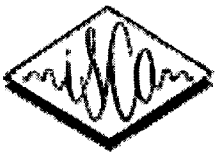
Experimental Linguistics



International Speech Communication Association

Proceedings of ISCA Tutorial and Research Workshop
on

Experimental Linguistics

28-30 August 2006, Athens, Greece.

Edited by Antonis Botinis

Antonis Botinis

University of Athens

Dynamic auditory representations and phonetic processing: The case of virtual diphthongs

Ewa Jacewicz, Robert Allen Fox and Lawrence L. Feth
Department of Speech and Hearing Science, Ohio State University, USA

Abstract

Auditory spectral integration in the perception of dynamic acoustic cues in speech was examined experimentally. The potential role of a dynamically changing center of gravity in the perception of diphthongs /ui/ (as in *we*) and /iu/ (as n *you*) was verified. Listeners identified the effective frequency changes well, showing that movement of the spectral center of gravity can provide the cues necessary for the identification of dynamic events in speech such as F2 transitions. Most models of vowel perception propose that vowels are identified on the basis of formant peaks. Our results indicate that perception of dynamic events in speech is to a large extent attributable to central auditory processes such as spectral integration.

Introduction

The concept of auditory spectral integration (ASI) refers to the improvement in detection or discrimination of complex sounds when their bandwidth exceeds a certain value (the "critical" band). Our present interest is in understanding how ASI functions in the coding of acoustic speech segments.

In speech perception, the effects of ASI have been studied primarily with reference to formants and perceived vowel quality. For example, early research demonstrated that two closely spaced formants could be matched to a single "intermediate" formant whose frequency depends on the specific relationship between the frequencies and amplitudes of the two close formants (Delattre *et al.,* 1952). The predictable shift in the matching frequency of the single formant occurs only within a larger bandwidth of about 3.5 bark (e.g., Chistovich and Lublinskaja, 1979). This center of gravity (COG) effect was interpreted as an indication of a central processing such as ASI.

A significant limitation of this early research was that the integration effects were examined only in static vowels. Auditory processing of these unnatural speech sounds is entirely focused on the frequency domain. However, human speech is inherently dynamic (in terms of both frequency and amplitude changes in time) and listeners are very sensitive to these dynamic changes. In a landmark study, Lublinskaja (1996) showed that the auditory system could attend to the dynamic spectral COG created by modifying relative formant amplitudes (but not formant frequencies) over time. The present paper assesses the efficacy of this "moving COG" in producing per-

Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics, 28-30 August 2006, Athens, Greece.

ceived dynamic frequency changes to signal glide (i.e., diphthong) differences.

Experiment: Perception of dynamic cues

This experiment examined and verified the potential role of a dynamically changing COG in the perception of the diphthongs /ui/ (as in *we*) with a rising F2 transition and /iu/ (as in *you*) with a falling F2 transition.

Stimuli: Actual F2 transitions (FT)

Nine basic F2 contours in diphthong stimuli were created using HLSYN with the .kld option. There were three tokens with a steady-state F2, three with rising F2, and three with falling F2. For all stimuli, F1 remained at 300 Hz for the entire token. F2 onset was 1800, 2000 or 2200 Hz; F2 offset was 1800, 2000 or 2200 Hz (see Figure 1). There were three different durations (50, 100 and 150 ms) for each of the nine formant patterns. F0 remained steady at 100 Hz for the entire token.

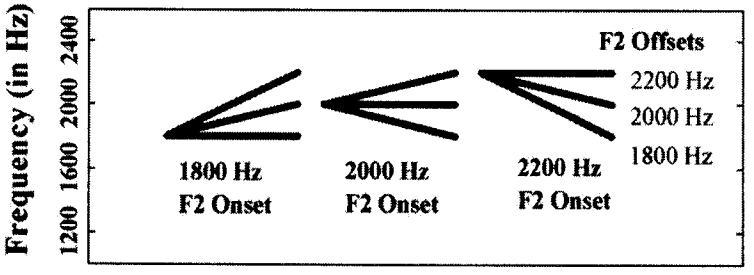


Figure 1. Schematic of nine basic F2 contours in diphthong stimuli. These basic formant patterns produced /i/, /u/, /ui/, or /iu/ percepts.

Stimuli: Virtual F2 transitions (VT or "virtual diphthongs")

A base token was created from the FT series which contained F1 only. Two "resonances" were then created. The first contained the 17th and 18th harmonics only; the second contained the 22nd and 23rd harmonics only of steady-state F2s created using the parallel branch of the synthesizer. Harmonics were isolated using sharp FIR filters. The amplitudes of these resonances were modified as appropriate to allow the COG to follow the frequency changes to the F2 in the FT series. Overall rms of these sets of resonances was matched to F2s in the FT series and inserted into the base token.

