DERIVING SPECTRO-TEMPORAL PROPERTIES OF HEARING FROM SPEECH DATA

Lucas Ondel[1,3], *Ruizhi Li*[1], *Gregory Sell*[1,2], *Hynek Hermansky*[1,2,3]

[1]Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, USA
[2]Human Language Technology Center of Excellence, The Johns Hopkins University, Baltimore, USA
[3]Brno University of Technology, FIT, IT4I Centre of Excellence, Czech Republic
iondel@fit.vutbr.cz, {ruizhili, gsell2, hynek}@jhu.edu

ABSTRACT

Human hearing and human speech are intrinsically tied together, as the properties of speech almost certainly developed in order to be heard by human ears. As a result of this connection, it has been shown that certain properties of human hearing are mimicked within data-driven systems that are trained to understand human speech. In this paper, we further explore this phenomenon by measuring the spectro-temporal responses of data-derived filters in a front-end convolutional layer of a deep network trained to classify the phonemes of clean speech. The analyses show that the filters do indeed exhibit spectro-temporal responses similar to those measured in mammals, and also that the filters exhibit an additional level of frequency selectivity, similar to the processing pipeline assumed within the Articulation Index.

Index Terms— perception, spectro-temporal, auditory, deep learning

1. INTRODUCTION

Auditory spectro-temporal cortical receptive fields (STRFs) are accepted as parts of higher levels of the mammalian hearing chain [1]. They describe linearized properties of cortical neurons in an auditory cortex, many of which exhibiting selective behavior by enhancing particular parts of the signal spectrum, particularly in terms of their different spectral and temporal modulations. Numbering in hundreds of millions, their existence suggests the capacity of the cortex to provide multiple views of the incoming acoustic signal for a further information extraction.

This frequency selectivity is also consistent with the Articulation Index concept (AI), which postulates extraction of speech messages in independent frequency bands [2]. This ability could account for the resiliency of speech communication in noisy environments, a robustness that multistream automatic speech recognition (ASR) techniques are attempting to emulate by creating multiple parallel processing streams for extraction of information in speech and selectively alleviating the corrupted streams.

Human speech developed after human hearing [3], and so it is likely the properties of human speech developed in order to align with the existing auditory system, as efficient communication would require that the signal and receiver match sufficiently for the transfer of information. As a result, the acoustic properties of speech and the response properties of the human auditory perceptual system are likely tied together in a fundamental way. Based on this observation, one might wonder if characteristics of human hearing such as the modulation-specific STRFs and multistream processing could be derived automatically in a data-driven fashion from the speech itself. Some indications that this might be possible are seen in earlier works [4, 5].

In this work, we explore the hypothesis that data-derived filters learned from speech data will exhibit spectro-temporal and frequency-selective multistreaming behaviors of the auditory cortex. Other work has demonstrated that data-driven training can yield systems that demonstrate human-like peripheral frequency resolution [5, 6, 7, 8, 9, 10, 11, 12, 13] and cortical-like sensitivity to modulations [5, 14, 15, 16], so here we extend this trajectory to explore the spectro-temporal properties. Toward this end, we first examine properties of filters learned in a data-driven fashion to identify the phonetic content of speech. We then replicate perceptual analyses on the learned filters and demonstrate the similarity of their responses to the human auditory system. We further explore the nature of the individual filters and observe patterns of frequency selectivity. In doing so, we demonstrate behaviors in the learned filters that mimic the modulation-selective and multistreaming properties of human auditory processing.

2. EXPERIMENTAL SETUP

We begin with a deep neural net (DNN) architecture that learns to classify context-independent phonemes based on their temporally-central mel frequency spectral frames. The first layer of the network is a convolutional layer that applies two-dimensional spectro-temporal (ST) filters in time (i.e., temporal convolution only). The outputs of these filters are then passed to two additional dense feed-forward layers with intermediate hyperbolic tangent nonlinearities and a final softmax non-linearity. The resulting output predicts the speech class from the 39 context-independent American English phonemes (including silence). However, the primary focus of this work is on the front-end ST filters that result from the training for phoneme classification, rather than the performance of the phoneme classifier itself.

