Auditory Processing of Speech: The COG Effect

Student Researcher: Daniel E. Hack

Advisor: Dr. Ashok Krishnamurthy

The Ohio State University
Department of Electrical and Computer Engineering

Abstract

The "COG effect" is an auditory phenomenon in which a human listener can perceive the spectral center of gravity (COG), or centroid, of a frequency band up to 3.5 Bark wide. Engineering applications of this effect include extracting centroid features for use in speech recognition algorithms. This work investigates the representation of the spectral centroid in the auditory system using a computational model of the auditory pathway. Specifically, a model of the representation of sound in the primary auditory cortex (the termination of the auditory pathway in the temporal lobe of the cortex) is used to derive a spectral centroid measure. This measure is then used to predict the results of two COG effect listening experiments, the first a vowel matching task and the second a pitch matching task. By demonstrating that the properties of a subset of the cortical representation of sound match those of the COG percept, this work concludes that cortical processing and the resulting cortical representation may represent the mechanism underlying the COG effect.

Introduction and Objectives

The human auditory system is superior to machines in nearly all speech related listening tasks such as speech recognition and speaker identification. In light of this fact, researchers have looked to the auditory system for insight in how to improve their algorithms. This trend has led to the incorporation of principles of auditory processing into state-of-the-art speech processing algorithms. For example, mel-frequency cepstral coefficient (MFCC) feature vectors are routinely used to encode the spectral features of speech signals as the first stage in speech recognition and speaker identification algorithms. The MFCC is based on two principles of auditory processing: first, the frequency analysis of the cochlea (organ of the human inner ear) may be simulated as a bank of bandpass "auditory filters," the output of which is typically called the "auditory spectrum;" and second, the auditory spectrum is subject to a second layer of processing in which properties of the spectral profile are extracted. Accordingly, the MFCC is calculated by passing the power spectrum of an input signal through an auditory filterbank (generate auditory spectrum), then applying logarithmic compression and a cosine transform (extract features of the auditory spectral profile). Thus, by imitating the processing of the auditory system, the MFCC has enabled increased performance in speech recognition and speaker identification algorithms.

This motivates the study of auditory processing itself, specifically its speech processing properties, as a means to inspire novel speech processing algorithms. One such avenue of research investigates spectral integration: how the auditory system combines information across frequency. A speech waveform is a broadband signal, containing frequency content from roughly 100 to 6000 Hz. The spectral profile of a speech signal typically contains several prominent peaks, called formants, which correspond to the resonant frequencies of the vocal tract at the time of production. The formant frequencies are labeled F1, F2, F3, etc., in order of increasing frequency. The idea of spectral integration in speech perception research has been formalized into the "center of gravity (COG) effect," which states that two closely spaced vowel formants are effectively "merged" into a single spectral prominence whose COG (mean frequency) determines the phonetic quality of the vowel. "Phonetic quality" refers to the properties of the vowel which influence the listener's determination of vowel identity, i.e., the phonemic decision. In other words, in vowels with closely spaced formants, the COG of the two formants is a salient cue which plays a significant role in vowel, and thus speech, recognition.

Independently, the engineering literature has recently incorporated a similar idea into speech recognition algorithms. A line of research led by Kuldip Paliwal (Paliwal 1998, Chen et. al. 2004) is investigating the

use of "spectral subband centroids" (SSCs) as features in speech recognition algorithms. SSCs are calculated by filtering the speech signal through a fixed number of spectral subbands, and calculating the centroids of the resulting power spectra. Results show that SSCs can achieve comparable recognition performance to MFCCs with clean speech, and better performance than MFCCs with noisy speech, validating their efficacy as speech recognition features. However, the evolution of SSC calculation has progressed heuristically, through engineering intuition, rather than incorporating knowledge of the auditory processing of speech. It is reasonable that better performance could be achieved by explicitly incorporating the parameters governing the COG effect into the calculation of SSCs.

