# Perception of vowels with missing formant peaks

Filip Nenadić,[1,a] Pamela Coulter,[2] Terrance M. Nearey,[1] and Michael Kiefte[2]

[1]*Deparment of Linguistics, University of Alberta, Edmonton, Canada*

[2]*School of Communication Sciences and Disorders, Dalhousie University, Halifax, Canada*

(Dated: 5 February 2022)

This is an Accepted Manuscript of an article published in
The Journal of Acoustical Society of America on October 8, 2020,
available online: https://asa.scitation.org/doi/10.1121/10.0002110.

Although the first two or three formant frequencies are considered essential cues for vowel identification, certain limitations of this approach have been noted. Alternative explanations have suggested listeners rely on other aspects of the gross spectral shape. A study conducted by Ito *et al.* [J ACOUST SOC AM (110), 2001] offered strong support for the latter, as attenuation of individual formant peaks left vowel identification largely unaffected. In the present study, these experiments are replicated in two dialects of English. Although the results were similar, quantitative analyses showed that when a formant is suppressed, participant response entropy increases due to increased listener uncertainty. In a subsequent experiment, using synthesized vowels with changing formant frequencies, suppressing individual formant peaks led to reliable changes in identification of certain vowels but not in others. These findings indicate that listeners can identify vowels with missing formant peaks. However, such formant-peak suppression may lead to decreased certainty in identification of steady-state vowels or even stable changes in vowel identification in certain dynamically specified vowels.

[a]nenadic@ualberta.ca

## I. INTRODUCTION

Peterson and Barney (1952) described the first two or three formant frequencies as essential cues when investigating vowel identification. The "formant hypothesis", also called the "target model", has been dominant ever since. This approach is supported by many studies (e.g., Klatt, 1982) or at least is always mentioned in studies exploring the role of formants and their characteristics in vowel perception (e.g., Kiefte *et al.*, 2010). Kiefte *et al.* (2013) provide an overview of arguments in favor of the notion that information near the high-intensity formant peaks should be the most robust and informative part of the signal and the formant hypothesis is also discussed extensively in reviews (e.g., Molis, 2005; Rosner and Pickering, 1994).

However, certain issues and limitations of the formant hypothesis have been noted (Bladon, 1982, 1983; Molis, 2005): (1) relying only on formant peaks represents a significant reduction of the signal, (2) determining formant frequencies is not always an easy or straightforward task, and (3) formant frequencies alone cannot fully account for certain empirical findings (see e.g., Fox *et al.*, 2010; Hillenbrand *et al.*, 2006). Another example of such an empirical finding is noted by Ito *et al.* (2001) where change in relative amplitude of adjacent formants — as in the center of gravity effect (Chistovich and Lublinskaya, 1979) — can affect vowel perception even if formant values are held constant. Additionally, engineering solutions for automatic speech recognition do not rely on extracting formant values as parameters (Yu and Deng, 2014). These and similar arguments support an alternative explanation in which not only formant peaks, but the overall spectral shape, acts as a cue to vowel identity. This "whole-spectrum hypothesis" might then provide a better fit to the data gathered from listeners (Bladon and Lindblom, 1981; Hillenbrand and Houde, 2003; Zahorian and Jagharghi, 1993).

Perhaps the strongest evidence against formant peaks as the only relevant cues for vowel identification comes from experiments conducted by Ito *et al.* (2001). In their first experiment, the authors synthesized a continuum of vowels varying by $F_1$ and $F_2$ values which were used as controls, as well as suppressed-formant variants in which either $F_1$ or $F_2$ peaks were flattened with as much of the remaining spectral shape as possible retained. Stimuli were presented in successive per-condition blocks to four listeners and responses showed that suppressing formant peaks did not radically change vowel identification. In the second and third experiment, Ito *et al.* also show that changing the amplitude ratios of $F_1$ relative to higher formants affects vowel perception. These results indicate that loss of formant frequency information can be compensated for by using information extracted from the gross spectral shape. Additionally, it seems that changes in relative formant amplitude (e.g., spectral tilt) can affect vowel identification even if formant frequencies are not manipulated.

