A STUDY OF AUDITORY MODELING AND PROCESSING FOR SPEECH SIGNALS

Woojay Jeon and Biing-Hwang Juang

Georgia Institute of Technology
School of Electrical and Computer Engineering
Atlanta, GA 30332, U.S.A.

ABSTRACT

In this paper, we study a modified version of a computational model of the human peripheral and central auditory system [1][2], and examine the validity of its output from two practical perspectives: one that considers the well-known Mel-Frequency Cepstral Coefficients (MFCC) as an approximate representation of the physiology-based early auditory processing result, and the other that allows the derivation of feature vectors from the dimension expanded cortical response of the central auditory system for use in a conventional phoneme recognition task. In addition to confirming the relevancy of the model under existing statistical speech recognition framework, we conduct a preliminary study of the cortical response in connection with known physiological studies, to find new possibilities in using the auditory model to perform cognitive functions based on a better understanding of the human auditory system. In particular, the cortical response may be a place-coded data set where sounds are categorized according to the regions containing their most distinguishing features. The results of this study encourage us to develop hierarchical, detection-based methods in which this mechanism may be utilized to simulate a variety of human perceptual and cognitive functions.

1. INTRODUCTION

Machine listening systems often employ rudimentary simulations of the human auditory system to mimic human perception and cognition of sound. For example, in the case of speech recognition, the Linear Predictive Coding (LPC) model spectrum is built on an all-pole model of the resonances of the vocal tract, while the MFCC is based on an approximation of the critical bands. Most of these front-end processing methods, however, are based on only crude approximations of the peripheral auditory system, with little or no consideration for the latter stages along the auditory cortex where signal representations may undergo further transformations. Building a machine that approaches the capability of humans is still far beyond our reach. It was shown in [3] that automatic speech recognition systems perform far worse than human listeners under noisy conditions. Hence, while much research is aimed at developing functional approximations to human capabilities, there is an intense interest in building computational models that accurately and extensively mimic human physiology. Studying such physiological models may also lead to a better understanding of human audition, thereby offering the possibility of inducing improved functional models.

Relatively recent developments by Shamma [1] include simulations of not only the peripheral auditory organs but also the neural encoding of the primary auditory cortex(A1) in the central auditory system. The one-dimensional auditory spectrum produced by the early stages of the model are transformed into a three-dimensional, data-redundant response in the A1, which may encode auditory features that are relevant to perception and cognition in a more explicit, place-coded manner. In this study, we develop a modified version of this model and examine its validity as a representation of auditory signals. Our study was carried out along two directions. First, we study the computational relationship between the MFCC and the auditory spectrum. The widely-used MFCC was proposed more than two decades ago as a representation based on a crude approximation of the auditory response, with little rigorous justification for its implementation. In contrast, the auditory spectrum is produced by a more elaborate, physiologically motivated (and perhaps better justified) early auditory model. Such a comparison would provide insights for developing more refined parametrization methods for speech and audio signals. Second, we (reverse) validate the new dimension-expanded cortical response model by deriving speech parameters from the model and applying them to a phoneme recognition task. We find the cortical response is capable of providing speech features that are comparable to the MFCC in terms of recognition accuracy.

We shall re-emphasize that the aim of our study is to strive for more thorough modeling of cognition and perception than to simply derive parameters for automatic speech recognition. The purpose of experimenting with a conventional HMM-based system is only to validate this new auditory model within *existing* recognition framework. By studying the dimension-expanded cortical response in connection with known physiological studies, we could gain an enhanced understanding of auditory physiology, and develop better functional and computational models for not only achieving more robust recognition but solving other interesting auditory analysis problems. A preliminary study of the variance of the cortical response hints at a place-coding mechanism that may offer us many new possibilities of utilizing the dimension-expanded data in a general, hierarchical framework for the detection of perceptual and cognitive cues.