However, there is dispute in the speech perception literature regarding the details of the COG effect: what is the bandwidth of spectral integration; where is the COG "encoded" in the auditory nervous system (peripherally or centrally); is the COG effect a fundamental property of the auditory system, or exclusive to speech perception? This work considers such questions in order to lay the ground work for future applications of the COG effect in speech processing.

Methodology

The monaural auditory pathway conveys information as a time-varying pattern of neural excitation. It consists of a complex network of serial and parallel neural connections, converging in several major neural nuclei, and terminating in the primary auditory cortex. It is generally divided into two sections: peripheral processing (ear and auditory nerve); and central processing (brainstem, midbrain, and cortex). Recent work investigating the physiology of the ferret auditory cortex has led Shamma and colleagues to propose a computational model of auditory processing, composed of peripheral (Yang et. al. 1992) and central stages (Wang et. al. 1995).

The peripheral stage simulates the mechanical to neural transduction of auditory information in the cochlea and a lateral inhibitory spectral sharpening process in the cochlear nucleus. The output of the peripheral stage is the "auditory spectrum," a spatial pattern of neural excitation at the output of the cochlear nucleus, which resembles a sharpened version of the acoustic input spectrum. The frequency axis of the auditory spectrum is called the "tonotopic" axis because in the auditory system frequency is mapped to location. Figure 1 shows the spectrum of an input signal and its corresponding auditory spectrum.

The central stage performs a spectral shape analysis, implemented as a multi-scale filtering of the auditory spectrum, resulting in the final cortical representation. This is a simplification of the central stages of the auditory pathway, but is functionally valid in that it has been shown to accurately predict cortical neural responses to vowel stimuli (Versnel et. al. 1998). The central stage models the auditory cortex as a population of neurons with receptive fields (RFs) parameterized along three dimensions: best frequency x_c (octaves re. 1 khz), the RF's center frequency along the tonotopic frequency axis; scale Ω_c (cycles/oct), which controls the RF's bandwidth; and symmetry ϕ_c (rad), which controls the RF's symmetry w.r.t. the best frequency. A receptive field is defined as:

$$RF(x, x_c, \Omega_c, \phi_c) = h(x - x_c)\cos\phi_c - \hat{h}(x - x_c)\sin\phi_c,$$

where *h(x)* is an even seed function, defined as the second derivative of a Gaussian, and $\hat{h}(x)$ its odd Hilbert transform. These are shown in figure 2. Sample RFs are shown in figure 3 for different scales and symmetries. Intuitively, the RF describes a neuron's excitatory and inhibitory regions. The response of a neuron with receptive field RF is given as the inner product of RF with the input auditory spectrum y(x):

$$\begin{aligned} r(x_c, \Omega_c, \phi_c) &= \langle y(x),\, RF(x; x_c, \Omega_c, \phi_c) \rangle \\ &= a(x_c, \Omega_c)\cos(\Psi(x_c, \Omega_c) - \phi_c) \end{aligned}$$

As this equation shows, the maximum response $a(x_c, \Omega_c)$ is achieved for symmetry value $\phi_c = \Psi(x_c, \Omega_c)$. If we take this value of ϕ_c for all (x_c, Ω_c) pairs, the response can be displayed as a 2D function of x_c and Ω_c. An example cortical response $r(x_c, \Omega_c)$ is shown in figure 4, for the same acoustic signal as figure 1.