For data to train the DNN, we selected the Wall Street Journal database, due to its clean and mostly well-articulated data. We conducted our experiments on a subset of the Wall Street Journal corpus usually referred as SI-284 [17], which is composed of roughly 37,000 sentences spoken by 284 speakers for a total of about 62 hours of data. 10% of this data was removed from training as a cross-validation set to monitor the progress of the DNN.

Input to our classifiers was provided by 7 Mel-frequency spectral filters [18], with windows of length 20ms extracted every 10ms (resulting in a frame sampling rate of 100Hz). ST filters were also given 600ms of context, resulting in each filter having a size of 7x60. The initial layer of the network had 32 ST filters, and the subsequent dense layers were of size 512. These parameters were selected

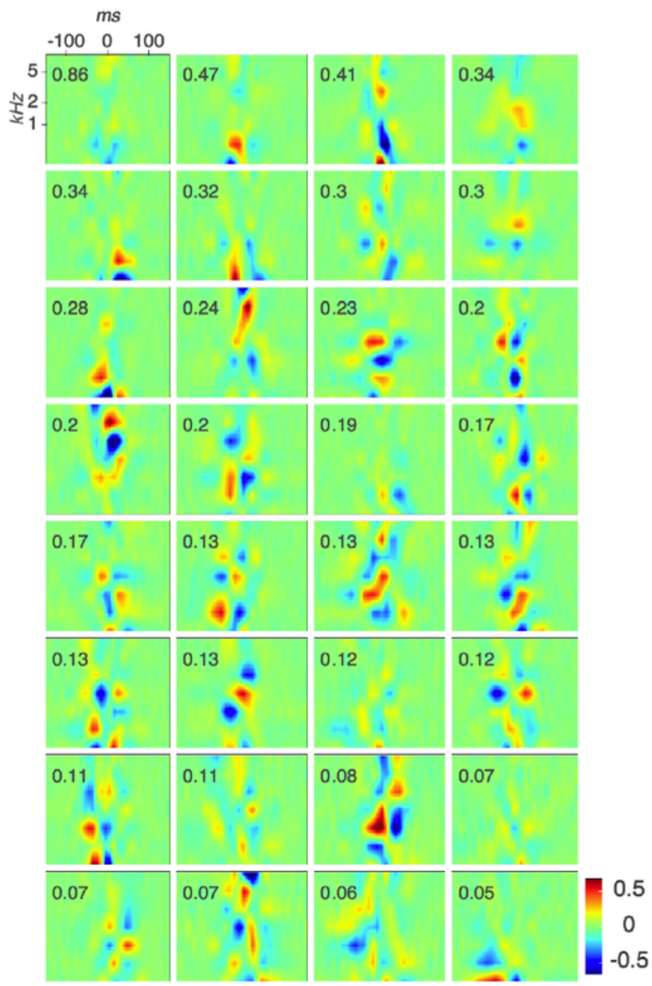

Fig. 1. Impulse response of the data-driven ST filters over time and frequency. Only 300 ms of the ST filters are shown as values near the filters' boundaries were close to zero. Labels in the upper left corner indicate the Mutual Information (MI) in bits between the filtered speech and the phone labels for each ST filter.