2. THE AUDITORY SPECTRUM AND THE MFCC

The auditory spectrum [2] is a one-dimensional spectral representation produced by a computational model of the mechanical and

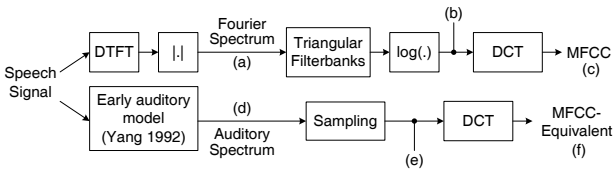
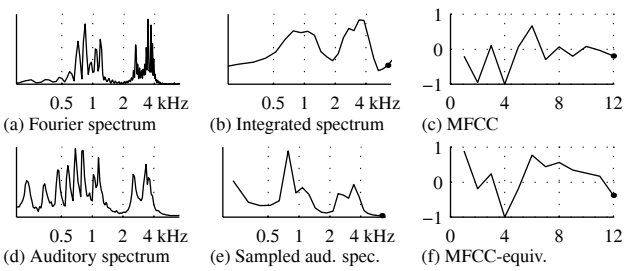


Fig. 1. Comparing the auditory spectrum with the mel-cepstrum.



(a) Fourier spectrum (b) Integrated spectrum (c) MFCC
(d) Auditory spectrum (e) Sampled aud. spec. (f) MFCC-equiv.

Fig. 2. Deriving an MFCC-equivalent from the auditory spectrum

Table 1. Approximate center frequencies (Hz) and bandwidths (Hz) used for the filterbanks of the MFCC (implemented by HTK software) and the MFCC-equivalent. Cochlear filterbanks designed at the Institute for Systems Research at the University of Maryland were used for the early auditory processing.

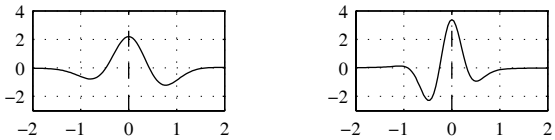
	MFCC							MFCC-e					
f_c	68	144	226	317	416	525	645	226	320	415	523	640	784
b/w	144	158	173	190	209	229	251	28	40	52	65	80	98
f_c	777	921	1080	1254	1445	1655	1886	932	1077	1245	1438	1661	1865
b/w	276	303	333	365	401	440	483	117	136	157	183	212	238
f_c	2139	2416	2721	3056	3423	3827	4270	2154	2418	2714	3047	3420	3839
b/w	531	583	640	702	771	846	929	278	313	354	400	453	513
f_c	4756	5289	5875	6519	7225			4309	4699	5274	5920	6456	
b/w	1020	1120	1229	1349	1481			584	645	738	848	944	

neural processing in the early stages of the auditory system. Since the widely-used MFCC is mostly based on a crude approximation of the peripheral auditory system, most notably the cochlear filtering action that affects human perception of pitch, the more accurate early processing model essentially encompasses the computation stages of the MFCC. Specifically, the integration of spectral energy via mel-scale filterbanks has much to do with the cochlear filtering followed by the nonlinear stages used to obtain the auditory spectrum, as illustrated in Figure 1. Hence, a crude counterpart to the MFCC could be extracted from the auditory spectrum by selecting the output of those channels corresponding to the MFCC's critical-band filters and applying the Discrete Cosine Transform on these sampled points, as shown in Figure 2. Note that only a crude match of the center frequencies was done to obtain the auditory spectrum channels, and the bandwidths differ significantly, as shown in Table 1. To more accurately demonstrate the relationship between the auditory model and the MFCC, some of the spectral integration at the cortical stages would have to be considered. Also note that cochlear channels corresponding to three of the MFCC's filterbanks are unavailable. Despite these simplifications, this feature can be useful in quantitatively studying the relationship between the MFCC and the physiological model by applying it in the phoneme recognition task described in Section 4. Further insights are provided in Section 5.