Effectively, the cortical response is calculated by integrating energy over the tonotopic frequency axis with a variety of RF weighting functions. A subset of these neural RFs, centered around scale $\Omega_c = 0.34$ cyc/oct, have excitatory bandwidths approximately equal to the bandwidth of spectral integration hypothesized to underlie the COG effect. Focusing on the profile of the cortical response along $\Omega_c = 0.34$ cyc/oct, we define the cortical perceptual COG as the best frequency x_c which gives the maximum response,

$$COG = \arg\max_{x_c} r(x_c, \Omega_c)\big|_{\Omega_c = 0.34 c/o}$$

Perceptual COG calculation is illustrated in figure 4, where the dotted line represents the cortical response profile along the $\Omega_c = 0.34$ cyc/oct contour (solid line), and the dashed vertical line represents the COG frequency. It is important to point out that the auditory spectrum is typically time-varying. Hence, the cortical response and perceptual COG are time varying as well. In all modeling, the input stimuli is split into 10 ms frames, and a cortical response and COG are calculated for each frame. The COGs are then averaged to produce an average COG, μ_{cog}.

Average COG values were used to model the results from two listening experiments. The first, reported in Assmann (1991), was a vowel perception experiment in which listeners matched the phonetic quality of multi-formant synthetic vowels. In this experiment, the reference stimuli were 6 formant vowels which had harmonics in the region of F2 and higher scaled by -20, -10, 0, +10, and +20 dB. This had the effect of lowering (-20 and -10) or raising (+10 and +20) the perceptual COG in the F1-F2 formant region, a region crucial in determining phonetic quality. Three different reference (F1,F2) conditions were tested: (350, 500); (450, 700); and (550, 800) Hz. The matching stimuli were 6 formant vowels with equal F1 and F2 amplitudes, where F2 = F1 + 250 Hz. Listeners adjusted F1 in order to match the phonetic quality of the reference vowel under consideration.

The second experiment, reported in Feth (1982), was a psychoacoustics experiment in which listeners matched the pitch of two-tone signals. In this experiment, the reference stimuli were composed of two sinusoids at frequencies F1 and F2, where $F1 = f_c - \Delta f/2$ and $F2 = f_c + \Delta f/2$, with amplitude ratios (A2/A1) of -3, -1, -0.5, 0.5, 1, and 3 dB. For $A2/A1 > 0$, the X_H stimuli, the listeners hear a pitch closer to F2, while for $A2/A1 < 0$, the X_L stimuli, listeners hear a pitch closer to F1. This is indicative of pitch being determined by some sort of COG mechanism. Fixed center frequencies f_c equal to 500, 1000, and 2000 Hz were used, along with Δf values equal to 10, 20, 50, and 100 Hz. The matching stimuli were composed of two equal amplitude sinusoids, at frequencies $f_c - \Delta f/2$ and $f_c + \Delta f/2$ Hz. For the matching stimuli listeners hear a pitch equal to f_c Hz. Listeners adjusted the matching stimuli f_c to match the perceived pitch of the reference stimuli.

Model predictions were obtained for both experiments by matching reference stimuli μ_{cog} to matching stimuli μ_{cog}. In both experiments, the matching stimuli μ_{cog} were monotonically increasing with increasing stimulus parameter (F1 in Assmann, f_c in Feth). For each reference stimulus, the matching stimulus with equal μ_{cog} was taken as the model prediction.

Results and Discussion

The results of the Assmann experiment are shown in figure 5. The plots show the experimental matched F1 frequencies for the subjects, along with model predictions. Dashed lines indicate the reference F1 and F2 frequencies. The results indicate that for the 0 dB condition, listeners matched the reference and matching F1 frequency, as expected. In the +10 and +20 dB conditions, listeners increased the matched F1 frequency, as expected if listeners were matching the perceptual COG of the F1/F2 region. However, in the -10 and -20 dB conditions, listeners matched the reference and matching F1 frequencies, rather than decreasing the matched F1 frequency to compensate for an expected decrease in the reference F1/F2 perceptual COG. Assmann interpreted this result as inconsistent with the COG hypothesis. However, the model predicts the same pattern of results in all conditions. This suggests that the reference stimuli perceptual COGs, in fact, did not decrease in the -10 and -20 dB conditions relative to the 0 dB condition. While the COG does decrease if calculated as the mean frequency of a 3.5 Bark region centered between F1 and F2, it does not decrease when calculated using the pattern of cortical responses as described in the previous section. We therefore conclude that a COG mechanism is not inconsistent with the results.