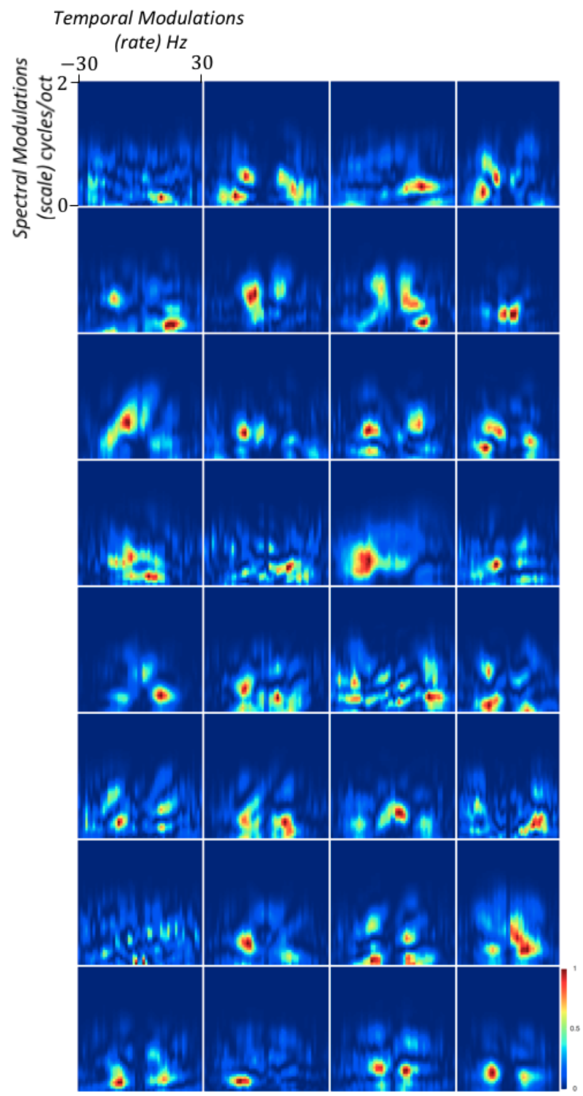

Fig. 2. Responses to ripple analysis for each filter, which are laid out in the same orientation as Fig. 1. In many cases, the responses show a highly localized selectivity in their spectro-temporal behavior, indicated by small regions of intensity.

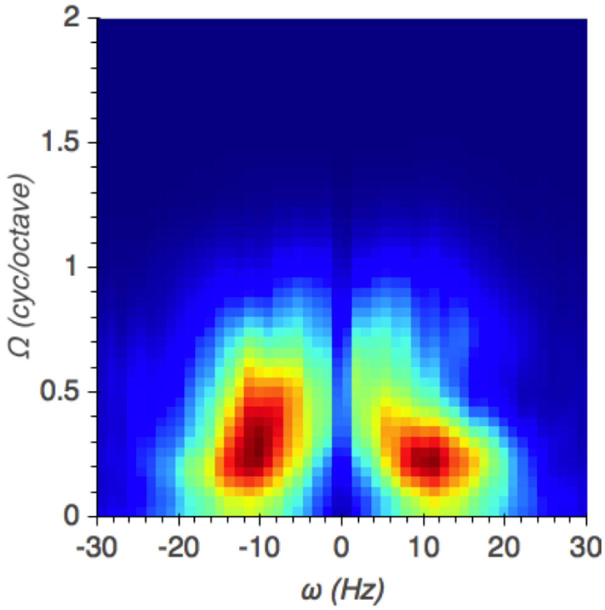

Fig. 3. The aggregate response of all 32 ST filters to the ripple analysis. The results are very similar to human responses found in [20], and the ranges of sensitivity (5-15 Hz in rate; up to 0.7 cycles/octave) are known to be critical to the information content of speech. [21, 22]

first to provide sufficient context for meaningful temporal modulation analysis (600ms), and then to optimize performance in phoneme classification on the held-out validation set, in the end resulting in a phoneme classification error rate of approximately 22%.

3. DATA-DERIVED ST FILTER ANALYSIS

Upon training the network, we performed a set of analyses of the front-end ST filters in order to explore the consistencies with properties of the auditory system. The analyses include inspection of the individual filters, ripple analysis to quantify modulation responses, and measurements of spectral responses to speech input.

3.1. Individual ST Filters

Time-frequency responses of the 32 ST filters trained on the 7 mel-spaced filter bank spectrum of the center of the phonemes are shown in Fig. 1. For display purposes, the filter values were linearly interpolated (on the mel-scale) from their original 7x60 grid. The filters typically show a temporal span of about 200ms and a nonuniform frequency response.

