VOWEL PERCEPTION WITH VIRTUAL FORMANTS

Robert Allen Fox, Ewa Jacewicz, Chiung-Yun Chang

Speech Perception and Acoustics Labs
The Ohio State University
fox.2@osu.edu, jacewicz.1@osu.edu, chang.552@osu.edu

ABSTRACT

This study examines the potential role of the auditory spectral integration in phonetic vowel quality decisions. Synthetic stimuli included a "virtual formant (F2)" which was produced by inserting two pairs of sine waves below and above a "perceptual target frequency" (the spectral center-of-gravity, COG). Intensity weighting across the pairs of sine waves created a virtual F2, i.e., an F2 percept which listeners formed although the formant was not physically present in the signal. Two different vowel series containing a virtual F2 were created by varying the intensity weightings of the sine wave pairs. The patterns of vowel identification decisions were similar with either the actual or virtual F2. The results are interpreted as evidence that the auditory system performs spectral integration across spectral components and can extract formant frequency information which in fact is not present in vowel spectrum.

Keywords: vowel perception, auditory spectral integration, center of gravity, virtual formant.

1. INTRODUCTION

The concept of formant averaging or spectral integration in speech perception research has its roots in the early experiments by Delattre *et al.* [5]. This work showed that the quality of synthetic two-formant back vowels (whose first two formants are close in frequency) can be approximated by a single intermediate formant located between the two.

This effect was further explored by Chistovich and colleagues (e.g., [2][3]) within the framework of the spectral COG hypothesis. Accordingly, when two formants differing in their relative amplitudes are integrated, the frequency of the perceived formant (F*) is closer to that of the stronger formant. This corresponds to a shift in vowel identification. The predictable shift in the frequency of the single formant in a matching task (in which a one-formant vowel was matched to a two-formant vowel) occurred when the two close formants fell within a bandwidth of about 3.5 bark. The effect disappeared when the frequency separation was larger than 3.5 bark. In such cases, F* was matched either to F1 or F2 or listeners showed chance performance.

It was postulated that the changes to the relative amplitude ratios between the two formants changed their combined spectral center of gravity and it was to this spectral COG that the frequency of the single formant was being matched. The COG effect was interpreted as indicating that the auditory system performs auditory spectral summation at a more central processing level.

Over the years, the work on auditory spectral integration has utilized limited stimulus sets and has provided mixed results [1]. In addition, little is known about the psychoacoustic characteristics of the process which, as suggested by experimental evidence, may not be restricted to phonetic processing of vowel spectra of simplified structure. Rather, the COG effect may reflect a general property of the auditory system and therefore may be equally manifested in processing of speech- and non-speech sounds [8].

The present experiment examines the potential role of the spectral integration in making phonetic vowel quality decisions. The focus is not on finding the limits of the integration bandwidth which was the primary research objective in the past. Rather, the study examines the extent to which spectral integration in static vowel-like sounds is uniform within and slightly beyond the putative 3.5 bark range. In particular, we examine the perceptual salience of "virtual formants" produced by modifying the spectral COG of two pairs of sine waves.

2. METHOD

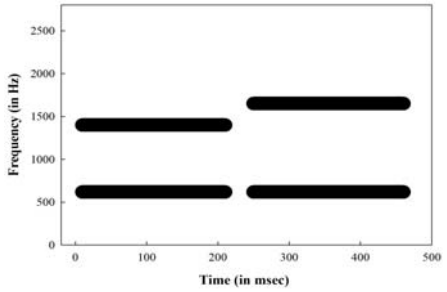
2.1. Stimuli

Two vowel pairs were selected, /ʌ/-/æ/ and /ɪ/-/i/. Two types of stimuli were created for each pair.