A subset of the results of the Feth experiment is shown in figure 6. The plots show the average difference between X_H and X_L matched f_c values for each f_c, Δf, and ΔI combination. For example, consider the upper-left-most data point (f_c = 500 Hz, Δf = 10 Hz, ΔI = 3dB), where the matched frequency difference is equal to 5.5 Hz. This means that the matched f_c for the +3dB (X_H) stimulus was 5.5 Hz higher than the matched f_c for the –3dB (X_L) stimulus. The model predictions are shown by the solid and dashed lines. As shown, the model predictions match the trends of the experimental results.

These results support the following conclusions. First, the mechanism underlying the COG effect might be a place coding mechanism in which the perceptual COG is represented as the peak excitation in a subset of cortical neurons. More specifically, the COG may be encoded as the best frequencies of the collection of neurons, tuned to a scale of approximately 0.34 cyc/oct and maximum symmetry, which have maximum excitation in response to an input stimulus. This conclusion is supported by the modeling results presented in this work. Second, the perceptual COG effect involves a COG calculation more involved than simply calculating the mean frequency of a region of the input stimulus spectrum, as demonstrated by the Assmann experimental and modeling results. This is significant because it suggests that SSC features, which rely upon calculating simple moments of the stimulus power spectrum, do not closely model the underlying perceptual process, and that they may be improved by incorporating the cortical COG calculation proposed in this work. Third, the COG effect is not a speech-specific effect, but a fundamental property of the auditory system, as demonstrated by the Feth results.

Figures

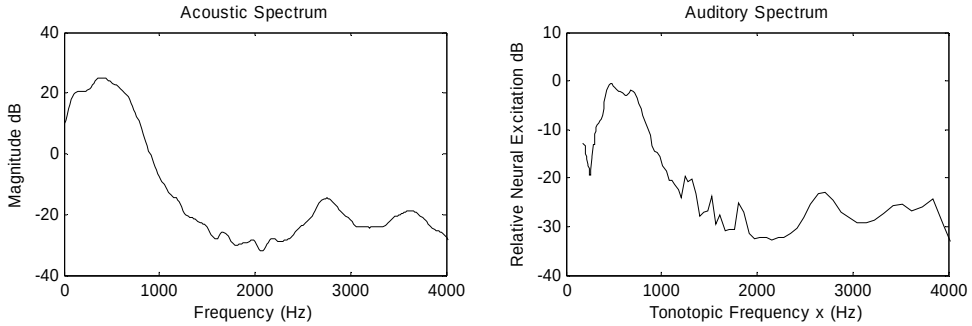

Figure 1 Comparison of Acoustic Spectrum (left) and auditory spectrum (right) of a 10 ms synthetic vowel.

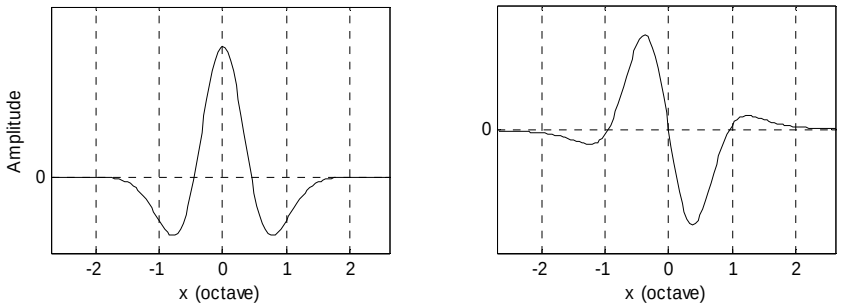

Figure 2 Plot of $h(x)$ (left) and $\hat{h}(x)$ (right), used to generate receptive fields.

