Dynamics of Speech Production and Perception
P. Divenyi et al. (Eds.)
IOS Press, 2006

Dynamic Center-of-Gravity Effects in Consonant-Vowel Transitions

Lawrence L. FETH, Robert A. FOX, Ewa JACEWICZ and Nandini IYER
Department of Speech and Hearing Science
The Ohio State University
Columbus, Ohio, USA

Abstract. Lublinskaja (1996) has reported that dynamic changes in the spectral center-of-gravity (c-o-g) in selected Russian vowels led to changes in vowel identification. Movement of the c-o-g was effected by simultaneous amplitude modulation of two formants placed at the end points of the desired frequency transition. Experiment 1 of the present study explored whether c-o-g effects extend into the processing of consonant-vowel transitions in /da/-/ga/. Three different stimulus sets were synthesized in which the F_3 transition was a formant, a frequency modulated (FM) tone, or a Virtual Frequency (VF) glide. Listeners' identification of /da/ or /ga/ was not affected by changing the means by which spectral changes were made to F_3. Experiment 2 examined whether subjects could identify the type of F_3 transition in /da/ (formant, FM tone, VF glide) after a short training period. Responses did not differ with the transition type, thus, processing of transition information does not depend on the method used to elicit the perception of a frequency change. Experiment 3 was conducted to eliminate a possible confounding of transition cues in the VF stimuli used in Experiments 1 and 2 and to test listener performance in a dichotic listening condition. The results indicate that the dynamic c-o-g effect is evident in identification of English CV's just as it was for Russian vowels. The results lend support to the proposition that neural activity rather than signal energy is summed in the spectral integration process.

Keywords. Spectral integration, center-gravity-effect, virtual frequency glide

Introduction

When two formant peaks of a vowel are close in frequency, their combined energy leads to a percept much like that of a single formant peak. This phenomenon has been variously called spectral integration, formant averaging, or the center-of-gravity (c-o-g) effect. Chistovich and Lublinskaja [3][2] have suggested that the frequency separation between the formants and the relative amplitude of the formant peaks are important factors in spectral integration. A recent review of this topic may be found in Xu *et al.* [13].

Lublinskaja [8] demonstrated spectral integration for speech sounds in which the spectral centroid was changed dynamically. She asked listeners to identify three-formant synthetic Russian vowels in which F_2 and F_3 were modulated so that as the amplitude of one formant increased over time, the amplitude of the other decreased. When the amplitude of F_2 decreased while the amplitude of F_3 increased, the resulting percept was

phonetically categorized as similar to a two-formant diphthongal vowel with a rising F_2. Conversely, when the amplitude of F_2 increased as that of F_3 decreased over time, listeners categorized the signal as similar to a two-formant diphthongal vowel with a falling F_2. Lublinskaja reported that the ability of listeners to identify these sounds was limited to critical separations between F_2 and F_3 of roughly 4.3 Bark (or 6 ERBu).[1]

Anantharaman [1] investigated spectral integration for dynamic sinusoidal signals suggested by the results reported by Lublinskaja. He generated two-component signals called virtual frequency (VF) glides. The amplitudes of the tones were modulated so that the intensity of one tone (usually the higher frequency component) increased while that of the other (lower frequency component) decreased. The changes in amplitude gave rise to the perception of a changing pitch, similar to that of a frequency modulated (FM) tone. Anantharaman's results show that listeners can match the rate of frequency transition in a virtual glide to the rate of change of an FM glide. These results were used to extend the Perceptual Centroid Model that had been previously applied to static signals. Further, he found that the dynamic c-o-g effect extended to at least 5.6 Bark (8 ERBu) and, thus, far exceeded the frequency separation reported by Lublinskaja.

