which is randomly sampled from the training dataset.

Results. The results suggested that the small convolutional neural network was able to fit a spectrogram well with a small amount of training data (less than one hour). One example of spectrogram reconstruction is shown in Figure 4. To reconstruct the audio sample, we also used the auditory spectrogram inversion introduced in the previous section to obtain audio clips corresponding to the spectrogram Audio examples are included in the project website, which sounds nearly identical to the ground truth.

Additionally, we observed that the learned STRF parameters are concentrated in regions low in both spectral and temporal modulation. This happened among all models including those trained during the parameter search that are not included in the results. We also observed some differences between the STRF parameters that are specific to the training speaker. In the model trained on the male speaker, the STRFs that are greater than 3 cycles/octaves in scale are generally concentrated among higher scale regions (5–8 cycles/octaves), while in the model trained on the female speaker, the scales of the STRFs are shifted to lower regions (3–7 cycles/octaves). Considering that scales in this range mostly represents harmonics that are related to pitch, this pattern is expected considering that the lower pitch of the male speaker would lead to more densely distributed harmonics, and therefore higher scale. This suggests that the STRFs are optimized for the specific distribution of the audio material.

Discussion. Our results replicated that of (Vuong et al., 2021). Unlike their model, however, we were not able to obtain satisfying reconstruction using a linear decoder alone. This is expected considering that our STRF feature extraction step is more than a simple convolution with finite-length filteres, and therefore cannot be reconstructed using a single convolutional decoder. Additionally, with the tradeoff between model simplicity and reconstruction quality, we opted for the latter in order to obtain high-quality audio using spectrogram inversion.

Case III. Speech Enhancement

In the previous section, we reported high-quality reconstruction of auditory spectrograms by a CNN from STRF features. A natural extension of the autoencoder structure is whether it can be used in speech enhancement. It has been postulated that STRF features are well-suited for scene analysis and source segregation. For example, STRF-based models have been successful in source segregation (Elhilali & Shamma, 2008). The Gabor-based STRF model has also been applied to voice activity detection (Vuong et al., 2020) and music tagging (Ma & Stern, 2022), tasks that separate different types of audio that differ greatly in temporal modulation (Ding et al., 2017), among other aspects. Furthermore, it has been shown that speech enhancement can benefits from loss not only on spectral domain, but also in STRF domain (Vuong et al., 2021). These successful results motivates us to use the autoencoder structure with STRF feature extractors to perform speech enhancement directly. Specifically, instead of clean auditory spectrograms as input, we corrupted the input spectrograms with additive white noise. The model is then trained to minimize the loss with the clean version of the spectrogram.

Additionally, we trained the autoencoder to perform a even harder task – instead of white, stationary noise, if the noise to be filtered out is another speaker, would the autoencoder be able to reconstruct the spectrogram with the "noise" speaker filtered out? This task is known as the cocktail party problem (Elhilali, 2017), and is much easier in normal-developing humans compared with machines. Particularly, when humans are presented in a cocktail party scene, the STRF responses are found to be stronger for the attended speaker relative to the unattended speaker, likely due to top-down feedback from higher-level cortical regions (Ding & Simon, 2012). While in neural networks that are feed-forward, top-down feedback is harder to implement, we can use the convolutional decoder to learn some speaker-specific selection. Additionally, inspired by the previous section, if STRFs display specificity for e.g. the pitch of the speaker, then allowing flexible STRFs to be fitted may also help in retaining the speech signal from one speaker while removing that of the unattended speaker.

Method. For both speech enhancement problems, we used the same autoencoder architecture, with only a slight increase in the number of channels in the CNN decoder: the three layers here have 10, 5, and 1 channels. Other than the change in the input-output pair, the training paradigm was the same as described in the previous section. For white noise enhancement, we trained models on only the male speaker (i.e. 40 min of speech). Separate models were trained with different signal-to-noise ratio (SNR) in the training data: -10, 0, and 10 dB, respectively.