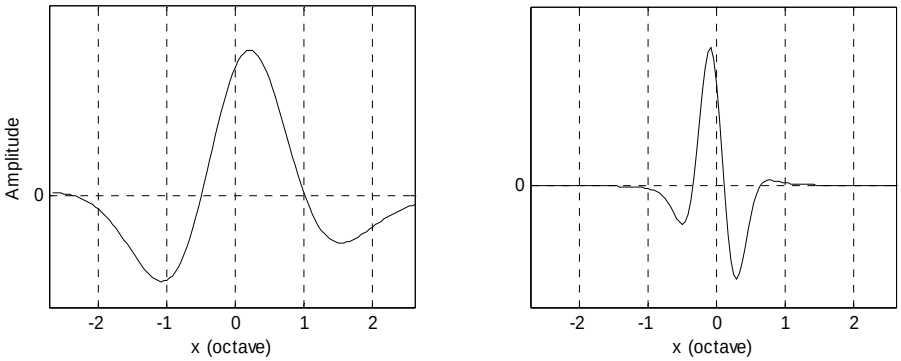

Figure 3 Example RFs. The left RF is tuned to x_c = 0 oct, Ω_c = 0.3 cyc/oct, and $\phi_c = \pi/6$ rad. The right RF is tuned to x_c = 0 oct, Ω_c = 1 cyc/oct, and $\phi_c = -\pi/4$ rad.