3. THE CORTICAL RESPONSE

3.1. The A1 Model

In the A1[1], the one-dimensional auditory spectrum is redundantly encoded by a set of neurons, each neuron possessing a "response area" that characterizes the amount of excitation induced by spectral components along the tonotopic frequency axis. Each response area is characterized by an excitatory range around the neuron's best frequency (BF) surrounded by inhibitory areas. The response areas are organized in roughly three dimensions: BF, bandwidth (scale), and symmetry (phase). The bandwidth dimension indicates the overall stretch of the neuron's excitatory and



(a) $s = 0.50$ cyc/oct, $\phi = -\pi/8$ (b) $s = 0.80$ cyc/oct, $\phi = \pi/4$

Fig. 3. Analysis functions with fixed BF (dashed lines) and varying scale and phase. Excitatory peaks are aligned to the BF's.

inhibitory areas, while the symmetry indicates the difference in inhibition above and below the BF.

In the original model[1], a seed analysis function $h_s(y)$ is defined on the tonotopic axis y to simulate these response areas. A sinusoidal interpolation between $h_s(y)$ and its Hilbert Transform $\widehat{h}_s(y)$ models the varying symmetry, parameterized by ϕ. Although this construction allows efficient computation of the cortical response, it also has the effect of causing the peak of each excitatory lobe to drift away from the BF as $|\phi|$ increases. Hence, in our model we added a translation factor $c(s,\phi)$ to every analysis function to compensate for this deviation and align the excitatory peaks of all analysis functions to their BF's. The resultant cortical response becomes a more direct encoding of the local symmetry around each point on the auditory spectrum, allowing easier interpretation of visual plots. Mathematically, the analysis function on the tonotopic domain y parameterized by x (best frequency), s (scale), and ϕ (symmetry) can be written as:

$$\begin{aligned} w(y;x,s,\phi) &= h_s\left(y-x+c(s,\phi)\right)\cos\phi \\ &+ \widehat{h}_s\left(y-x+c(s,\phi)\right)\sin\phi \end{aligned} \quad (1)$$

The zero-scale correction factors $c(0,\phi)$ can be found by numerically solving the following equation:

$$dw'(y;0,0,\phi)/dy = 0 \quad (2)$$

for each ϕ where $w'(y;x,s,\phi)$ is the same as $w(y;x,s,\phi)$ but without the alignment factor $c(s,\phi)$. We can then compute $c(s,\phi)$ by dilating $c(0,\phi)$ as follows:

$$c(s,\phi) = c(0,\phi)/\alpha^s \quad (3)$$

where α is the dilation factor in [1]. Two examples of the resultant analysis functions are shown in Figure 3. Only the range $-\frac{\pi}{2} < \phi < \frac{\pi}{2}$ is used for the symmetry axis in our model. The cortical response is obtained by computing the inner product of these functions with the auditory spectrum $p(y)$.

$$r_3(x,s,\phi) = \int_{\mathcal{R}} p(y)w(y;x,s,\phi)dy \quad (4)$$

An example of the auditory spectrum and its cortical response is shown in Figure 4 for the vowel /aa/. Note that, compared to [1], it is easier to visually track the change in symmetry as the local spectral components at each tonotopic point are encoded from fine scale to broad scale (e.g., along the line drawn at around 3 kHz) due to our alignment of the best frequencies.

3.2. Feature Extraction

To experimentally corroborate the validity of the cortical response as an auditory representation, we use two well-known dimension reduction methods, Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA) [4], to derive feature vectors for use in a phoneme recognition task.

Since the cortical response contains too much data for direct application of PCA and LDA, we first apply a simple method of data reduction by discarding regions that are found to have relatively high variance for all phonemes. The reasoning here is that the responses in these locations will be weakly correlated with the

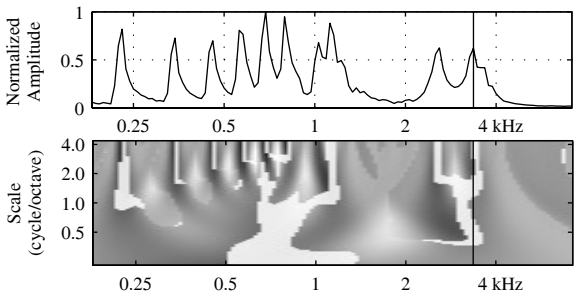


Fig. 4. The auditory spectrum and cortical response of the /aa/ vowel. Only the maximum cortical response along each ϕ-axis is plotted using the color convention in [1]. Readers are asked to refer to the website `http://www.ece.gatech.edu/~wjjeon` to view the plots in color.

