SPEECH PROCESSING WITH A CORTICAL REPRESENTATION OF AUDIO

Nima Mesgarani¹,², Shihab Shamma²

¹Department of Neurological Surgery, University of California San Francisco
²Institute for System Research, University of Maryland College Park

(mnima,sas)@umd.edu

ABSTRACT

Neurophysiological studies in the primary auditory cortex have recently demonstrated a rich diversity of responses that provide an explicit multidimensional representation of phonemic acoustic features (Mesgarani 2008). Specifically, distinct subsets of cortical neurons are activated by articulatory gestures and dynamics that are characteristic of different phonemes. Here we use a computational cortical model to illustrate how these phonetic features appear in such a multiresolution representation. We also review how this representation has been successfully applied in variety of speech processing tasks including robust speech discrimination, speech enhancement and phoneme recognition.

1. INTRODUCTION

Humans reliably identify many phonemes and discriminate them categorically despite considerable natural variability across speakers and distortions in noisy and reverberant environments that limit the performance of even the best speech recognition algorithms [3]. There is experimental evidence for cortical encoding of acoustic phonetic features regarded as critical for distinguishing phonemes [6]. Recent neurophysiological studies have shown that neuronal responses to continuous speech in the primary auditory cortex provide a multidimensional representation that is sufficiently rich to support the discrimination of many phonemes [1], and is made possible by the wide range of spectrotemporal tuning in auditory cortex to stimulus frequency, bandwidth (or scale) and dynamics (or rate) [1, 7]. In previous studies, we demonstrated the applications of this model in a variety of speech processing and recognition tasks [2, 4, 5]. The emphasis of this paper is on describing the encoding of acoustic phonetic features in a computational model of cortical representation of sound. In particular, we shall highlight how acoustic phonetic features appear explicitly in this multi-resolution framework, and the advantages these views afford us compared to the usual spectrogram.

2. CORTICAL REPRESENTATION OF SPEECH

The computational auditory model is based on neurophysiological investigations at various stages of the auditory system [8]. It consists of two basic stages. An early stage models the transformation of the acoustic signals into an auditory spectrograms modeling the cochlear filter bank, hair cell and lateral inhibitory network. The central stage analyzes the spectrogram to estimate the content of its spectral (scale) and temporal (rate) modulations using a bank of modulation selective filters mimicking those described in a model of mammalian primary auditory cortex. This is mathematically equivalent to a two-dimensional wavelet transform of the auditory spectrograms, with wavelets resembling a 2D Gabor function. We assumed a bank of directional selective filters tuned to different rate and scales, and can have symmetric or asymmetric shapes (Figure 1). The output of this stage is the convolution of the auditory spectrogram \((y(t,f))\) with STRFs (Figure 1):

\[
r_{\pm}(t,f,\omega,\Omega,\Phi) = y(t,f) * STRF_\pm(t,f,\omega,\Omega,\Phi)
\]

where \(\omega\) is rate (temporal modulation in Hz), \(\Omega\) is scale (spectral modulation in Cycle/Octave) and \(\Phi\) is the phase of the filter and \(\pm\) specifies the direction of tuning (up vs. down sweep). This multiresolution output provides an explicit representation of different spectrotemporal features of the input spectrogram based on how they match various tuning properties of the filters as explained in next section.

2.1. Analyzing speech along various dimensions

The cortical representation of speech is a multidimensional tensor that makes its visualization rather challenging. Here, we describe the encoding of information along various dimensions by reducing the output in ways that highlight them separately. We start with pure temporal or pure spectral decomposition of the spectrogram and extend this framework to the full spectrotemporal analysis. Figure 2 shows spectrogram of a sentence along with its filtered temporal modulations at 2,4 and 10Hz. The modulation filters were tuned only to different rates with no scale or
direction sensitivity. Speech components that vary more slowly show up in lower rates (2 Hz), while the fast changing spectral features such as onsets of plosives activate faster filters (10 Hz). To obtain rate vs. time representation (rategram), we reduced each filtered spectrogram at each time instance to its maximum value, ignoring the frequency where the maximum occurred. The result of this operation is a two-dimensional display showing energy at different rates (figure 3b) for different times. The same operation can be done using filters with spectral only tunings to estimate scale vs. time (scalegram, figure 3c). To compute the up-down directions, the full spectrotemporal filtering of the spectrogram is required. This results in two subgroups of filters each capturing the sound energy that has upward or downward pattern such as prosody or formant transitions.

3. PHONETIC FEATURES MANIFESTED IN THE SCALEGRAMS AND RATEGRAMS

The scalegrams and rategrams highlight information about phonemic spectral shape and dynamics that are often only implicitly represented in the spectrogram. Such features can serve as robust cues for their encoding and detection in ASR and other speech and audio applications as we elaborate later.