Following these findings, Kiefte and Kluender (2005) compared relative contributions of the second formant frequency and spectral tilt in an experiment that finely manipulated them in synthesized /i/ to /u/ continua. Second-formant variation proved to be a significant cue for determining which vowel was heard, but so did spectral tilt, albeit with a smaller effect size (expressed as $D^2$). Both the results of Ito *et al.* (2001) and Kiefte and Kluender (2005) may result from effects of simultaneous masking as acknowledged by Kiefte *et al.* (2010). However, Kiefte and Kluender (2005) found that very different results are obtained when using /ai/ and /au/ stimuli in which formant-frequency parameters change — even by very small amounts — throughout the duration of the stimulus unlike the stimuli used by Ito *et al.* (2001) in which the synthesized formant values were kept constant. In these circumstances, spectral tilt did not have a significant effect on vowel identification, prompting the conclusion that spectral tilt may be informative only for vowels that have unchanging spectral characteristics. English has a number of diphthongs wherein formant frequencies change substantially as the vowel unfolds (Hillenbrand *et al.*, 1995; Hillenbrand and Nearey, 1999). Moreover, recordings of many English vowels regarded as monophthongs also show changing formant patterns that are important for their perception.

Besides using vowels with steady formant peaks, Ito *et al.* (2001) made other design decisions that could have affected the outcome of their study. Only four participants were tested and substantial individual differences can be seen in their responses. All participants heard each stimulus a very large number of times; that is, they had prolonged exposure to the stimuli. The study was conducted in Japanese which has only five vowel categories, so less robust acoustic cues (e.g., spectral tilt) might suffice to distinguish vowels in this sparse choice set. Finally, the three types of stimuli (original, $F_1$-suppressed, and $F_2$-suppressed) were presented in separate blocks, which may have allowed listeners to more easily adapt to formant peak attenuation within each condition and focus their attention on other cues.

The above considerations raise questions as to the importance of gross spectral shape cues when identifying vowels in a more ecologically valid setting. As both Molis (2005) and Kiefte *et al.* (2013) note, the formant hypothesis and the whole-spectrum hypothesis are not mutually exclusive — the whole-spectrum approach also necessarily includes information about the location of local formant peaks. It is clear that formant frequencies seem to be sufficient for reliable vowel identification in certain contexts, such as in pattern-playback speech (Delattre *et al.*, 1952) or when only three harmonics corresponding to formant peaks are preserved (Kakusho *et al.*, 1971; Kiefte *et al.*, 2010). This, however, does not mean that they are necessary in more naturalistic speech, nor that other spectral characteristics cannot be informative as well given the right circumstances (see, e.g., Chistovich and Lublinskaya, 1979; Ito *et al.*, 2001; Kiefte and Kluender, 2008). The question is rather what are the circumstances in which (1) formant-frequency information can be distorted without impeding vowel identification, and (2) other spectral characteristics (most notably amplitude information, e.g., spectral tilt), are utilized by listeners.[1]

Although Kiefte and Kluender (2005) investigated the same effects as Ito *et al.* (2001), they did not strictly replicate the original experiment. The present study more closely follows the methods of Experiment 1 conducted by Ito *et al.* Our Experiments 1 and 2 involve a larger number of listeners from two dialects of English, both of which have larger vowel inventories than Japanese. This may limit the listeners' ability to benefit from broadly tuned spectral characteristics in distinguishing phonetically similar vowels. Our last two experiments explore more ecologically valid situations: Experiment 3 investigates how stimulus blocking affects which cues listeners rely on, as in this experiment stimuli with a suppressed formant are presented together with original full-formant stimuli in randomized order, simulating situations where formant peaks are possibly masked or attenuated by the listening environment. Finally, in Experiment 4, we synthesize vowels with changes in their formant values across time to test how loss of formant information affects perception if that formant is also variable in time. Our expectations are that formant-peak manipulation should have more detrimental effects in our experiments than those recorded by Ito *et al.*[2]

## II. EXPERIMENT 1

### A. Method

Fifteen native speakers of Eastern Canadian English (22 – 32 years; M = 25.7; SD = 2.92; 67% females) were recruited from the Dalhousie University School of Communication Sciences and Disorders in Halifax, Canada. Participants received no compensation for taking part in the study. All participants completed an undergraduate university phonetics course and thus had some knowledge of English vowel phonology as well as the ability to respond using IPA vowel symbols. None of the participants reported any hearing impairment, and their measured hearing thresholds were normal.