2.1.1. Two-formant synthetic series

Two-formant synthetic series were created, one for each vowel pair, using HLSYN [4] (.kld option, parallel synthesis). In each series only the frequency of F2 was modified, and F0 was kept constant at 120 Hz. For /ʌ/-/æ/ series, F1 was held constant at 620 Hz while F2 was increased in ten 28-Hz steps from 1400 Hz (the /ʌ/ endpoint) to 1650 Hz (the /æ/ endpoint). For /ɪ/-/i/ series, F1 was held constant at 400 Hz while F2 was increased in ten 39-Hz steps from 1800 Hz (the /ɪ/ endpoint) to 2150 Hz (the /i/ endpoint). All vowels were 210 ms in duration and were on-ramped and off-ramped over 20 ms. The schematic endpoints of two-formant synthetic /ʌ/-/æ/ series are shown in Fig. 1.

Figure 1: Endpoints of two-formant /ʌ/-/æ/ series.



2.1.2. Virtual F2 vowel series

Virtual versions of these two-formant tokens were then created for each vowel pair. In these tokens, no actual F2 appeared. Instead, a "virtual F2" (i.e., the percept of a formant) was produced by inserting two pairs of sine waves, the frequencies of which were multiples of the F0 (120 Hz). For /ʌ/-/æ/ series, the frequencies of the individual sine waves in the pair were varied as shown in Table 1a. The amplitude of each individual sine wave within a pair was identical. A pair of sine waves instead of a single sine wave was used to achieve a more natural-sounding synthetic vowel.

The two pairs of sine waves whose frequencies were both below and above the "perceptual target frequency" (or a "virtual F2") were then added together and the overall intensity of the composite was adjusted to match the intensity of the original F2. An estimate of the perceived frequency of these two pairs of sine waves (resulting from auditory spectral integration, i.e., the spectral COG) was computed as an intensity-weighted average of their center frequencies (the intensities of these sinusoidal components being proportional to the square of their amplitudes). The composite was then inserted into the base token that consisted of a single formant (F1) at 620 Hz (see Fig. 2). The resulting stimulus tokens were matched in overall intensity (within .2 dB) to the original two-formant vowel tokens.

Figure 2: Virtual F2 of /ʌ/-/æ/ series.

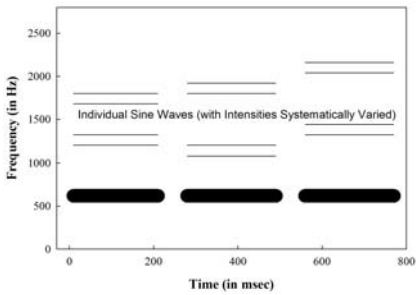


Table 1a: Frequencies of individual sine waves in creation of virtual F2 for /ʌ/-/æ/ series. ERB calculated using [6]; bark calculated using [7].

Set	ERB/Bark difference	Lower pair of sine waves (Hz)		Higher pair of sine waves (Hz)	
1	3.31/2.65	1200	1320	1680	1800
2	4.69/3.75	1080	1200	1800	1920
3	3.09/2.48	1320	1440	1800	1920
4	3.60/2.88	1320	1440	1920	2040
5	4.07/3.27	1320	1440	2040	2160

Table 1b: Relative amplitudes of lower and higher pair of sine waves of endpoint stimuli for /ʌ/-/æ/ series; intermediate steps determined using linear interpolation.

Set	/ʌ/ endpoint		/æ/ endpoint	
	Lower	*Higher*	*Lower*	*Higher*
1	0.900	0.100	0.100	0.900
2	0.850	0.150	0.150	0.850
3	0.900	0.100	0.100	0.900
4	0.850	0.150	0.150	0.850
5	0.800	0.200	0.200	0.800

Five different sets of 10-step virtual F2 stimuli were created for /ʌ/-/æ/ series. These sets varied in terms of the frequencies of the sine wave pairs and the frequency separation between them. The frequencies of the sine waves are shown in Table 1a. The relative amplitudes of the lower and higher sine wave pair in the composite were varied in equal steps (see Table 1b) to

create the virtual F2 frequency for each step of each stimulus set.

Virtual /ɪ/-/i/ series were created in a similar way. The frequencies of the individual sine waves in the pair were varied as shown in Table 2a. The composite was inserted into the base token with F1 at 400 Hz. Four different sets of 10-step virtual F2 stimuli were created for /ɪ/-/i/ series. The relative amplitudes of lower and higher pairs of sine waves were varied as shown in Table 2b.