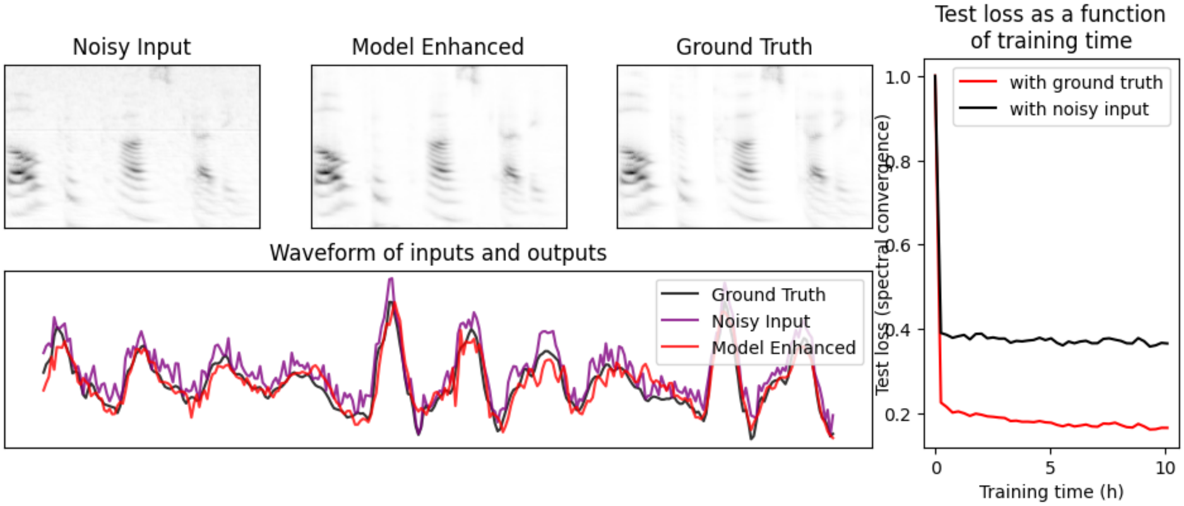

Figure 5. Results on speech enhancement against additive white noise. Top left: the auditory spectrograms of speech with additive noise (Noisy Input), the output of the trained autoencoder (Model Enhanced), and the clean version of the speech (Ground Truth). Bottom left: a zoom in of the waveform of a fraction of a vowel (0.188 s, 300 frames). Right: Test loss as a function of training time. Spectral convergence with the noisy input (black) and with the clean output (red) are plotted. The -10 dB SNR model was used for all plots in this figure.

For the cocktail party paradigm, we trained models on reconstructing the male speaker with the female speaker as noise. We trained separate models with 0, 10, and 20 dB SNR.

Results. As shown in Figure 5, the autoencoder model performed well on the speech enhancement task. The enhanced auditory spectrogram closely resembles the ground truth. By plotting the spectral convergence, we can also observe that the reconstructed spectrogram is much more similar to clean speech than the noisy version. We also found that when all else is held the same, models that were trained on noisier inputs learned to remove noise better, which suggests the possibility that the SNR levels in the training data could be further decreased.

The audio example for the cocktail party paradigm is displayed online. In testing, while the model cannot eliminate the female speaker in the reconstructed speech, the female speaker did become less intelligible and quieter. Of note, although the model is only trained on one male and female speaker pair, when tested on a novel female speaker, the model was also able to perform a similar level of enhancement, selecting the same male speaker from the unseen female speaker. This suggests that the model was able to generalize to some novel speakers.

Discussion. The results suggest great potentials for the STRF-based model to perform speech enhancement. Our autoencoder model has the advantage of being lightweight and requiring little training data. The total number of parameters for the STRF feature extractor is only 60, and approximately 4–5k for the convolutional decoder. The small size of the also makes the required training data small. As a result, this architecture has potential applications in low-resource settings, such as enhancing one particular speaker's speech in a video conference or in AI-based headphone denoising, where only a few minutes of the target speaker speech is available. Additionally, as the encoder only involves filtering and the decoder is fully convolutional, a trained model can be applied to an arbitrary duration of speech, which can be tailored to the hardware's capacity.

General Discussion

In this project, we contributed a differentiable model of auditory processing. Using JAX, we were able to not only combine traditional signal processing with deep learning, but also utilize GPU acceleration in signal processing to achieve spectrogram inversion in real time. Additionally, the spectrotemporal model showed promising results on speech enhancement and source separation.

There are some immediate future directions to be pursued. For one, in the current experiments for speech enhancement, we have performed a very simple task (additive white noise) and a very difficult task (cocktail party problem). While white noise has completely different spectrotemporal modulation properties compared with speech, speech is maximally similar with speech itself. It also seems like the models easily handled white noise but had space for improvement for the cocktail party case. To explore the capacity of the STRF model in speech enhancement, it would therefore be sensible to try something in between, such as music source separation or speech enhancement from other types of noise.