Listeners and procedure

Six native speakers of American English aged 19-28 years listened to the signals over headphones while seated in a sound-attenuating booth. A single-interval 4AFC identification task was used with the response choices */i/ /u/ /ui/* and */iu/* displayed on the computer screen. The tokens were blocked by stimulus type (FT and VT). In each block there were 270 stimuli presented randomly in each session (3 durations x 9 formant patterns x 10 repetitions).

Results

Table 1. Percentage of /i/, /u/, /ui/ and /iu/ responses to each stimulus token for the steady-state (S), rising F2 (R), and falling F2 (F) series.

	Onset (Hz)	**Offset** (Hz)	**/i/**	**/u/**	**/ui/**	**/iu/**	**/i/**	**/u/**	**/ui/**	**/iu/**
			Actual F2 Transition				**Virtual F2 Transition**			
S	1800	1800	5.6	**83.9**	1.1	9.4	6.1	**91.1**	1.7	1.1
	2000	2000	63.9	19.4	6.1	10.6	44.4	45.0	9.4	1.1
	2200	2200	**95.6**	0.0	3.9	0.6	**95.0**	2.2	2.8	0.0
R	1800	2000	13.3	7.8	75.6	3.3	3.9	15.6	80.0	0.6
	1800	2200	2.8	0.6	**95.6**	1.1	0.6	0.0	**99.4**	0.0
	2000	2200	55.0	0.0	43.3	1.7	30.0	0.6	69.4	0.0
F	2000	1800	2.2	31.7	1.1	65.0	0.6	38.9	0.0	60.6
	2200	1800	3.9	12.8	0.0	**83.3**	0.0	1.1	0.0	**98.9**
	2200	2000	33.3	9.4	2.8	54.4	21.1	18.9	0.6	59.4

In both series, listeners most often identified the 1800 Hz steady-state tokens as /u/ and the 2200 Hz steady-state tokens as /i/ (see Table 1). A paired-samples t-test showed no significant difference as a function of stimulus type. However, for the 2000 Hz tokens, there was a significantly greater number of /i/ responses to the FT tokens than to the VT tokens showing that the amplitude variation in the two resonances for VT did not produce as high a perceived F2 as in the FT condition.

As expected, the percentage of /ui/ responses for both FT and VT was highest when the decrease in F2 frequency was the largest. An ANOVA of the number of /ui/ responses with the within-subject factors series (FT and VT) and token showed a significant effect of token ($F(2,10)=12.9$, $p=0.002$, $\eta^2=0.546$), but no significant main effect of series nor a significant series by token interaction.

The percentage of /iu/ responses for both FT and VT was highest when the increase in F2 frequency was the largest. An ANOVA of the number of /iu/ responses with the within-subject factors series (FT and VT) and token showed a significant main effect of token ($F(2,10)=7.99$, $p=0.008$, $\eta^2=0.615$). The number of /iu/ responses was significantly higher for the 2200-1800 Hz token, while the remaining two were not different.

Discussion and conclusions

The results show 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 did provide the cues necessary for the identification of F2 transitions comparably with the actual formant transitions.

For both types of signals (i.e., FT and VT), the highest proportion of expected responses was obtained for the greatest frequency differences between the diphthongal onsets and offsets, which produced the clearest percepts of either /ui/ (the 1800-2200 Hz token) or /iu/ (the 2200-1800 Hz token). The proportion of the expected responses to spectral changes decreased with each smaller frequency separation between onsets and offsets, and the signals were identified as stationary vowels /i/ or /u/ when there was no frequency change.

Most models of vowel perception propose that vowels are identified on the basis of formant peaks. This approach will not work with the signals utilized here. Our approach to modelling is to examine dynamic auditory excitation patterns thought to result from the acoustic signals presented to the listener. Perception of dynamic events in speech is to a large extent attributable to central auditory processes such as auditory spectral integration explored here. Moreover, the perception of formants in vowels is almost certainly a result of the spectral integration of the energy of their harmonics.

Acknowledgements

Work supported by NIH R01DC00679-01A1 (L.L. Feth, PI). We thank Marc Smith for help with data collection.

References

Chistovich, L. A. and Lublinskaja, V. 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.

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

Lublinskaja, V. V. 1996. The 'center of gravity' effect in dynamics. In Ainsworth, W. and Greenberg, S. (eds.), Proc. of the Workshop on the Auditory Basis of Speech production, 102-105, ESCA.