19th INTERNATIONAL CONGRESS ON ACOUSTICS
MADRID, 2-7 SEPTEMBER 2007

SPECTRAL INTEGRATION OF VIRTUAL CUES IN SPEECH PERCEPTION

PACS: 43.71.Es

Fox, Robert Allen; Jacewicz, Ewa; Wackler, Lisa
Speech Perception and Acoustics Laboratories, The Ohio State University, 1070 Carmack Rd., Columbus, OH, 43210, USA; fox.2@osu.edu; jacewicz.1@osu.edu; wackler.4@osu.edu

ABSTRACT
A number of experiments have established that the alveolar and velar place distinction in initial stop consonants ([d] and [g]) can be cued by the slope of the third formant transitions: Rising F3 transitions cueing /g/ and falling F3 transitions cueing /d/. Replacing this F3 transition with a tone glide has no significant effect on the percept of the place distinction as long as the non-speech transition provides the appropriate dynamic spectral information. However, this tone glide may be replaced by a "virtual" glide produced by dynamically modifying the amplitudes of two pairs of steady-state sine waves (producing a dynamic change in the spectral "center-of-gravity" of these sine waves). This demonstrates that the perceptual system depends upon the movement of auditory excitation (rather than frequency change, per se) in processing auditory speech cues. However, the question remains as to the source of this virtual change: Does it occur in the auditory periphery or as a function of central auditory processing? The present experiment examined the salience of actual and virtual F3 transitions in both diotic and dichotic conditions. Both virtual and actual F3 transitions produced similar identification functions but the efficacy of the virtual transition was somewhat reduced in the dichotic condition.

INTRODUCTION
The syllable-initial formant transition in the syllables [dɑ]-[gɑ] is a good example of dynamic acoustic cues used in the perception of speech. In particular, when a stop release burst is not present, it is the direction of F3 transition alone that signals the place of articulation of the initial plosive: a falling or flat F3 transition gives the percept of [dɑ] and a rising F3 transition leads to the perception of [gɑ] (see Figure 1). That the direction of F3 transition can cue the place of articulation distinction of a stop consonant has been well demonstrated with both synthetic stimuli and with tone glides (e.g., [7] [3] [4]).

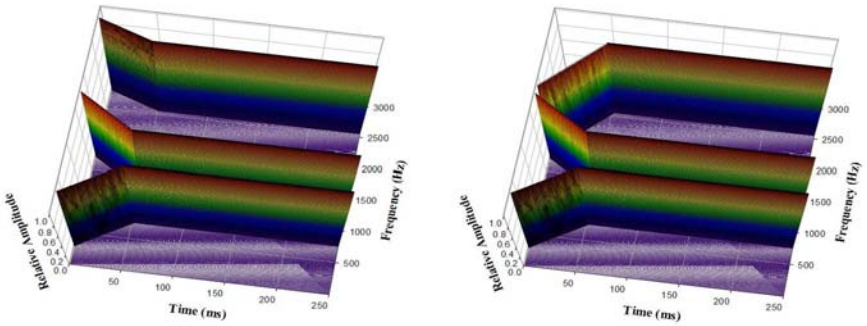

Figure 1. Schematic representation of three-formant syllables [dɑ] (left) and [gɑ] (right) consisting of a 50-ms transition portion and a 200-ms steady-state vowel portion.

The present experiment was conducted to verify whether a "virtual" F3 transition—a transition percept produced by moving the spectral center of gravity (COG) of an acoustic cue rather than changing the frequency of any of its components—can also serve as a cue to place of articulation. The role of the spectral COG in speech perception was first explored by Chistovich

and colleagues (e.g., [1] [2]). The early experiments used static synthetic two-formant vowel approximations. The changes to the relative amplitude ratios between the two closely spaced formants of a vowel changed their combined COG and listeners perceived this change as a vowel category shift. The COG effect was interpreted as an indication that the auditory system performs a type of auditory spectral integration at some level (perhaps a more central level) of auditory processing.

More recently, work on the COG effect showed that modifying the amplitude ratios of two relatively closely spaced formants over time can produce a diphthong percept [6]. These results suggest that the auditory system can attend to a spectral COG that is dynamic, tracking changes in effective frequency over time even if no actual frequency change is present in the acoustic signal. The perception of such "virtual diphthongs" was further investigated in [5] showing that listeners were equally sensitive to both the actual and the virtual frequency changes in making their vowel identifications. The differences between responses to the dynamic formant transitions and virtual transitions were not significant, indicating that movement of the spectral COG could, in fact, provide the cues necessary for the identification of F2 transitions comparably with the actual formant transitions.