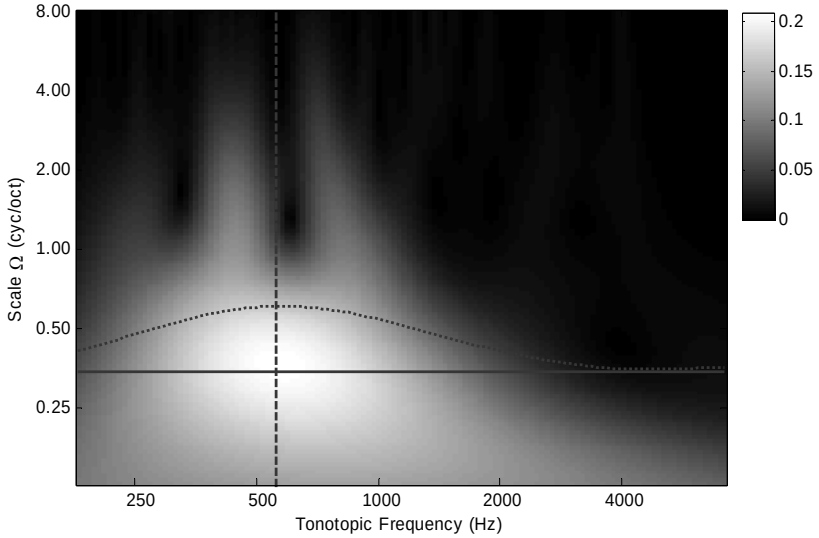

Figure 4. Example Cortical Response. The solid line indicates

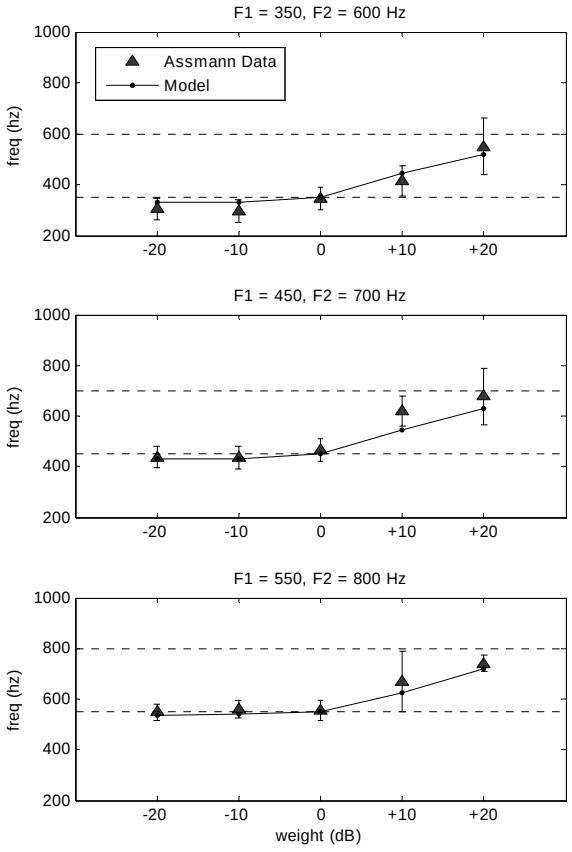

FIG. 5 Assmann (1991). Experimental results show the mean matched F1 frequencies, with bars representing +/- 1std. Model predictions are shown by solid lines.

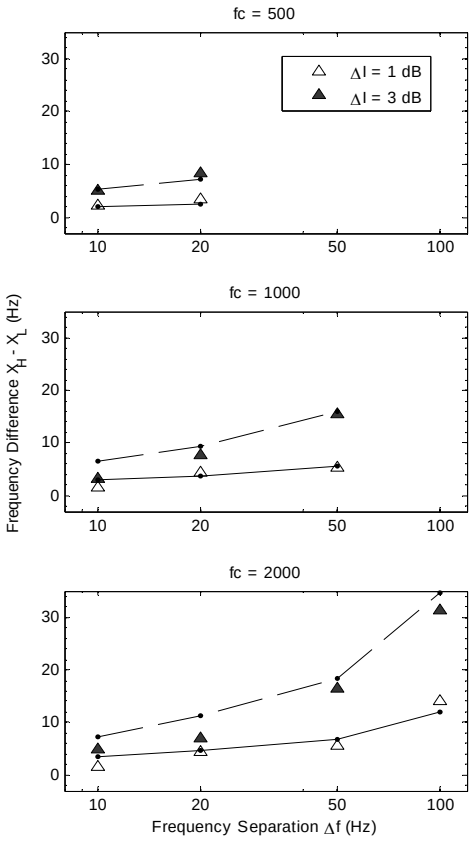

FIG. 6 Feth (1982). Experimental results show the mean matched frequency difference between corresponding X_H and X_L stimuli. Model predictions are shown by the solid and dashed lines.

Acknowledgments
The author would like to thank OSGC for support, Ashok Krishnamurthy and Larry Feth for guidance, and Shihab Shamma and Taishi Chi for help in understanding the model.

References
1. Peter F. Assmann. The Perception of Back Vowels: Centre of Gravity Hypothesis. *Quart. Journal Exp. Psych*., 43A(3):423-448, 1991.
2. Jingdong Chen et al. Recognition of Noisy Speech Using Dynamic Spectral Subband Centroids. *IEEE Signal Processing Letters*, 11(2):258-261, Feb. 2004.
3. Lawrence F. Feth et al. Pitch of Unresolved, Two-component complex tones. *J. Acoust. Soc. Am*, 72(5):1403-1412, Nov. 1982.
4. Kuldip K. Paliwal. Spectral Subband Centroid Features for Speech Recognition. *Proc. of ICASSP '98*. 2:617-620, May 1998.
5. Huib Versnel and Shihab A. Shamma. Spectral-ripple Representation of Steady-State Vowels in Primary Auditory Cortex. *J. Acoust. Soc. Am*., 103(5):2502-2514, May 1998.
6. Kuansan Wang and Shihab A. Shamma. Spectral Shape Analysis in the Central Auditory System. *IEEE Trans. Speech Audio Process.*, 3(5):382-395, Sept. 1995.
7. Xiaowei Yang and Shiha A. Shamma. Auditory Representations of Acoustic Signals. *IEEE Trans. Info. Theory*, 38(2):824-839, March 1992.