Iyer [6] further investigated the processing of VF signals in temporal acuity and temporal masking experiments. The goal was to determine what is integrated in the "spectral integration" process. If the process is mediated near the periphery, signal energy may be the variable of interest; however, at more central locations the signal is represented by patterns of neural activity. Listener performance in a step- vs. linear glide discrimination task [9] was compared for FM and VF signals. Just-discriminable step durations were measured for three conditions: type of signal (FM and VF), center frequency (1000 and 4000 Hz), and frequency separation (2, 5, and 8 ERBu). The just discriminable step was found to be approximately 10 ms for both signal types at 1 kHz at 5- and 8 ERBu separations. All estimates were higher at 2 ERB (approximately 14 to 19 ms) and the 4 kHz result at 8 ERBu was about 12 ms. Iyer offers plausible explanations for the differences from Madden and Feth [9] and, in general, concludes that the temporal acuity task indicates that both FM and VF signals undergo similar processing in the auditory nervous system.

Using FM and VF glides as maskers for brief tones located at the center frequency of the glide led Iyer to very different results. When the probe was placed near the onset or the end of either masker, results were predictable from the power spectrum model of masking [11]. However, when the probe was placed at the temporal center of the glide signal, the shift in threshold was very different for the FM and VF maskers. As expected, the FM glide increased the threshold of the probe by approximately 30 dB for all three frequency separations. While there was substantial masking by the VF signal at 2 ERBu, there was much less at 5 ERBu and almost none at 8 ERBu. Thus, FM and VF signals are processed quite differently at the level of the inner ear, where direct masking effects are assumed to be mediated.

The large difference between masking produced by equivalent VF and FM maskers indicates that, at the periphery, the representation of these two signals is quite different. However, the similarity of temporal acuity results leads us to infer that they are represented by similar neural activity patterns higher in the central auditory system. Thus, the locus of spectral summation must be higher in the central auditory system. Said another way, it is neural activity, rather than signal energy, that is summed in this spectral integration process.

The study reported here addresses the phonetic processing of VF signals for consonant-vowel (CV) transitions in /da/ and /ga/. It builds on the work reported by Lublinskaja [8] and replicated by Iyer *et al.* [7]. The main question is whether VF and FM transitions are processed in a way that makes them perceptually equivalent to that of a synthetic formant transition. If so, then we may conclude that when neural activity moves from one frequency location to another, the percept will be the same for a variety of stimulus configurations.

Extending the investigation of the dynamic c-o-g effect observed in diphthongal vowels to CV transitions is dictated by the dynamics of speech and the role that amplitude changes may play in addition to frequency changes in processing of larger units of speech. Selecting /da/ and /ga/ as CV units for the present investigation is particularly well-suited because the direction of F_3 transition has shown to differentiate /da/ from /ga/ perceptually in three-formant synthetic approximations of the syllables [4][10][12]. The auditory distinction between /da/ and /ga/ is determined by the slope of the transition at the onset of the third formant (F_3): a rising transition specifies /g/ and a falling transition leads to the percept of /d/. Replacing the F_3 transition with an FM- or VF-glide should not cause a change in the perception of syllables as /da/ or /ga/ in response to either stimulus type. Furthermore, if the synthetic CV's are well-matched, listeners may not be able to easily discern the type of the transition.

1. Method

1.1 Subjects

There were 13 listeners in Experiment 1, 11 listeners in Experiment 2, and 8 listeners in Experiment 3. Some individuals participated in all three experiments, some in two of the three and a few participated in only one of them. All were native speakers of American English and were graduate students or research affiliates at The Ohio State University. All subjects reported normal hearing.

1.2 Stimuli

The test signals consisted of synthesized versions of the consonant-vowel (CV) tokens of /da/ and /ga/ based on those described in Fox *et al.* [4]. A 'base' token (the CV base) was generated at a sampling rate of 10 kHz using the parallel version of the Klatt synthesizer. It consisted of a 50-ms transition portion and a 200-ms steady-state portion. The transition portion consisted of F_1 and F_2 only, whereas the steady-state portion contained the first three formants of the vowel. For the F_1 transition, the initial frequency was 279 Hz and the final was 765 Hz. For the F_2 transition, the frequencies were 1650 Hz and 1230 Hz, respectively. The final frequency of both formants in the transition portion of the CV base corresponded to the frequencies of the steady-state portion. Fundamental frequency changed linearly from 120 Hz to 110 Hz. The base token did not contain a stop release burst.