The present experiment used a different type of dynamic signals than diphthongal vowels. The initial formant transitions in the syllables [dɑ] or [gɑ] are shorter and occur at a faster rate than the diphthongal transitions. We are investigating whether listeners can attend to changes in the spectral COG in such short transitions to make place of articulation decisions on the basis of the direction of the "virtual" F3 transition. In addition—as an initial step in determining the level of auditory processing at which this spectral integration occurs—we examine whether the mode of presentation of the stimuli (i.e. diotic or dichotic) with the virtual F3 transition affects the salience of the virtual F3 cue.

METHODS

Stimuli

The stimuli were designed for presentation in two listening conditions: 1) diotic (the same signal was presented to both ears), and 2) dichotic (two different signals were presented simultaneously, one to each ear). For diotic listening, the entire syllable [dɑ] or [gɑ] was presented to both ears. In the dichotic condition, the syllable was partitioned so that different components of the syllable were presented simultaneously to different ears.

Stimuli for diotic condition

Two series of stimuli were created for the diotic condition. The first series consisted of tokens containing an actual F3 transition. In the second series, the actual F3 transition was replaced by a virtual F3 transition. These sets will be called the *Actual F3* and *Virtual F3* series, respectively.

The *Actual F3* stimuli were created using the parallel branch of the Klatt synthesizer (the .kld option) in HLSYN (Sensimetrics, 1997) with a sampling rate of 11025 Hz. The duration of each token was 250 ms (including an initial 50-ms transition and a 200-ms steady-state vowel portion). The F1 and F2 transitions, as well as the 200-ms steady-state portion for all three formants (F1, F2, and F3) were identical for all tokens in the series. For the 50-ms transition, F1 increased from 443 to 700 Hz and F2 decreased from 1520 to 1220 Hz. For the 200-ms vowel portion, the frequencies of F1, F2, and F3 remained constant at 700, 1220, and 2600 Hz, respectively. The tokens differed only in terms of the initial F3 transition. The F3 onset frequency varied in nine equidistant steps ranging from 1800 to 2600 Hz each with an offset frequency of 2600 Hz. No stop bursts were present in the signal so that only the F3 transitions provided place of articulation information for the [dɑ]-[gɑ] contrast.

For the *Virtual F3* stimuli, the virtual F3 transition was created separately and then inserted into the base token. The virtual transition consisted of two 50-ms sine wave pairs whose frequencies remained constant. Frequency values of the sine waves were multiples of the fundamental frequency (F0=120 Hz), and were just above and just below the frequencies where the actual F3 transitions occurred. The percept of a dynamic virtual transition was created by changing the

relative intensities of the pairs of sine waves. This was done for each of the nine steps (see Table I). The sine wave pairs were then combined and the intensity adjustments over time (50 ms) gave rise to the perception of a spectral change, similar to that of the frequency change in the actual F3 transition. Reiterating, there was no frequency change per se in the virtual F3 transition and only the changes in the intensity ratio between the lower and higher sine wave pairs produced the perceived change in F3 frequency. The overall mean rms of the combined pairs of sine waves was then adjusted to the average rms value of the actual F3 transitions and was then inserted into the base token (with the onset of the transition synchronized with the onsets of F1 and F2 transitions in the base token).

Table I. Relative intensities of lower and higher pairs of sine waves for each of the nine steps of *Virtual F3* stimulus series. The frequencies of the sine waves remained unchanged throughout.

Step	**Lower pair** 1680 Hz 1800 Hz	**Higher pair** 2640 Hz 2760 Hz
[gɑ] endpoint	0.937	0.063
2	0.833	0.167
3	0.729	0.271
4	0.625	0.375
5	0.521	0.479
6	0.417	0.583
7	0.313	0.687
8	0.208	0.792
[dɑ] endpoint	0.104	0.896

Stimuli for dichotic condition

In the dichotic listening condition, the signal was delivered in two separate channels which were used to simultaneously present different parts of the token to each ear.

The *Actual F3_1* stimuli were constructed with a base token (containing F1 and F2 with the same frequencies and durations previously used) in one channel and the nine different F3 steps (containing both the 50-ms F3 transition and the 200-ms steady-state portion of F3) in the other channel. All frequency values were identical to those used in the diotic condition.