Table 2a: Frequencies of sine waves in creation of virtual stimuli for /ɪ/-/i/ series.

Set	ERB/Bark difference	Lower pair of sine waves (Hz)		Higher pair of sine waves (Hz)	
1	3.58/2.88	1560	1680	2280	2400
2	4.77/3.83	1440	1560	2400	2560
3	4.37/3.30	1560	1680	2400	2560
4	4.49/3.61	1560	1680	2560	2680

Table 2b: Relative amplitudes of lower and higher pairs of sine waves of endpoint stimuli for /ɪ/-/i/ series; intermediate steps determined using linear interpolation.

Set	/ɪ/ endpoint		/i/ endpoint	
	Lower	*Higher*	*Lower*	*Higher*
1	0.900	0.100	0.100	0.900
2	0.850	0.150	0.150	0.850
3	0.925	0.075	0.075	0.925
4	0.850	0.150	0.150	0.850

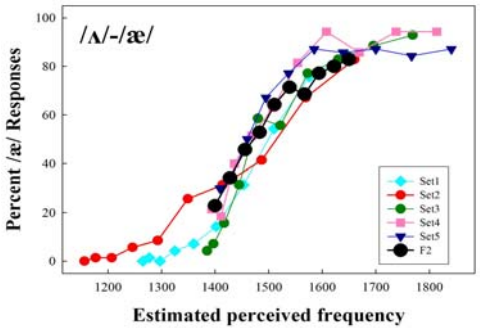
2.2. Listeners and procedure

Seven listeners aged 20-34 years who were native speakers of American English participated in the experiment. The stimuli were presented diotically via Sennheiser HD600 headphones at 70 dB HL to a listener seated in a sound-attenuating booth. A single-interval 2AFC identification task was used with the response choices either /ʌ/ and /æ/ (for the /ʌ/-/æ/ pair) or /ɪ/ and /i/ (for the /ɪ/-/i/ pair) displayed on the computer monitor. A short practice was given prior to the experiment. The presentation of stimuli was blocked only by vowel pair. For each vowel pair, all stimuli from the different stimulus sets including two-formant synthetic tokens and all virtual sets were presented randomly in one block. Listeners completed the task in two 1-hr sessions.

3. RESULTS

Listeners' mean responses to the /ʌ/-/æ/ series for all virtual sets and for two-formant stimuli are shown in Fig. 3. For both stimulus types, the lowest F2 frequency gave rise mostly to the identifications as /ʌ/ and the highest F2 to the identifications as /æ/. The consistency of responses across all sets is noteworthy, especially in terms of the steepness of the identification functions for F2 range between 1400 -1650 Hz.

Figure 3: Responses to /ʌ/-/æ/ series.



The results were analyzed in terms of category boundary differences among the different sets of stimuli. First, the /ʌ/-/æ/ category boundary (representing the 50%-crossover point) for each set was calculated for each listener using PROBIT analysis. These boundaries were then analyzed using a within-subject analysis of variance (ANOVA). Results showed no significant differences in the category boundaries among the sets. Additional Bonferroni-adjusted pairwise t-tests comparing the mean category boundary of each virtual set with the two-formant set showed no significant differences. Interpreting these results, it is clear that listeners identified the tokens as instances of either /ʌ/ or /æ/ equally well and in a similar way regardless of stimulus type used. This indicates that a virtual formant provides as strong a cue to vowel identification as an actual formant frequency does.

The mean responses to the second vowel pair, /ɪ/-/i/, are shown in Fig. 4. In these displays, the lowest F2 frequency corresponds to the greater number of identifications as /ɪ/ and the highest F2 to the identifications as /i/.