Three types of F_3 transitions were added to the CV base to obtain three different continua: (a) Klatt-synthesized formant transition, (b) FM glide, and (c) VF glide. These parameters were changed for Experiment 3 for reasons discussed below. For the first

continuum, the F_3 transitions were generated using the Klatt synthesizer. The final frequency of F_3 was 2527 Hz, which also corresponded to the F_3 frequency of the steady-state portion. The initial frequency of F_3 transitions ranged from 2018-2818 Hz in 80-Hz steps, creating an 11-step continuum.

For the second continuum, the F_3 transition was replaced by an FM glide, which had the same parameters (i.e., duration, and initial and final frequencies) as the Klatt-generated formant transition. The FM glides were generated in Matlab 5.3 and added to the base signal. The initial frequency of the glides was changed in 80-Hz steps to obtain an 11-step continuum.

For the third continuum, a VF glide was generated by simultaneously modulating the amplitudes of two fixed-frequency tones. For Experiments 1 and 2, the frequencies of the tones corresponded to the initial and final frequencies of the Klatt-synthesized F_3 transitions. Since the final frequency of the synthesized F_3 transitions was always 2527 Hz, the frequency of one of the tones in the VF glide was always 2527 Hz. The frequency of the other tone was varied in 80-Hz steps to obtain an 11-step continuum. In order to achieve a rising F_3 transition (as in /ga/), the relative amplitude of the lower frequency was 14 dB more intense; over time, the amplitude of the lower frequency decreased, while that of the higher frequency increased, until the higher frequency was 14 dB more intense at the end. The amplitude differences created an effect of c-o-g "movement" over the 50 ms transition portion. A downward F_3 transition (as in /da/) was obtained by reversing the amplitude modulations. Similarly, the relative amplitudes of the two tones changed over a 14-dB range. The virtual glides were generated and added to the CV base using Matlab 5.3. The amplitudes of the FM glides and the VF glides were adjusted so that the rms amplitude of the transition portion matched the rms amplitude of the Klatt-synthesized tokens.

For Experiment 3, the frequencies of the tones were fixed at the initial and final frequencies of the Klatt-synthesized F_3 transitions (2018 Hz and 2658 Hz, respectively). To produce a rising VF glide, the amplitude of the 2018 Hz tone decreased linearly over the 50 ms duration as the amplitude of the 2658 Hz tone increased. For a falling VF glide, the amplitude of the higher frequency tone declined as the lower frequency tone increased in amplitude. Initial and final amplitudes were chosen to make the signal c-o-g traverse the desired frequency range for each of the 8 tokens generated. The virtual glides were generated and added to the CV base using Matlab 5.3. The amplitudes of these VF glides were adjusted so that the rms amplitude of the transition portion matched the rms amplitude of the Klatt-synthesized tokens.

Experiment 3 was conducted several months after the first two experiments in order to minimize possible confounding in the VF signals used for the first two experiments. Because the frequency of the lower tone in the VF pair was changed for each token in Experiment 1, it was suggested that listeners might base their identification on the locus of that tone even if no formant-like transition was produced by the amplitude modulations.

Figure 1 displays spectrograms of the three token types for /ga/. The top panel (a) shows the full CV. In the lower panels, the F_3 transition has been isolated and displayed in a separate channel: (b) the Klatt-synthesized F_3 transition, (c) the FM glide, and (d) the VF glide. For the dichotic presentations in Experiment 3, the second channel was delivered to the earphone contra-lateral to the one containing the CV-base signal. Signals were played at a 10 kHz sampling rate, via the TDT system II, with low-pass filtering at 5 kHz.