The dichotic *Virtual F3* stimuli were presented in two separate channels as well. However, there were three types of dichotic stimuli, differing by the ways the stimulus components were divided between the ears. In the first stimulus type (*Virtual F3_1*), the base token was delivered in one channel and the virtual F3 transition (and steady-state portion of F3) in the other channel. The virtual F3 transitions were created in the same manner as those in the diotic condition, with the intensities of two sine wave pairs being manipulated in order to change the spectral COG. In the second stimulus type (*Virtual F3_2*), the virtual F3 was partitioned so that the lower frequency sine wave pair was inserted into the base token and presented in one channel while the higher frequency sine wave pair and the steady F3 were presented in the other channel. In the third stimulus type (*Virtual F3_3*), the higher frequency sine wave pair was inserted into the base token and presented in one channel while the lower frequency sine wave pair and the steady F3 were presented in the other channel. All frequency values of the sine waves and the intensity weighting adjustments remained the same as in the *Virtual F3* stimuli for the diotic condition.

Listeners

Fifteen listeners (six men and nine women) with no known history of hearing impairment participated in the experiment. All listeners were native speakers of American English and ranged in age from 19 to 28 years. They were college students enrolled in a variety of majors at The Ohio State University and were paid for their effort.

Procedure

All signals were presented via TDH-49 headphones at a comfortable listening level to a listener seated in a sound-attenuating booth. In both diotic and dichotic presentations, a single-interval

two-alternative forced choice identification task was used with the response choices */da/ as in 'dot'* and */ga/ as in 'got'* displayed separately on two halves of a computer monitor. In response to each token, listeners were asked to indicate whether they heard an instance of [dɑ] or [gɑ] by clicking with a mouse button on the appropriate section of the display. There were 135 stimuli presented randomly in each of six sets (9 tokens x 15 repetitions) blocked by the token type. There were two sets presented in the diotic condition (*Actual F3* and *Virtual F3*) and four sets in the dichotic condition (*Actual F3_1*, *Virtual F3_1*, *Virtual F3_2*, and *Virtual F3_3*). Each listener participated in two separate one-hour sessions, with three sets of 135 stimuli presented in each session. The presentation order of sets was counterbalanced across listeners. A short familiarization task was presented prior to each stimulus set along with a fifteen-item practice with no feedback. After the familiarization task, the actual listening task began. Each of six sets lasted approximately twenty minutes.

RESULTS

Diotic listening

Shown in Figure 2 (left panel) are the identification responses to both the actual and virtual F3 transition series for the diotic listening condition. It is clear that the two identification functions are similar, indicating that the listeners were able to identify the stimuli as instances of either [dɑ] or [gɑ] in a similar fashion regardless of the stimulus type. The primary difference is that the slope of the identification function of the *Actual F3* series is steeper than that of the *Virtual F3* series and the number of *da* responses do not reach the ceiling for the *Virtual F3*. The locations of the [dɑ]-[gɑ] category boundary (the 50% cross-over point) along the F3 onset axis for each individual subject were calculated using PROBIT analysis. An ANOVA with the within-subject factor stimulus type conducted on these data showed that the difference in the mean category boundary between *Actual F3 (*mn=2229 Hz) and *Virtual F3* (mn=2121 Hz) was significant ([F(1, 12)=13.2, p=.003, η^2=.523]). This indicates that the boundary was shifted towards the [dɑ] endpoint for the *Virtual F3* series as compared to the *Actual F3*. The slopes of the identification functions for each listener were determined using linear regression. The slope of the *Actual F3* identification function (mn=.130 pct/Hz) was higher than that of the *Virtual F3* identification function (mn=.080 pct/Hz). This matched the results of a two-way ANOVA of /d/ responses with the within-subject factors stimulus type and F3 onset which showed a significant stimulus type by F3 onset interaction (F(8,104)=16.45, p<.001, η^2=.559). These analyses support the claim that although the virtual F3 transition does provide the same type of information regarding the place of articulation distinction as the actual F3 transition, it is not as salient a cue.

