ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 5aSCb: Production and Perception II: The Speech Segment (Poster Session)

5aSCb35. Perception of vowel-inherent spectral change
Kathleen Chiddenton and Michael Kiefte*

*Corresponding author's address: School of Human Communication Disorders, Dalhousie University, 1256 Barrington St., Halifax, B3J 1Y6, Nova Scotia, Canada, mkiefte@dal.ca

One family of theories regarding vowel perception suggests that onset and offset formant-frequencies are important for identification but that the shape of the transitions themselves is not perceptually important [e.g., Morrison and Nearey (2007). J. Acoust. Soc. Am. 122, EL15-EL22]. The present study determined just-noticeable-differences in deviations for linear formant trajectories. Diphthong-like stimuli were manipulated by inserting a point of inflection into the otherwise linear transition. Several parameters were manipulated including vowel duration, location of the inflection point in time, direction of formant change, and fundamental frequency. Data from the first experiment indicate that listeners are largely insensitive to deviations from linearity of formant trajectory but that deviations that fell outside the range of frequencies spanned by the onset and offset were detectable. However, a second experiment in which only the first half of stimuli were presented gave different results. Results from these experiments along with several hypotheses are presented. [Work supported by SSHRC.]

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

The way in which listeners perceive vowels has been debated since the earliest investigations of speech perception and production by human listeners. With some exceptions, vocal-tract resonances, or formant frequencies, have been commonly identified as the most important characteristics of a vowel in identification (see Kiefte et al., 2012; Rosner and Pickering, 1994, for reviews). Many studies have shown that the first two or three formant frequencies ($F_{1,2,3}$) are the main spectral properties used for vowel identification. However, it remains unclear how human listeners track these formants in the ever changing environment of natural speech production.

The simple target model of vowel perception characterizes vowels as discrete points in a two- or three-dimensional space represented by the first few formant frequencies. However, many pattern-recognition studies report better identification of vowels in models that incorporate the spectral change of the vowel rather than a measurement sampled at a single time (e.g., Nearey and Assmann, 1986). Hillenbrand and Nearey (1999) reported a study in which a pattern classifier was significantly more accurate when trained on two points drawn from the formant pattern of the vowel rather than a single point.

In addition, while it has long been established that diphthongs are characterized by changing formant frequencies over time, there is a growing volume of research that shows traditional monophthongs in English may be diphthongized in the sense that they too show vowel inherent spectral changes (VISC; Nearey and Assmann, 1986; Hillenbrand et al., 1995). This VISC has been shown to be essential in vowel identification and discrimination.

Three hypotheses of how VISC is detected by listeners have been proposed (Nearey and Assmann, 1986; Gottfried et al., 1993; Morrison and Nearey, 2007). All three are in agreement that the initial formant frequency is important in identifying vowels but they differ in what other information is necessary. The onset-offset hypothesis suggests that formant frequencies at the beginning and the end of the vowel are the important spectral properties in perception. The slope hypothesis claims that perceptual cues are based on the rate of change over time for the formant frequencies irrespective of offset frequencies. Similar to the slope hypothesis, the direction hypothesis suggests that the direction of formant movement is the relevant piece of information irrespective of vowel duration. Several studies have been conducted to evaluate which of these hypotheses can better represent how listeners perceive VISC. Results in favor of the onset-offset hypothesis were presented by Nearey and Assmann who used silent-center vowel stimuli to determine how identifiable the vowels were. With only brief onglides and offglides, high identification rates were still obtained.

Until now, it has never been conclusively shown that listeners perceptually follow formant frequencies across time. Most results have been in favor of the onset-offset hypothesis for vowel identification such that formant tracking over time is not actually required. Morrison and Nearey (2007) reported a study to determine which aspects of formant change in vowels were relevant in perception. They used three types of synthetic vowels to differentiate among the three different VISC hypotheses: Formant trajectories were either straight transitions across the duration of the vowel, elbowed with an initial steady state for the first quarter of the vowel duration followed by gliding formant transitions, or constant with no formant-frequency changes throughout the vowel duration. Similar to Nearey and Assmann (1986), their results provided support for the onset-offset hypothesis indicating the perceptual importance of the initial and final formant frequencies with little attention being given to intermediate changes—i.e., listeners identified the straight and elbowed formant trajectory as the same vowels. Further, anecdotal evidence suggested that listeners were not even able to detect differences between these stimuli suggesting that they were not at all sensitive to the formant manipulations used to test vowel identification.
Extending this research, we tested the limits of the onset-offset theory. One can assume that there is a point at which listeners must be able to detect a change in formant frequency in the middle of the vowel if it deviates from the straight trajectory sufficiently. By examining the threshold at which listeners detect differences in the stimuli in which the deviation of the elbow differs, we may gain some insight into the perceptual processes involved in formant perception and tracking. A better understanding of how listeners perceive important characteristics of vowels can also offer a better understanding of speech perception in general.

In addition, although listeners may not have been sensitive to the manipulations used in the study by Morrison and Nearey (2007), it is useful to know the smallest deviation from a straight formant transition that is detectable by listeners. This information would help to evaluate the practical limits of the theories that have been proposed for vowel perception.

**METHODS**

Fifteen normal-hearing participants were recruited to participate. All subjects were 18 or older and had normal hearing as determined by a hearing screening. Participants were seated in an IAC double-walled sound-treated booth and were fit with a pair of Beyer-Dynamic DT150 headphones. All stimuli were presented via an Edirol UE-25EX external AD/DA attached to an iMac. Presentation was controlled by PsychoPy.