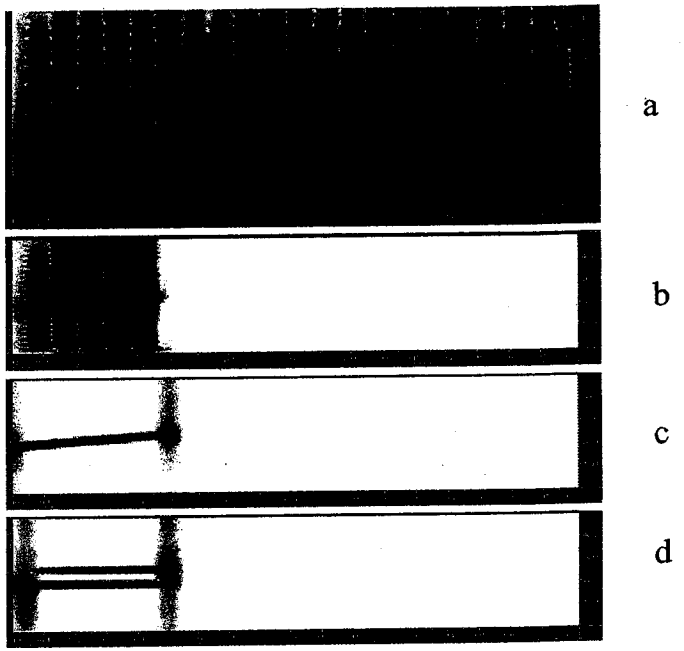

Figure 1. Spectrograms of the synthetic speech stimuli used in the experiments. Panel (a) displays the base Klatt-synthesized /ga/ CV. Notice the "missing" F_3 transition. The Klatt-synthesized F_3 transition alone is shown in panel (b). Panel (c) shows a frequency-modulated tone substituted for the formant on FM trials, and panel (d) shows a virtual frequency transition used in VF trials.

1.3 Procedure

For Experiments 1 and 2, signals were presented monaurally via Sennheiser HD 580 headphones to a subject seated in a sound-attenuating booth. In Experiment 1, a single-interval 2-AFC identification task was used with the response choices /da/ and /ga/ displayed on the computer screen. Subjects were asked to indicate whether they heard a /da/ or a /ga/ for each token presented. There were 660 stimuli presented randomly in 3 experimental blocks (3 transition types x 11 tokens x 20 repetitions). In Experiment 2, subjects responded in a single-interval, 3-AFC identification task indicating whether the token of /da/ they heard was generated with a Klatt formant transition (K), an FM Tone transition (T), or a VF transition (V). The three choices (K, T, or V) were displayed on the screen. Here, 120 stimuli were presented in one block (4 /da/-tokens, 3 transition types x 10 repetitions). Subjects were trained prior to the task by listening to 80 trials of /da/-tokens blocked by each transition type and presented in the following order: Klatt-synthesized formant transition, FM transition, and VF transition.

For Experiment 3, signals were presented binaurally via Sennheiser HD 580 headphones to a subject seated in a sound-attenuating room. The response task was identical to that used in Experiment 1. Half of the presentations were diotic (same signal to both ears). For the other half of the blocks of trials, the CV base signal was delivered to one ear and the Klatt, FM or VF transition was delivered to the contra-lateral ear. Six of the eight tokens were presented 20 times each in diotic and dichotic listening conditions. Results are averaged over the eight subjects.

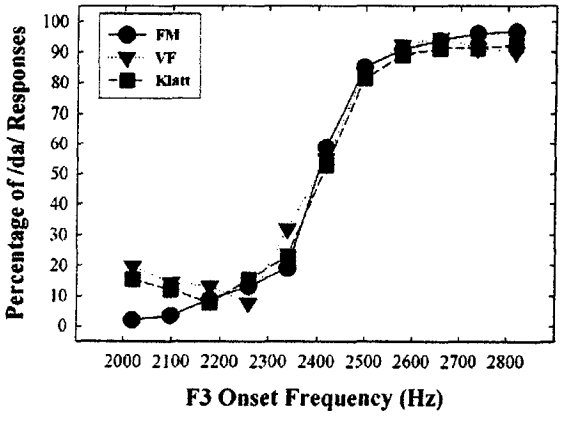

Figure 2. Identification functions for the CV stimuli presented in Experiment 1. Squares denote Klatt transitions; circles represent FM transitions; and inverted triangles are plotted for VF transitions. Results are averaged for 13 listeners.