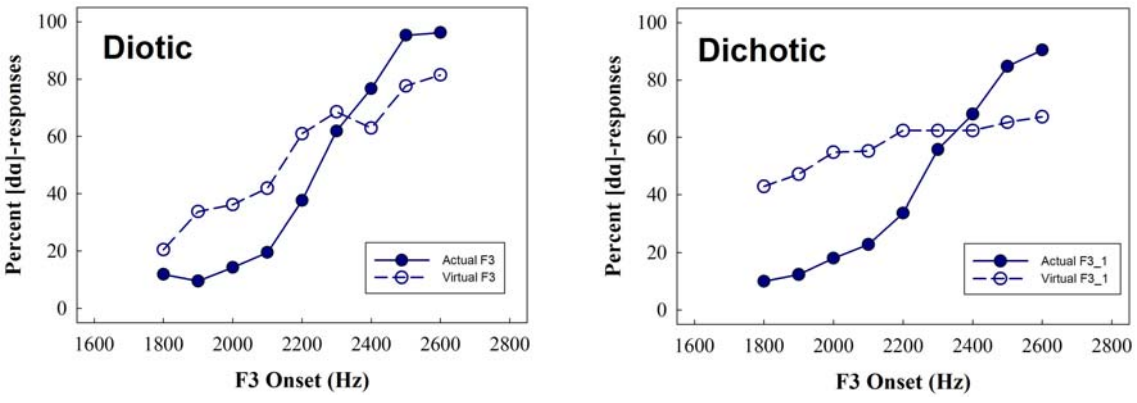

Figure 2. Identifications as [dɑ] and [gɑ] in response to the *Actual F3* and *Virtual F3* stimulus series in diotic (left) and dichotic (right) presentation.

Dichotic listening

The right panel of Figure 2 shows the identification responses to the actual and virtual F3 transition (*Virtual F3_1*) series for the dichotic condition. The identification function for the *Actual*

F3 series is very similar to the *Actual F3* series for the diotic listening condition in the left panel. This indicates that listeners were able to correctly identify the syllables as either [dɑ] or [gɑ] when different components of the syllable were presented to different ears. It needs to be underscored that, although the subjects heard two separate signals and no ear received complete spectral information of a syllable, the crucial frequency information which cues the [d]-[g] distinction, the rising F3 transition, was presented to a different ear.

The virtual F3 transition (*Virtual F3_1*) provided a much weaker cue to place identification. As can be seen in the right panel of Figure 2, the identification function for the *Virtual F3_1* series was much shallower and the difference in responses to the endpoint stimuli was small. Although the ANOVA examining location of the category boundary conducted on PROBIT means showed no significant difference of stimulus type (the mean category boundary was 2273 Hz for *Actual F3_1* ad 2206 for *Virtual F3_1*), these results must be interpreted with caution because the shapes of the identification functions were different. As in the diotic condition, the slope of the identification function for the *Actual F3_1* (mn = .116 pct/Hz) was significantly greater than for the *Virtual F3_1* (mn=.033 pct/Hz). A two-way ANOVA of /d/ responses with the within-subject factors stimulus type and F3 onset showed a significant stimulus type by F3 onset interaction ($F(8,104)=20.85$, $p<.001$, $\eta^2=.616$). As in the diotic condition, these analyses support the claim that although the virtual F3 transition does provide the same type of information regarding the place of articulation distinction as the actual F3 transition, it is not as salient a cue. The mean slopes of the actual F3 identification functions in the diotic and dichotic conditions were almost identical (0.130 vs. 0.116 pct/Hz), but the mean slopes of the virtual identification functions were significantly and more profoundly different (0.080 vs. 0.033 pct/Hz). This suggests that the virtual cue is even less salient in the dichotic condition than in the diotic condition.

Shown in Figure 3 are the identification functions for the three dichotic virtual F3 series (i.e., *Virtual F3_1* redrawn from Figure 2, *Virtual F3_2* and *Virtual F3_3*). It can be seen that the response pattern to *Virtual F3_1* series is different from either *Virtual F3_2* or *Virtual F3_3.* Although the identification function is generally shallow for *Virtual F3_1,* the shape of the identification function generally follows that of the *Actual F3_1* (although its slope is much shallower) which is not the case for the two other virtual F3 conditions. The slope of the *Virtual F3_1* stimulus type (mn=.030 pct/Hz) was greater than both the slope of the *Virtual F3_2* (mn=.009 pct/Hz) and the *Virtual F3_3* (mn=.017), although only the *Virtual F3_1* and *Virtual F3_2* difference was statistically significant. The pattern of the results suggests that the ability of the listener to successfully integrate the two sinewave pairs into dynamic F3 cue is affected by whether the two pairs of sinewaves are heard in the same ear.