identity of each phoneme. For example, in Figure 5 we can see that the response for many different samples of the vowels /iy/ and /uw/ has high variance at the upper left regions. As discussed in [1], the responses in these regions are usually manifestations of pitch, which, naturally, is highly variant and does not contribute to discriminating between the actual identity of vowels. In the same manner, we assume that cortical regions that have overall high variance across all phoneme classes do not contribute much in distinguishing between the phonemes, and discard such regions from the cortical response. As a result, we can vastly reduce the data to a more manageable size where PCA and LDA can be more readily applied. In the case of LDA, we circumvented the rank-deficiency of scatter matrices by reducing the data to an intermediate size by PCA before applying LDA[5].

4. EXPERIMENTAL VALIDATION

Recognition experiments were run using phonemes extracted from the TIMIT database. From [6], we arbitrarily chose (for the sake of simplicity) only one phoneme from each of the seven groups where within-group confusions were not counted, and excluded the group of closures, resulting in a total of 38 phonemes. 16 kHz samples of phoneme utterances were taken from all "si" and "sx" sentences in the TIMIT database, resulting in a total of 82,881 total training phone samples and 30,335 testing samples. Clean data was used for training, while Gaussian white noise was added to the test data to vary the SNR. The auditory model response consisted of 128 channels on the tonotopic axis, 21 channels on the scale axis (with a resolution of 4.7 channels per octave), and 11 channels on the phase axis linearly spaced between $-\pi/2$ and $\pi/2$. Each raw feature consisted of 12 points, and frame-by-frame energies and delta and acceleration coefficients were appended to form 39-point vectors for the recognition task. The recognizer was a 5-state HMM (3 emitting) implemented with HTK software. Table 2 shows the recognition ratios achieved with increasing Gaussian mixtures per state using the MFCC, the MFCC-equivalent (MFCC-e), $\mathbf{r_p}$ derived from the A1 response by PCA, and $\mathbf{r_d}$ by LDA. The robustness of the physiological model toward noise[7] seems to contribute significantly to its performance under low SNR, especially when using $\mathbf{r_p}$. Note that the MFCC-e is only a very crude derivative of the auditory spectrum based on the computation of the MFCC, with no added attempt to optimize it for the recognizer. Moreover, the pitch-related harmonics of the signal are greatly accentuated in the auditory spectrum[2], which would further harm the MFCC-e's robustness. The A1-derived features are also somewhat arbitrary features used simply to quantitatively validate the physiological model under conventional HMM-based recognition framework. Extracting features to compete with the MFCC for speech recognition is not the objective of this paper.

Table 2. Phoneme recognition rates (%) for varying features, SNR(dB, C=Clean), and number of mixtures per state.

	MFCC				MFCC-e				$\mathbf{r_p}$				$\mathbf{r_d}$			
	1	4	8	32	1	4	8	32	1	4	8	32	1	4	8	32
C	47.2	55.3	58.3	62.7	36.3	41.9	45.6	49.7	39.5	47.2	50.9	56.5	42.1	49.6	52.9	58.8
20	41.3	49.3	50.9	54.3	35.6	41.3	44.7	48.9	39.3	46.6	50.4	55.7	40.6	48.0	50.8	55.1
15	36.4	43.9	45.6	48.1	35.0	40.3	43.5	47.7	38.4	45.8	49.4	54.2	34.8	41.2	43.5	47.1
10	29.8	35.5	36.8	39.0	33.1	38.0	40.5	44.2	36.7	43.1	46.5	50.7	22.5	26.6	28.5	32.4
5	22.6	24.8	25.7	26.7	28.5	31.8	33.5	35.2	32.5	36.5	39.0	41.5	15.6	16.4	17.3	18.2