2. Results

Figure 2 presents the averaged identification functions for 13 subjects in Experiment 1. The abscissa shows starting frequency for the F_3 formant transition. Each datum corresponds to a point on the 11-step /ga/ - /da/ continuum. The ordinate is percentage of /da/ identifications. Thus, the lowest starting frequency was rarely identified as a /da/ sound while the highest one was most often labeled as /da/. Squares represent the Klatt-synthesized transition; circles represent the FM tone transitions; and inverted triangles represent VF glide transitions.

The identification data were analyzed in two separate ways. First, the total number of /ga/ responses (summed across all 11 steps) for each subject were analyzed using a repeated-measures analysis of variance with the factor stimulus type (Klatt, FM glide, VF glide).

There was no significant effect of stimulus type ($F[2,24] = 2.726$, $p > .08$). Second, the category boundaries were analyzed in the same manner. The category boundary (the 50% cross-over point) for each subject for each stimulus type was calculated using PROBIT analysis. Again, there was no significant effect of stimulus type ($F[2,24] = 0.026$, $p > .90$). Both results appear to support the claim that the precise manner in which perceived frequency changes in the F_3 transition were elicited did not affect the phonetic category of the stimulus.

Experiment 2 examined whether listeners can identify the F_3 transition type (Klatt, FM glide, VF glide) after a short training in responding to each stimulus type. The last four tokens on the 11-step continuum with a gradually falling F_3 transition were selected for identification as /da/. Identifications averaged over 11 subjects reached 37% correct for the Klatt-synthesized transition, for the FM glide 32.3% correct, and for the VF glide 32% correct. Since chance performance equals 33.3%, the results indicate that the subjects did not differentiate between the transition type in their identifications of the stimuli as /da/. Results are shown in Table 1.

Stimulus/ Response	FM	VF	KL
FM	32.3	29.3	28.9
VF	34.3	32.0	34.1
KL	33.4	38.2	37.0

Table 1. Results of Experiment 2. Averaged performance for 11 subjects using the last four tokens of the continuum. Listeners were asked to identify the type of transition: Klatt-synthesized (KL), Virtual Frequency (VF), or Frequency modulated tone (FM). Entries are the percentage of responses for each stimulus type.

Experiment 3 was conducted to investigate the possibility that identification of a CV with the VF glide substituted for F_3 could have been cued by the difference in the frequency of the lower tone. Recall that the lower frequency for each virtual glide was changed to produce the F_3 "virtual" transition in the signals used in Experiments 1 and 2. For Experiment 3, the frequency of the lower tone was fixed, and the virtual transition was generated solely by changing the depth of amplitude modulations imposed on the tone pair. Signals representing six of the first eight steps (steps 1,2,4,5,7,8) in the original 11-step /da/ - /ga/ continuum were used in this experiment. For the diotic listening condition, only FM and VF transitions were tested since there was such good agreement between the Klatt signals and the FM transition signals in the earlier tests. Eight listeners heard 20 repetitions of each token in both diotic and dichotic listening modes. Results are shown in Figure 3 (a – diotic; b – dichotic). The psychometric functions plot the percentage of /da/ responses for each token averaged over eight listeners. For the diotic condition, the identification function for the FM transition is very similar to that shown in Figure 2, but the function for the VF transition is shifted to the left in frequency. The slope of the VF ID function is also less steep than that of the FM signals.