The experiment was a three-alternative force-choice 1-up 3-down adaptive staircase design which converges on the .841 probability on the psychometric function (Levitt, 1971). Using a numeric key pad, participants selected either 1, 2, or 3, corresponding to the stimulus they believe sounded different from the other two. Inflection points were varied in steps of 5 Hz, either up or down. Each run was terminated after 12 reversals of which the last eight were used to determine the resultant threshold.

The default stimulus used was 100 ms long and the vowel was modeled after the diphthong /ei/. All stimuli were generated using a MATLAB implementation of the KLATT80 speech synthesizer (Kieft, 2002; Klatt, 1980). All synthesis parameters were maintained at their software defaults as described by Klatt with the exceptions described below.

In the default stimulus series, first formant (F1) onset frequency was set at 425 Hz and offset frequency was 394 Hz. Second formant (F2) transitioned from 1900 Hz at onset to 1971 Hz at offset (100 ms). For the deviant stimuli, an inflection point or bend was introduced at 25% of the vowel duration or 25 ms. The deviation of this frequency from the linearly interpolated frequency between onset and offset was varied until responses converged on a detection threshold. Fundamental frequency f0 was set at 125 Hz at onset and fell to 88 Hz at offset. However, several other series were also tested.

In addition to a deviation introduced at 25 ms, another series was generated in which the bend occurred at 50 ms. In another, the duration of the vowel was extended to 200 ms with the inflection point still at 25% of the duration or 50 ms. In a fourth series, the formant transitions for both F1 and F2 were reversed in time (offset and onset values switched) such that the stimuli sound like /i/ instead of /ei/. Finally, another series was identical to the first (100-ms duration and inflection point at 25 ms) but with the fundamental frequency decreased to 110 Hz at onset and 77 Hz at offset.

The purpose of these variations was to examine the effects of changing the slope of F2 transition (100 versus 200-ms duration) and the direction (default versus reversed). By placing the inflection point at 50% of the duration, we can also examine the effects of changing transition slope independent of duration. The final condition in which the fundamental frequency was altered was introduced to examine the possible interaction between formant and fundamental frequencies.
RESULTS

The results for each condition are represented graphically in Fig. 1 and are based on the average threshold across participants. Error bars give the standard error of the mean. Each subfigure gives the original series as a comparison (rising $F_2$, 100-ms duration, 125-Hz onset $f_0$, 25-ms inflection point) where either the blue or green line shows the comparator stimulus (straight formant transition) and the red line gives the threshold at which listeners detected a difference between the deviated formant transition and the comparator. Also in each figure, the comparator stimulus for a second condition is represented by a straight blue line [In Figs. (a) and (b) this is the same as for the comparator stimulus in the original series]. The angled brown line represents the threshold or smallest deviation necessary from the straight transition for participants to detect a stimulus change.

Figure 1 (a) compares the original series with a 50-ms inflection point. As shown, $F_2$ had to deviate from the original path to an absolute threshold of 1831 Hz which is approximately the same value for the original series. However, the difference between the inflection point and the frequency of $F_2$ at the same time in the comparator stimulus was much larger than for the original series (−104 versus −87 Hz, respectively). Figure 1 (b) shows that changing $f_0$ had virtually no effect on the threshold indicating that there was little measurable interaction between formant and fundamental frequencies.

Figure 1 (c) shows that reversing the direction of $F_2$ had little effect on the absolute threshold of the inflection point (1839 Hz). Nonetheless, the deviation of the inflection point from the straight formant transition in the comparator stimulus was much larger (−114 versus −87 Hz, respectively). This suggests that the magnitude of the deviation in $F_2$ needed to be much larger for the /i/ stimulus but that the absolute threshold of the frequency deviation was approximately the same as for /ε/.

Finally, Fig. 1 (d) shows that changing the slope of the formant does little to change the absolute frequency of the threshold for the deviation. The threshold for the 200-ms condition was 1817 Hz.

DISCUSSION

The purpose of this study was to test the limits of the onset-offset theory of vowel perception. By introducing a sufficiently large deviation, we were able to show that listeners can perceive differences in formant frequency over time. As shown in Fig. 1, the absolute threshold reached for each of the conditions was similar despite several differences between them. The $F_2$ in these stimulus conditions varied in slope, direction, onset/offset frequency. Although relative thresholds (differences between threshold frequency at the inflection point and the frequency at the linear formant transition) were very different, thresholds in absolute frequency were largely the same.

In order to understand these results, it is necessary to examine variables that are held constant across these conditions. The primary similarity across stimulus series is range of $F_2$ frequencies across the duration of the stimulus irrespective of other parameters. It is possible that this is the primary factor that influences the magnitude of the deviation necessary for listeners to perceive a change. If this is true, it implies that listeners’ ability to detect these deviations can only be determined after the stimulus is complete as listeners must hear the offset of the stimuli before the range of $F_2$ frequencies can be determined.

In order to test this theory, continued research will be done on stimuli similar to these, but which are truncated such that the range of $F_2$ frequency is reduced while preserving the slope and onset frequencies of the formants. We expect the absolute threshold to be affected if this is in fact the deciding factor.
FIGURE 1: Results from experiment. In each subfigure, the blue line (—) illustrates the formant transition for the reference $F_2$ (linear transition) and the brown line (—) gives the threshold. As a reference, in each figure, the red lines (—) give the threshold for the original series with the inflection point at 25 ms. In Figs. (c) and (d), the green line (—) gives the reference stimulus for the original series. In Figs. (a) and (b) this is the same as for both series illustrated.
ACKNOWLEDGMENTS

This work was supported by a fellowship from the Social Sciences and Humanities Council of Canada to the first author.

REFERENCES