Stimuli were vowels synthesized in a manner similar to that of a cascade-type Klatt synthesizer (Klatt, 1980) and following the procedure described in Ito *et al.* (2001). Fundamental frequency, $F_0$, was set at 125 Hz and the first two formants of the vowels were varied systematically in 125 Hz increments, ranging from 250 to 1250 Hz for $F_1$ and from 750 to 2250 Hz for $F_2$. Higher formants were set to 2500, 3500, and 4500 Hz and the remaining synthesis parameters are given in Table 1 of Ito *et al.* (pp. 1142). Vowels which had $F_1$ and $F_2$ within 200 Hz of each other were excluded as unnatural, so the final number of synthesized vowels was 96. These control vowels were then modified to suppress either the F1 or F2 peak, while retaining as much of the remaining spectral shape as possible. After the stimuli were generated via cascade synthesis at a 10-kHz sampling rate, 80 samples (8 ms) corresponding to one pitch period were extracted from a window 100 ms following the onset. This frame was analyzed via Fourier transform such that each component in the spectral domain gave the amplitude and phase of each harmonic. To excise a formant peak, two harmonics were found — one on either side of the target formant peak — such that a straight line between them in dB/ERB (Glasberg and Moore, 1990) would fall below all intermediate harmonics in amplitude as well as the two harmonics immediately outside that range on either side. The amplitudes and phases of the intervening harmonics were then linearly interpolated between these two harmonics in dB/ERB. Experimental stimuli were then resynthesized from the modified spectra via inverse Fourier transform. The resulting 80-sample segment was then repeated to produce a 400-ms stimulus. The onset and offset of the stimulus was weighted by a 4-ms half-Hamming window. Sample spectra of a single vowel in each of the three conditions are given in Figure 1 and the stimuli are available in our supplementary material.

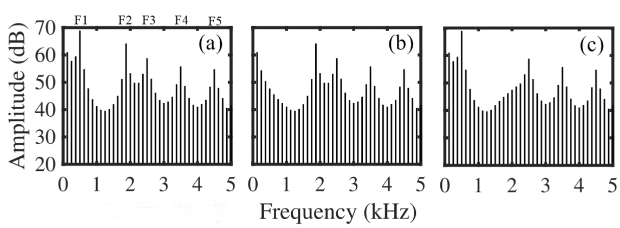


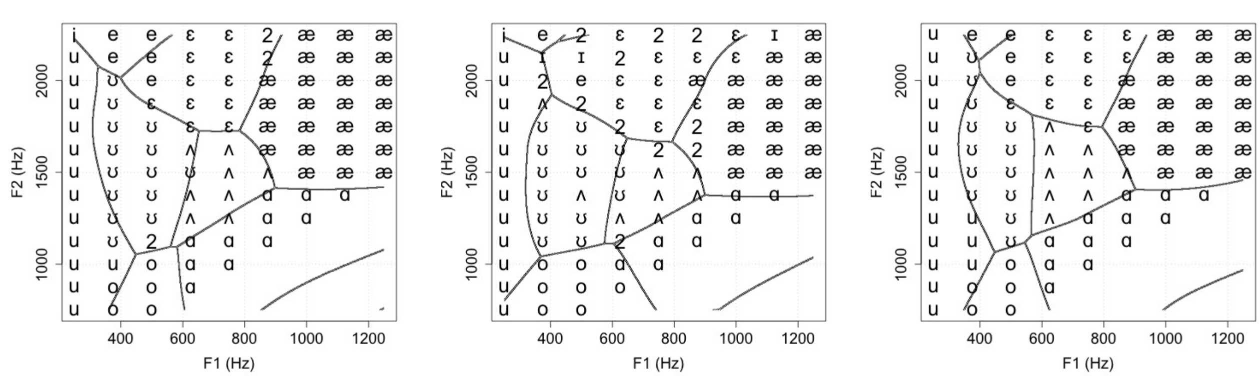

FIG. 1. Sample vowel in its (a) original form, (b) with $F_1$ suppressed, and (c) with $F_2$ suppressed.