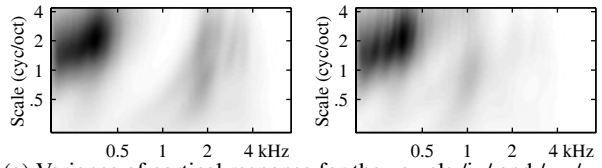
5. TOWARD A HIERARCHICAL AND CATEGORY-BASED DETECTION FRAMEWORK

While we have focused on implementing and validating a model of the peripheral and central auditory system in this study, a more in-depth investigation of the dimension-expanded auditory representation that it provides can lead us to new directions in using it for perceptual and cognitive tasks.

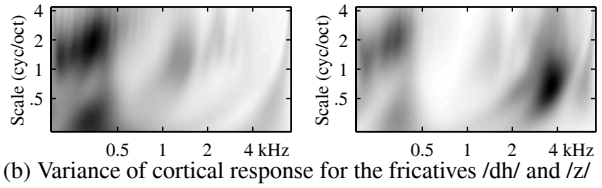
An interesting starting point is to study the variance of the cortical response to various phonemes. Figure 5 shows the variance of the zero-phase response for several different phonemes. If we consider the statistics of individual neurons only (without regard to correlations among neighboring neurons), it is conceivable that the identifying features of each phoneme lie in the light-shaded low-variance regions, as was discussed in Section 3.2. Moreover, one notices that many phonemes can be grouped together according to the similarity of their low variance regions. In Figure 5, for example, such a grouping can be found with several phonemes categorized as vowels, fricatives, affricates, and plosives.

One possible implication of this phenomenon is that the cortical response serves as a place-coded data set where phonemes sharing common characteristics have a common "identifying region" where their differentiating features are most strongly present. Hence, in order to detect the presence of a certain phoneme, one must analyze the identifying region pertaining to its category. It is already implied by many physiologists that the auditory system is composed of specialized processing stations. For example, evidence is shown in [8] that distinct regions of the brain process syllables while others process phonemes. [9] states that the left hemisphere of the brain may be specialized in processing acoustic transients. If we were to extend this notion of specialization to the processing of speech phonemes, we may hypothesize that the cortical response consists of distinct identifying regions that are specific to certain categories of phonemes, from which we can extract data to make cognitive decisions. Note that we need not limit ourselves to speech signals, but other complex audio signals in general. For example, different identifying regions may exist for different categories of musical instruments.

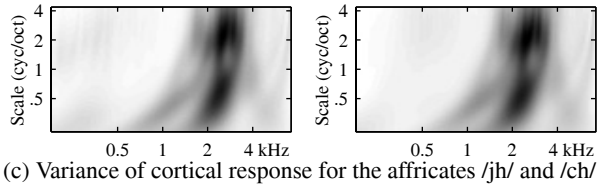
Furthermore, many physiological studies imply that the functions of the brain, including the auditory system, are organized hierarchically. For example, [10] shows through functional magnetic resonance imaging (fMRI) that pure tones primarily activate the core of the human auditory cortex, while complex sounds such as narrow-band noise usually stimulate the belt areas, implying a hierarchical process of sound being decomposed into basic features and later integrated into more complex stimuli. This inspires us to hypothesize that the cortical response may be an intermediate auditory representation from which higher-level processing



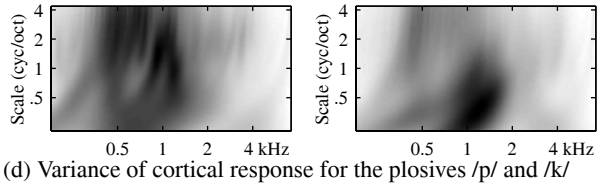
(a) Variance of cortical response for the vowels /iy/ and /uw/



(b) Variance of cortical response for the fricatives /dh/ and /z/



(c) Variance of cortical response for the affricates /jh/ and /ch/



(d) Variance of cortical response for the plosives /p/ and /k/

Fig. 5. Variance (dark=high) of cortical response at $\phi = 0$.