3.1. Scalegrams and rategrams of continuous speech

Scalegrams suppress the frequency and rate dimensions to highlight the best scales of analysis as a function of time. The ordinate roughly measures the bandwidths of the spectrum at any moment. Vowels and consonants in continuous speech occur in a wide variety of contexts that influence the details of their representation. However, there are prominent features that transcend this variability and reflect fundamental properties of speech such as its syllabic, rhythmic, and prosodic structure. For example, consider the scalegram of Figure 3c where we see strong activity in the highest (finest resolution) scales (>1 cyc/oct) whenever the speech is voiced reflecting the presence of resolved harmonics. By contrast, strong activity in the lowest scale (<0.25 cyc/octave) indicates epochs with compact spectra such as the vowels /ae, ow/ (discussed later).

3.2. Spectral information in the averaged scalegrams

Scalegrams indicate the instantaneous spectral bandwidth of speech, or more accurately its compact/diffuse nature. We illustrate this point with the series of mean vowel scalegrams in Figure 4a. Back vowels (e.g., /ao, ah, a/)) generally have closely-spaced formants which result in a single compact spectral envelope when viewed at moderate resolution. By contrast, front vowels (e.g., /ey, ih, iy/) have
widely separated formants and hence exhibit multi-peaked diffused spectral envelopes (Fig. 4a). Other vowels (e.g., /ae, eh, ax, ux, and ow/) are transitional in character with diffuse but mostly single differentiated peak in their envelopes. These and other spectral features are readily reflected in the averaged scalegram patterns of Figure 4a. First, voicing in all vowels is indicated by the strong activation of the highest scales due to the presence of closely spaced harmonics. Compact spectral envelopes activate best the lowest (broadest) scales, while the multiple peaks of the diffused spectra activate the intermediate scales. Transitional vowels have a more varied activation that combines both of the patterns above.

3.3 Dynamic information in the rategrams

Most phonemes also possess distinctive dynamics due to their place and manner of articulation that are sometimes difficult to discern in spectrograms (Figure 4). Rategrams make these dynamics explicit as illustrated in the averaged patterns in Figure 4. In these plots, the ordinate indicates approximately the speed with which the spectral features change.

Vowels are usually described as steady spectra (over short times) and little is noted about their dynamics in the traditional spectrograms. Rategrams however reveal a different picture (Fig. 4a). At their onsets, vowels usually activate the fastest rate (or speed) channels because of the rapid rise in energy relative to preceding epochs. Slightly later, more prolonged activation occurs that extends down to intermediate rates (for short vowels such as /ix, ax, ux, ih/), to lower rates (for longer vowels such as /ao, ah, owl/), and finally to the lowest rates (for the longest vowels /aa, ae, eh, ey, iy/). Such a distribution also distinguishes the longer labile consonants (/p, b, m/) from others (/t, d, k, g, n, ng/) that activate only the highest rates (Fig. 4b).

4. SPEECH PROCESSING USING CORTICAL REPRESENTATION

Speech has a distinct pattern of spectrotemporal modulations that differ from those of most sounds and environmental noise and distortions. As a result, the representation of speech in noise is often more separable in the cortical representation than at the level of spectrograms. Therefore, learning the signature of speech in this representation can be used to discriminate speech from many other sounds including music, animal vocalizations and environmental noise even when the quality of the recordings are severely degraded [2]. Figure 5a illustrates the classification result for speech in noise and reverberation compared to two commonly used systems. Red curve represents the cortical-based model that remains robust down to -15dB SNR, where the other two systems break down at around 20dB. Since the cortical transformation is
invertible, one can use the spectrotemporal separation of speech and noise for speech enhancement [4]. Figure 5b illustrates spectrogram of speech in jet noise (6dB SNR, left) and its cleaned spectrogram on the right, after applying a filter to the spectrotemporal representation of noisy speech (Fig 5b, middle). Objective and subjective tests demonstrated the improved quality of the cleaned speech compared to the original noisy samples [4]. Finally, the cortical representation fits well in the multistream speech recognition framework [5]. In this approach, different spectrotemporal modulations of the signal are processed and classified in separate processing channels in order to adaptively eliminate the corrupted channels while preserving the uncorrupted ones for further processing. Figure 5c displays the recognition accuracy for different combinations of spectrotemporal modulations for four phoneme classes of vowels, nasals, fricatives, and plosives. As shown in Figure 5c, various spectrotemporal modulations contribute differently to the recognition of different phoneme classes. For example, mid-rate, high-scale streams perform better for vowels, while high-rates, high-scale streams have higher accuracy for nasals.

Fricatives are better classified with low-scale, mid-rate streams. Finally, the fast streams achieve a higher accuracy in plosive classification. A system that combines these results from all streams demonstrated improved recognition over a baseline system [5].

5. DISCUSSION

A multidimensional representation of speech was made possible by the wide range of spectrotemporal selective filters applied to the speech auditory spectrogram. One advantage of such multidimensional representation is that there is always a unique sub-population of features that captures the distinctive acoustic features of a given phoneme and hence encodes that phoneme in a high-dimensional space independent of speaker and context. In addition, speech and many noises are well separated in the cortical modulation representation which can be quite beneficial for robust speech processing algorithms. Future work includes extending the model to account for additional neural mechanisms such as adaptation, top-down modulations, and rapid plasticity.

6. ACKNOWLEDGMENT

This result was partly funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL) and NIH (R01 DC005779). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U.S. Government.

7. REFERENCES