Another potential direction is to perform further enhancement on the spectrogram inversion task. Currently, the inverted audio was very similar to the original, but artifacts can be identified when listening closely. One potential way to further enhance the generated audio is to pass it through a small neural network (e.g., an RNN) that transforms the reconstructed speech into, ideally, the original speech. Further improvment on the audio quality can make this model a better candidate for tasks such as speech enhancement and speech synthesis.

In the long term, this model can also be applied towards audio personalization. For example, given cortical response measurable through e.g. EEG, parameters of the auditory model can be estimated. In this project, we already performed fitting for the STRF parameters. Under the same framework, the cochlear parameters can also be fitted. This includes the filter parameters for each cochlear filterbank in frequency decomposition, amount of cochlear compression, and the decay rate in leaky integration. Obtaining individualized parameters for each listener or the same listener in different listening scenario can help create personalized listening experiences. Additionally, for cochlear-related abnormalities, our model can also help build hearing aid devices that is optimally fit to an individual. As a precursor, Drakopoulos and Verhulst (2023) used a deep model to fit hearing aid parameters to compensate for two different types of hearing loss. While their model stops at the auditory spectrogram, our model may be able to continue to the brain, and therefore use non-invasive brain recording techniques like EEG and MEG to collect response for supervision.

Acknowledgement

This work is supervised by Dr. Ramani Duraiswami. I would like to also thank Drs. Naomi Feldman and Shihab Shamma for helpful discussions and suggestions, and members of the PIRL lab for their technical help and feedback.

References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., … Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., … Zhang, Q. (2018). *JAX: composable transformations of Python+NumPy programs.* Retrieved from `http://github.com/google/jax`

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*(2), 887–906.

Decorsière, R., Søndergaard, P. L., MacDonald, E. N., & Dau, T. (2014). Inversion of auditory spectrograms, traditional spectrograms, and other envelope representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(1), 46–56.

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, *81*, 181–187.

Ding, N., & Simon, J. Z. (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of neurophysiology*, *107*(1), 78–89.

Drakopoulos, F., & Verhulst, S. (2023). A neural-network framework for the design of individualised hearing-loss compensation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Elhilali, M. (2004). *Neural basis and computational strategies for auditory processing* (Unpublished doctoral dissertation). University of Maryland, College Park.

Elhilali, M. (2017). Modeling the cocktail party problem. *The auditory system at the cocktail party*, 111–135.

Elhilali, M., & Shamma, S. A. (2008). A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. *The Journal of the Acoustical Society of America*, *124*(6), 3751–3771.

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS computational biology*, *5*(3), e1000302.

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, *32*(2), 236–243.

Kumar, K., Kumar, R., De Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., … Courville, A. C. (2019). Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, *32*.

Ma, Y., & Stern, R. M. (2022). Learnable front ends based on temporal modulation for music tagging. *arXiv preprint arXiv:2211.15254*.

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2009). Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *Journal of neurophysiology*, *102*(6), 3329–3339.

Mesgarani, N., Slaney, M., & Shamma, S. A. (2006). Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Transactions on audio, speech, and language processing*, *14*(3), 920–930.

Meyer, B. T., & Kollmeier, B. (2011). Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition. *Speech Communication*, *53*(5), 753–767.

Mohamed, A., Lee, H.-y., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., … others (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*.

Paul, D. B., & Baker, J. (1992). The design for the wall street journal-based csr corpus. In *Speech and natural language: Proceedings of a workshop held at harriman, new york, february 23-26, 1992.*

Vuong, T., Xia, Y., & Stern, R. (2020). Learnable spectro-temporal receptive fields for robust voice type discrimination. *arXiv preprint arXiv:2010.09151*.

Vuong, T., Xia, Y., & Stern, R. M. (2021). A modulation-domain loss for neural-network-based real-time speech enhancement. In *Icassp 2021-2021 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6643–6647).

Wu, H., Zheng, B., Li, X., Wu, X., Lee, H.-Y., & Meng, H. (2022). Characterizing the adversarial vulnerability of speech self-supervised learning. In *Icassp 2022-2022 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 3164–3168).

Zotkin, D. N., Chi, T., Shamma, S. A., & Duraiswami, R. (2005). Neuromimetic sound representation for percept detection and manipulation. *EURASIP Journal on Advances in Signal Processing*, *2005*, 1–15.