The /da/-/ga/ category boundaries (representing the 50% cross-over point) for the tone and virtual glide continua were calculated for each subject using PROBIT analysis. These boundaries were then analyzed using a within-subject analysis of variance. Results showed a significant difference in the category boundaries between the two stimulus sets [$F(7,1)=39.2$, $p<.001$, $\eta = 0.85$]. The mean category boundaries for the tone and virtual glide continua were 2346 Hz (= 38.5) and 2185 Hz (= 37.5), respectively.

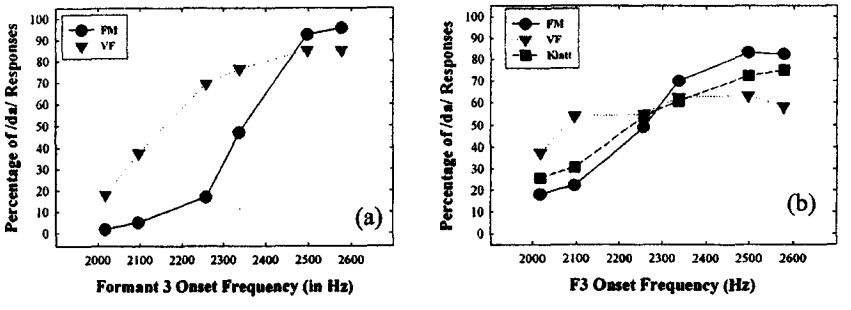

Figure 3. Identification functions for the CV stimuli presented in Experiment 3. Only FM and VF transitions were tested in the diotic (a) listening condition. Klatt, FM, and VF transitions were used in the dichotic (b) listening condition. Symbols are the same as in Figure 2.

For the dichotic listening condition, the three ID functions are shown in Figure 3b. The coordinates are the same as those for Figure 3a. Here, the slope of the function for FM transitions is steeper than that for the Klatt-synthesized transition, and the function for the VF transition is much flatter.

3. Discussion

This study addressed the question of whether VF and FM CV transitions are processed in a way that makes them perceptually equal to that of a synthetic formant transition. F_3 transitions in a /da/ - /ga/ continuum were replaced by FM and VF glides, respectively. Results from Experiment 1 showed no significant difference in the identification responses as a function of stimulus type. Similarity in perception and processing of VF and FM signals has been manifested in the phonetic processing of a CV unit, whose identification as /da/ or /ga/ was affected only by the direction of F_3 transition and not by the means by which the dynamic information was delivered to the listener's central nervous system. This shows that phonetic processing as well as interpretation of transitional information are independent of method used to elicit perception of frequency change. Or, how "excitation" is moved from one place to another has little effect on the listener's identification of the CV.

The third experiment was conducted to account for possible confounding in the VF transitions used in the first experiment. In addition, a dichotic listening condition patterned after Whalen and Liberman [12] was added. The identification function for VF transitions was shifted to the left of that for the FM (and Klatt) transitions. This may reflect a small bias in the calculation of the dynamic center-of-gravity for each token and require some minor modifications to the Perceptual Centroid Model. Listeners do hear the stimuli as belonging to the /da/ - /ga/ continuum. The effect is further diminished in the dichotic condition, where the FM transitions produce greater differences between the end points of the continuum than the original Klatt-synthesized signals. Clearly, further work is needed before we can assert that there is no difference in processing of VF, FM and Klatt-synthesized CV sounds.

Experiment 2 was conducted to ensure that perception of frequency change was not a consequence of presenting all three token types in the same experimental block for a phonetic identification as /da/ or /ga/. It could be argued that the nature of the task provided details which otherwise would not come into play in processing a particular type of F_3 transition. The listeners were asked to respond directly to the type of the transition after a reasonable amount of practice with each token type, ignoring the phonetic content of the syllable /da/. The results showed that listeners could not differentiate between the type of transition, not being able to indicate whether they heard a synthetic Klatt-version, an FM glide-version, or a VF glide-version of the transition in /da/. This implies that the dynamic character of F_3 transitions take perceptual precedence over the amount of information about the transition itself, be it frequency change of a formant, frequency change of a tone, or amplitude modulation of a signal. These details tend to be ignored if they do not contribute crucially to the identity of a phonetic unit.