The contribution of each filter to the phoneme classification task was also quantified by computing the mutual information (MI) between the individual filter outputs and the frame labels [19]. The filters are ordered in Fig. 1 in decreasing MI (with MI values printed in the upper left of each response).

3.2. ST Filter Ripple Analysis

Visual inspection of the derived ST filters suggests that the filters are enhancing particular spectral resolutions (scales) and particular modulation frequencies (rates), but a better quantification of this effect is desirable. To evaluate the range of enhanced spectral and temporal modulations, we followed the protocol used in evaluating sensitivity in human listeners in [20], in which systems are presented with a series of probe signals with spectro-temporal ripples of varying rate and scale. For this analysis, we projected spectra with a variety of spectro-temporal ripples on the filters and derived the sensitivities by computing the energy ratio of the output to the input for the particular signal. The response for each filter is shown in Fig. 2, ordered in the same grid as in Fig. 1 so locations are consistent.

Inspection of the responses shows the selectivity of many of the individual filters, with a number of them focusing on particular regions in the modulation space (indicated by localized regions of high response). So, the outcome of this analysis suggests that the filters are specializing in their responses, with many of them focusing on particular spectro-temporal modulations. This sort of behavior was hypothesized as learnable from speech based on the selectivity of cortical processing previously discussed, and we see here that it indeed can be automatically learned.

Considering the response of the filters in aggregate is also of interest. We see in the previous analysis that the individual filters are obeying properties expected based on auditory processing, but overall properties of mammalian ST responses are also known, and so their relation to the filters should be explored as well. With this in mind, the average energy response of the filters to the ripple probe signals is shown in Fig. 3. This response is similar to the responses shown in [20], though there are some differences. As seen, the filters enhance temporal modulation frequencies between 5 and 15 Hz (higher than reported for human listeners [20]) and smooth the spectrum rather heavily by attenuating frequency modulations greater than approximately 0.7 cycle per octave. These modulations are dominant for carrying linguistic information in speech [21, 22] and are consistent with observed human hearing sensitivities to spectro-temporal modulations [23]. So, in aggregate, we see that, like with the individual ST filters, the data-derived system is behaving similarly to our expectations based on auditory cortical processing.

Note that this analysis was repeated with 30 mel filters to ensure the spectral smoothing was not simply a product of the small number of filters. The analysis produced very similar results, with the ST filters derived with the increased frequency resolution actually showing a slight preference to greater spectral smoothing and nearly identical responses to temporal modulation. The analysis was also repeated with non-central frames for phones included in the training, and, again, the overall response of the filters to ripple analysis was consistent. The stability of the responses in these subsequent experiments suggests the human-like spectro-temporal behavior is indeed fundamental to the data-driven learning, and not simply a lucky artifact of a particular study.

3.3. ST Filter Response to Speech

To evaluate the frequency selectivity of the derived ST filters, the training data was projected frame-by-frame on each filter, with a few selected examples shown in Fig. 4. We then measured the ratio of the averaged energy of the speech by filter k to that of the original speech within each spectral band, given by

$$H_{k,i} = \frac{\frac{1}{N}\sum_n |(x * f_k)_i[n]|^2}{\frac{1}{N}\sum_n |x_i[n]|^2}$$

where $*$ is the convolution operator.