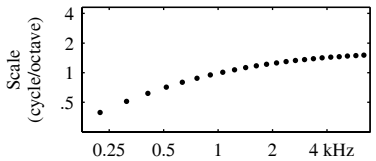


Fig. 6. Conceptual representation of MFCC filterbanks.

stations extract data to form cognitive decisions in multiple stages, starting with a broad categorization of sound (e.g. vowels vs. consonants or string vs. brass instruments) followed by more specific cognitive decisions (e.g. vowel /aa/ vs. /iy/, trumpet vs. tuba).

This inspires us to develop new, hierarchical systems where the presence of categories or classes of sound are measured and detected at distinct levels of processing. We can also take full advantage of the place-coded cortical response by using category-based identifying regions to make more accurate cognitive decisions. Note that in our recognition task, we applied a single low-variance region for the recognition of all phonemes, which completely ignored this category-based place-coding. Furthermore, by reducing the data set into a 12-point vector, we have taken away the dimension-expanded data redundancy that may be important for robustness. In the case of the MFCC, the mel-scale filterbanks can be conceptually represented on the zero-phase cortical plane by finding neural response areas with corresponding BF's and scales, as shown in Figure 6. Although this is a very simplified one-to-one correspondence, it conceptually demonstrates that the MFCC, in some respect, represents only one small subset of the auditory response. This implies the existence of other, more refined speech features that are based on more relevant areas of the cortical response. We also note that future studies should allow the model to incorporate explicit processing of temporal information, which plays an important role in the perception of pitch or timbre.

6. CONCLUSION

In this study, we implemented and studied a computational model of the peripheral and central auditory system based on the work by [1] and [2]. As a means of indirect validation of the model, we derived a crude equivalent to the MFCC from the more accurate peripheral model and used this feature in a speech recognition task to make a quantitative comparison. We also derived feature vectors from the dimension-expanded A1 model using well-known pattern recognition techniques and used them in a speech recognition task to verify the model's validity under conventional recognition methodology. By studying physiological models, we have gained some insights into human perceptual and cognitive processes and new approaches to simulating them. In this study, features were derived from the auditory model only to carry out experimental validation. Category-distinct low variance regions of the cortical model, in connection to existing physiological studies, suggest that a hierarchical, category-based system of detecting perceptual and cognitive cues may provide a more robust, elegant framework for solving various auditory analysis problems in future studies.

7. ACKNOWLEDGEMENTS

Thanks to Dr. Kuansan Wang, Dr. Shihab Shamma, and Rungsun Munkong for their help in implementing the auditory model.

8. REFERENCES

[1] K. Wang and S. A. Shamma, "Spectral shape analysis in the central auditory system," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 5, pp. 382 – 395, Sept. 1995.

[2] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 824 –839, Mar. 1992.

[3] R. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, Mar. 1997.

[4] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley and Sons, Inc., 2001.

[5] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, no. 8, pp. 831–836, Aug. 1996.

[6] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 11, pp. 1641–1648, Mar. 1989.

[7] K. Wang and S. Shamma, "Self-normalization and noise-robustness in early auditory representations," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 3, pp. 421 – 435, July 1994.

[8] W. T. Siok, Z. Jin, P. Fletcher, and L. H. Tan, "Distinct brain regions associated with syllable and phoneme," *Human Brain Mapping*, vol. 18, pp. 201–207, 2003.

[9] I. S. Johnsrude, R. J. Zatorre, B. A. Milner, and A. C. Evans, "Left-hemisphere specialization for the processing of acoustic transients," *NeuroReport*, vol. 8, pp. 1761–1765, 1997.

[10] C. M. Wessinger, J. VanMeter, B. Tian, J. V. Lare, J. Pekar, and J. P. Rauschecker, "Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging," *Journal of Cognitive Neuroscience*, vol. 13, no. 1, pp. 1–7, 2001.