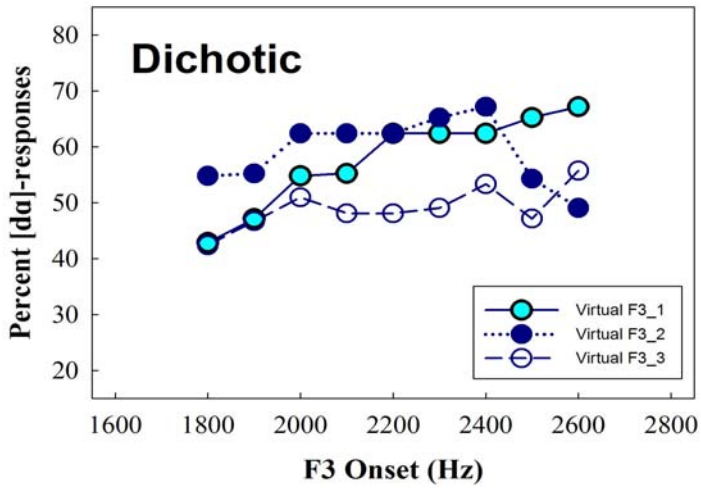

Figure 3. Identifications as [dɑ] and [gɑ] in response to the three types of *Virtual F3* stimuli in dichotic listening condition.

CONCLUSION

In the present experiment we examined whether dynamically modifying the amplitudes of two pairs of steady-state sinewaves (which produces a dynamic change in their spectral COG) could effectively cue place distinctions in initial plosives. These stimuli were presented to the listeners both diotically and dichotically to determine the level of auditory processing that might be involved in perception (and the process of auditory spectral integration) of virtual F3 cues.

The results for the diotic listening condition indicate that when dynamic frequency changes were unavailable to the listeners, dynamic amplitude changes could produce equivalent perceptual cues to the place distinction. Although the identification function for the *Virtual F3* stimuli was shallower than that for the *Actual F3*, the general shape of the function and the direction of listeners' responses were comparable.

However, in the dichotic condition, the virtual transitions provided less salient cues to the place distinction. The *Virtual F3_1* stimuli, in which the virtual F3 transition was presented in one channel and the base token in the other, yielded the best identification function and in the direction expected. This shows that the listeners were able to identify some of the stimuli as [gɑ] despite the apparent difficulty in integrating the information across the two channels. No such indication can be seen in the responses to the two other types of virtual stimuli, *Virtual F3_2* and *Virtual F3_3.* Clearly, the two identification functions show chance performance, indicating that listeners were unable to integrate the spectral information across the channels. The question arises whether they failed to integrate the two pairs of sine waves in order to perceive the virtual F3 transition or to integrate the virtual F3 transition with the base token. Further research will be directed at determining whether the COG effects examined here are a product of peripheral and/or central auditory processing.

ACKNOWLEDGMENTS

This study was supported by the research grant No. R01 DC006879 from the National Institute of Deafness and Other Communication Disorders, National Institutes of Health. The authors wish to thank Jeff Murray for editorial help.

References: [1] Y. A. Bedrov, L. A. Chistovich, R. L. Sheikin: Frequency location of the 'center of gravity' of formants as a useful feature in vowel perception. Akusticheskiĭ zhurnal **24** (1978) 480-486 (Soviet Physics - Acoustics **24** 275-282).
[2] L. A. Chistovich, V. V. Lublinskaja: The 'center of gravity' effect in vowel spectra and critical distance between the formants: psychoacoustical study of the perception of vowel-like stimuli. Hearing Research **1** (1979) 185-195.
[3] R.A. Fox, J. Gokcen, S. Wagner: Neurophysiological and behavioural evidence for a phonetic processor. Proceedings from the Panels of the Chicago Linguistic Society's 33rd Meeting **33-2** (1997) 311-332.
[4] J. Gokcen, R.A. Fox: Evidence for a special speech processing module from electrophysiological data. Brain and Language **78** (2001) 241-253.
[5] E. Jacewicz, R.A. Fox, L. L. Feth: Dynamic auditory representations and phonetic processing: The case of virtual diphthongs. Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics (2006) pp. 153-156.
[6] V.V. Lublinskaja: The 'center of gravity' effect in dynamics. Proceedings of the Workshop on the Auditory Basis of Speech production (1996) pp. 102-105.
[7] D.H. Whalen, A.M. Liberman: Speech perception takes precedence over nonspeech perception. Science **237** (1987) 169-171.