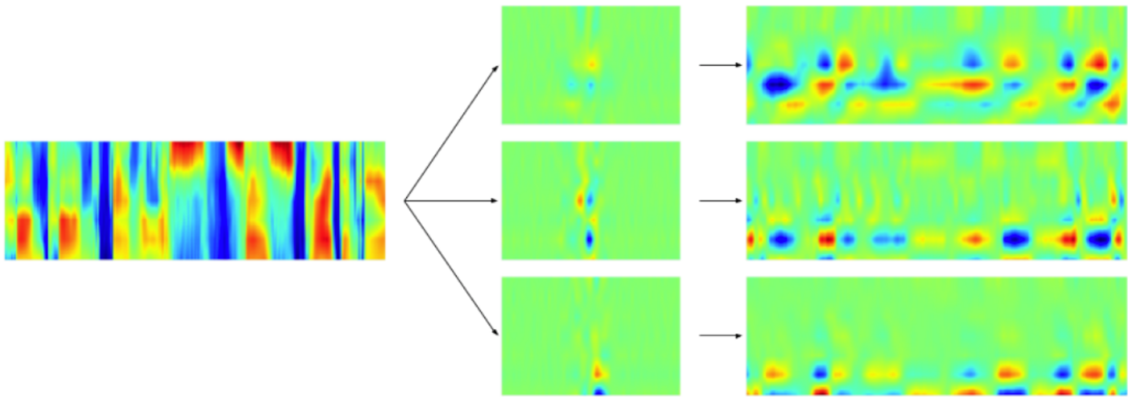

Fig. 4. A selection of filter output responses to speech input. The average ratio of the responses over time to the inputs are shown in Fig. 5

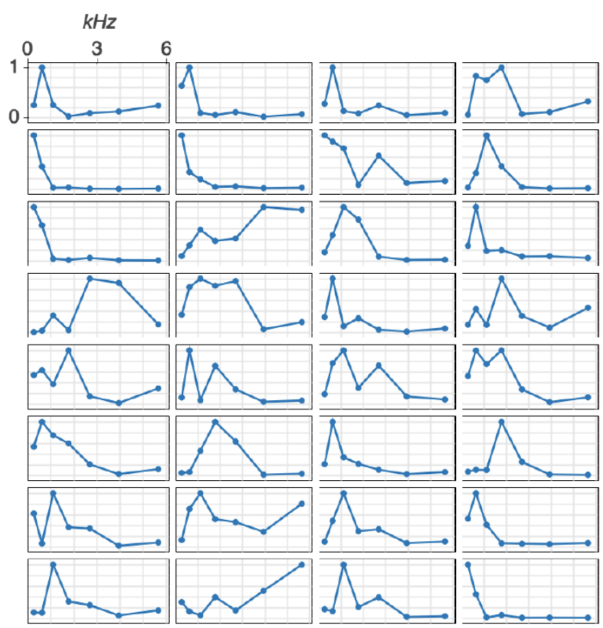

Fig. 5. Spectral energy ratios of the filter output to input speech for the learned filters from Fig. 1. Spikes in the responses reveal the frequency selectivity of the filters.

As seen in Fig. 5, the filters are often highly selective, with strong peaks in the spectral energy ratio. This was already demonstrated to some extent with the ripple analysis in Fig. 2, but this is noteworthily different in that this is showing frequency selectivity in addition to the already demonstrated spectro-temporal modulation selectivity. Furthermore, this confirms that the spectro-temporal selectivities exhibited in Fig. 2 align with particular frequency-limited events within speech.

4. CONCLUSION

The initial exploration presented here yielded data-derived spectro-temporal filters that qualitatively resemble spectro-temporal cortical receptive fields observed in mammalian cortices. Most filters enhance certain carrier frequencies within speech and focus on ranges of modulations which are dominant for carrying linguistic information in speech. Ripple analysis of the resulting ST filterbank indicates that the filterbank enhances modulation frequencies in the 5-15 Hz range and spectral scales up to 0.7 cycle/oct. Such sensitivities are consistent with observed human hearing sensitivities to spectro-temporal modulations.

This work demonstrates that auditory processing and data-driven methods are not necessarily as divergent as they would often appear. In the future, we hope to continue these analyses in the presence of noisier and more challenging training speech in order to study the changes in the front-end filters learned for improved robustness. The analyses presented here also suggest that data-derived networks are, to some extent, automatically learning multistreaming behavior, and so architectures that encourage this sort of processing pipeline will be explored to increase the network's capacity for multiple views of the data.

5. ACKNOWLEDGEMENT

The work was supported by the National Science Foundation under EAGER Grant No. 1743616 and No. 1704170, by a faculty gift from Google Inc and by JHU Human Language Technology Center of Excellence. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or Google Inc.