Results from these experiments indicate the importance of dynamic information in speech processing. Extending the investigation of the c-o-g effect observed in diphthongal vowels to CV transitions revealed that the dynamic change caused by amplitude modulation is a phenomenon comparable to a frequency change. This "virtual" frequency change is processed similarly in acoustic and speech signals and the processing occurs more centrally along the auditory system.

Acknowledgments

This work was supported by a grant from The Ohio State University, College of Social and Behavioral Sciences to L. L. Feth and an INRS Award from NIH to R. A. Fox.

Note

1. Estimates of the width of the peripheral auditory filter, once known as the critical bandwidth, were initially given in units named "Barks," based primarily on work done in Zwicker's laboratory. More recent work, characterized by that from the laboratories of Patterson and Moore in Cambridge, has shown that the earlier estimates were systematically wider than now thought. The more recent unit for this estimate of auditory filter width is commonly called the Equivalent Rectangular Bandwidth unit, or ERBu. The original 3.5 Bark noted by Chistovitch and Lublinskaja [3] is approximately equal to 5 ERBu. Chapter 10 in Hartmann's book [5] has a discussion of the historical and technical details.

References

[1] Anantharaman, J.N. *A dynamic multi-channel perceptual spectral-centroid model for the processing of speech and other complex sounds.* Doctoral dissertation, The Ohio State University, Columbus, OH, 1998.

[2] Chistovich, L.A. "Central auditory processing of peripheral vowel spectra." *J. Acoust. Soc. Am.* 77: 789-804, 1985.

[3] Chistovich, L.A. and Lublinskaja, V.V. "The 'center of gravity' effect in vowel spectra and critical distance between formants: Psychoacoustical study of perception of vowel-like stimuli." *Hearing Research* 1: 185- 195, 1979.

[4] Fox, R.A., Gokcen, J. and Wagner, S. "Neurophysiological and behavioral evidence for a phonetic processor." *Proc. Chicago Linguistic Society's Thirty-third Meeting,* Vol. 33-2, pp. 311-332, 1997.

[5] Hartmann, W.M. *Signals, Sound and Sensation.* Woodbury, NY: American Institute of Physics, 1997.

[6] Iyer, N. *Temporal Resolution and Masking Patterns Using Frequency Modulated and Virtual Frequency Signals.* Doctoral dissertation, The Ohio State University, Columbus, OH, 2001.

[7] Iyer, N., Jacewicz, E., Feth, L.L. and Fox, R.A. "Center of gravity effects in the perception of virtual formant transitions." *J. Acoust. Soc. Am.* 109: 2294, 2001.

[8] Lublinskaja, V.V. "The 'center of gravity' effect in dynamics." *Proc. Workshop Auditory Basis of Speech Perception,* pp. 102-105, 1996.

[9] Madden, J.P. and Feth, L.L. "Temporal resolution in normal-hearing and hearing-impaired listeners using frequency-modulated stimuli." *J. Speech Hearing Research* 35: 436-442, 1992.

[10] Mann, V.A. and Liberman, A.M. "Some differences between phonetic and auditory modes of perception." *Cognition* 14: 211-235, 1983.

[11] Patterson, R.D. and Moore, B.C.J. "Auditory filters and excitation patterns as representations of frequency resolution." In *Frequency Selectivity in Hearing,* B.C.J. Moore (ed.), London: Academic Press, 1986.

[12] Whalen, D. and Liberman, A.M. "Speech perception takes precedence over non speech perception." *Science* 237: 169-171, 1987.

[13] Xu, Q., Jacewicz, E., Feth, L.L., and Krishnamurthy, A.K. "Bandwidth of spectral resolution for two-formant synthetic vowels and two-tone complex signals." *J. Acoust. Soc. Amer.* 115: 1653-1664, 2004.