A repeated-measures ANOVA of the category boundary for the /ɪ/-/i/ pair revealed a significant main effect of set [$F(1.8, 9.2) = 5.27$, $p = 0.032$]. Post-hoc analyses revealed that the mean category boundary for Set 1 was significantly higher than for the remaining virtual sets. Again, Bonferroni-adjusted pairwise comparisons between each virtual set and the two-formant set showed no significant differences. Overall, the responses for the /ɪ/-/i/ pair were more variable across the sets than were

responses to the /ʌ/-/æ/ series although the general direction of the identification functions is as expected. This may be due to the overall difficulty of the task, as listeners found it more difficult to respond to the /ɪ/-/i/ pair when all stimuli were presented randomly on one block. Also, since there were only seven listeners in the experiment, we cannot exclude that this result was due to the lack of power.

Figure 4: Responses to /ɪ/-/i/ series.

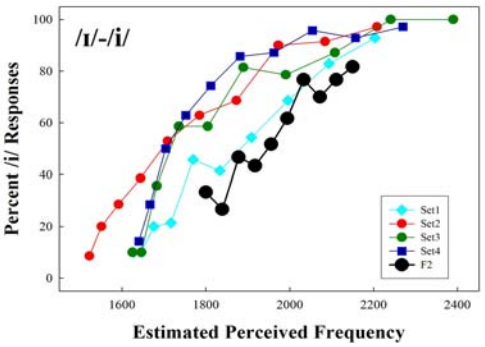


Figure 6: Responses to two-formant /ɪ/-/i/ series.

4. DISCUSSION

The present results show that listeners are able to make vowel identification decisions when the F2 is represented by a virtual formant. It needs to be underscored that the perceived frequency of the virtual formant represents the spectral COG of a set of sinusoidal components whose individual frequencies did *not* match the perceived F2. There were no marked discontinuities in any of the identification functions near the midpoints of any stimulus series, showing that listeners have no difficulty integrating the spectral information in the frequency range of the F2. There was no decline in performance with increased frequency separations between the lowest and the highest sine waves (which varied from 3.09 ERB/2.48 bark to 4.77 ERB/3.83 bark). Rather, identification functions were similar for separations smaller or greater than 3.5 bark, showing no indication of a limit of the integration bandwidth.

Patterns of listeners' responses were consistent with operation of auditory spectral integration producing the COG effect in that greater intensity of the lower pair of sine waves caused the perception of an /ʌ/ (or an /ɪ/) and greater intensity of the higher pair caused the perception of an /æ/ (or an /i/). Variation in intensity weighting across the pairs of sine waves produced variations in the perceived frequency of the virtual F2. Identification responses to the virtual formants were consistent with those for the two-formant stimuli although no actual formant occurred at this frequency in the stimulus token.

Comparing responses to the virtual stimuli and to the true two-formant stimuli we found the identification functions to be remarkably similar (in terms of both slope and category boundaries). Although the results were more variable for the /ɪ/-/i/ pair, here also the slopes and category boundaries were similar across the series for each vowel set.

In conclusion, the present experiment provides evidence that the auditory system performs spectral integration across spectral components. In turn, listeners perceive a virtual frequency corresponding to the COG of the components (such as sine waves used here) which can be determined by their intensity ratios. The results generally support earlier research findings in terms of manifestation of an auditory process known as COG effect. However, no indication of a 3.5-bark integration limit has been found at present.

ACKNOWLEDGMENT

Work supported by a research grant NIH/NIDCD R01 DC006879.

REFERENCES

[1] Assmann, P. F. 1991. The perception of back vowels: Centre of gravity hypothesis. *Quart. J. Experimental Psych.* 43A, 423-448.

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

[3] Chistovich, L., Lublinskaja, V. 1979. 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, 185-195.

[4] HLSYN. High-lever parameter speech synthesis system, v.2.2. 1997, Sensimetrics Co.

[5] Delattre, P., Liberman, A., Cooper, F., Gerstman, L. 1952. An experimental study of the acoustic determinants of vowel color. *Word* 8, 195-210.

[6] Moore, B., Glasberg, B. 1983. Suggested formulae for calculating auditory filter bandwidths and excitation patterns. *J. Acoust. Soc. Am.* 74, 750-753.

[7] Traunmüller, H. 1990. Analytical expressions for the tonotopic sensory scale. *J. Acoust. Soc. Am.* 88, 97-100.

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