6. REFERENCES

[1] Taishih Chi, Powen Ru, and Shihab A Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[2] Norman R French and John C Steinberg, "Factors governing the intelligibility of speech sounds," *The journal of the Acoustical society of America*, vol. 19, no. 1, pp. 90–119, 1947.

[3] Michael C Corballis, *From hand to mouth: The origins of language*, Princeton University Press, 2003.

[4] F. Valente and H. Hermansky, "Discriminant linear processing of time-frequency plane," IDIAP-RR-20-2006, also published in ICSLP 2006.

[5] Pavel Golik, Zoltán Tüske, Ralf Schlüter, and Hermann Ney, "Convolutional neural networks for acoustic modeling of raw time signal in lvcsr," in *Interspeech*, 2015.

[6] Hynek Hermansky and Narendranath Malayath, "Spectral basis functions from discriminant analysis," in *ICSLP*, 1998.

[7] Biem and Katagiri, "Filter bank design based on discriminative feature extraction," in *ICASSP*, 1994.

[8] Srinivasan Umesh, Leon Cohen, Nenad Marinovic, and D Nelson, "Frequency-warping in speech," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. IEEE, 1996, vol. 1, pp. 414–417.

[9] Terri Kamm, Hynek Hermansky, and Andreas G Andreou, "Learning the mel-scale and optimal vtn mapping," in *Center for Language and Speech Processing, Workshop (WS 1997). Johns Hopkins University*, 1997.

[10] Kuldip K. Paliwal, Benjamin J. Shannon, James G. Lyons, and Kamil K. Wójcicki, "Speech-signal-based frequency warping," *IEEE Signal Processing Letters*, vol. 16, pp. 319–322, 2009.

[11] Dimitri Palaz, Ronan Collobert, and Mathew Magimai-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," *CoRR*, vol. abs/1304.1018, 2013.

[12] Tara N. Sainath, Brian Kingsbury, Abdel-rahman Mohamed, and Bhuvana Ramabhadran, "Learning filter banks within a deep neural network framework," in *ASRU*. 2013, pp. 297–302, IEEE.

[13] Zoltán Tüske, Pavel Golik, Ralf Schlüter, and Hermann Ney, "Acoustic modeling with deep neural networks using raw time signal for lvcsr," in *INTERSPEECH*, 2014.

[14] Sarel van Vuuren and Hynek Hermansky, "Data-driven design of rasta-like filters," in *EUROSPEECH*. 1997, ISCA.

[15] Jan Pešán, Lukáš Burget, Hynek Heřmanský, and Karel Veselý, "Dnn derived filters for processing of modulation spectrum of speech," in *Proceedings of Interspeech 2015*. 2015, vol. 2015, pp. 1908–1911, International Speech Communication Association.

[16] Fabio Valente and Hynek Hermansky, "Discriminant linear processing of time-frequency plane," Idiap-RR Idiap-RR-20-2006, IDIAP, 0 2006, Published in ICSLP 2006.

[17] F. Kubala, J. Bellegarda, J. Cohen, D. Pallett, D. Paul, M. Phillips, R. Rajasekaran, F. Richardson, M. Riley, R. Rosenfeld, B. Roth, and M. Weintraub, "The hub and spoke paradigm for csr evaluation," in *Proceedings of the Workshop on Human Language Technology*, 1994, pp. 37–42.

[18] Steven B. Davis and Paul Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–66, 1980.

[19] Brian C. Ross, "Mutual information between discrete and continuous data sets," *PLoS ONE*, vol. 9, no. 2, pp. e87357, Feb. 2014.

[20] Taishih Chi, Yujie Gao, Matthew C. Guyton, Powen Ru, and Shihab Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2719–32, November 1999.

[21] Robert V Shannon, Fan-Gang Zeng, Vivek Kamath, John Wygonski, and Michael Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[22] Noboru Kanedera, Takayuki Arai, Hynek Hermansky, and Misha Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Communication*, vol. 28, no. 1, pp. 43–55, 1999.

[23] Taishih Chi, Yujie Gao, Matthew C. Guyton, Powen Ru, and Shihab Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2719–2732, 1999.