The experiment was conducted in a sound-attenuated booth and began with participants' hearing screening. Stimuli were presented using MATLAB, a digital signal processor Edirol UA-25EX, and circumaural headphones (Beyerdynamic DT 290) at 75 dB SPL. In response, participants used a DX1 system by ErgoDex to input their selection from a choice of 10 buttons, each programmed for one of the vowel choices. The input system has an image of the English vowel quadrilateral with the buttons placed at the conventional vowel positions and marked with both an IPA symbol and an orthographic representation of an /hVd/ word.

A practice session consisting of 20 stimuli with both formants preserved was first completed to familiarize participants with the task. Next, three blocks (original, $F_1$-suppressed, $F_2$-suppressed) were presented in random order. Stimuli were ordered randomly within each block. Participants only heard each stimulus once to avoid both extensive familiarization and fatigue; the larger number of responses per participant used by Ito *et al.* (2001) was replaced by an increase in participant sample size.

### B. Results

Contour plots of participants' synthesized vowel classifications are presented in Figure 2 (see supplementary materials for two additional sets of differently generated contour plots and a more detailed description of how each of these sets of plots were generated). The figures label the empirical modal response for every stimulus ($F_1$-$F_2$ combination) and the numeral 2 if two responses tied (and more rarely 3 when three responses tied). The original synthesized vowels show plurality response regions in the $F_1$-$F_2$ plane in roughly the expected places, with the exception of /i/ and /ɪ/ which received very few responses. Responses for $F_1$-suppressed and $F_2$-suppressed vowels show broadly similar patterns to those observed in original stimuli.




(a) (b) (c)

FIG. 2. Phoneme boundaries and modal responses for the (a) original synthesized vowels, (b) vowels with $F_1$ suppressed, and (c) vowels with $F_2$ suppressed in Experiment 1. The number 2 is used when two responses tied.

Importantly, we find distinctions between vowel responses are largely preserved along the frequency axis of the suppressed formant. In Figure 2 (b), which shows the $F_1$-suppressed condition, we see differences between /u/, /ʊ/, and /ʌ/, all of which have the same range of $F_2$ values around 1500 Hz, and are apparently still distinguished primarily by the suppressed peak $F_1$. A similar distinction is made between /ɛ/ and /æ/, which share $F_2$ values, but remain differentiated by the suppressed $F_1$ value. In Figure 2 (c) we see that vowels /ɛ/, /ʌ/, and /ɑ/ have similar $F_1$ values (around 600 to 850 Hz), but different $F_2$ values, even though this formant is suppressed, and the same can be observed for vowels /ʊ/ and /o/. In other words, the overall response patterns for vowels with a suppressed formant qualitatively resemble that of the original synthesized vowels.

However, we also wanted to quantify the variability present in listener responses. We used Shannon (informational) entropy (Shannon, 1948) calculated over relative frequencies of each phoneme response to a given synthesized vowel. This is calculated as $H$ (in nats) as shown in Equation 1, where a synthesized vowel $v$ has $n = 10$ different potential responses (i.e., the ten English vowels) with each being chosen as the response with a probability of $p(v_i)$. Higher Shannon entropy values indicate more disperse, varying responses.

$$H(v) = -\sum_{i=1}^{n} p(v_i) log_2 p(v_i), \tag{1}$$

We then analyzed these data by treating the Shannon entropy of each stimulus as a case in three repeated conditions (original, $F_1$-suppressed, and $F_2$-suppressed), effectively calculating a by-stimulus repeated measures ANOVA. There were significant differences in participant response entropy across conditions ($F(2, 190) = 42.06$, $p < .001$). Pairwise comparisons with Bonferroni correction showed that $F_1$-suppressed vowels have higher response entropy values than the original ($t(190) = -8.75$, $p < .001$) and $F_2$-suppressed condition ($t(190) = 6.76$, $p < .001$), indicating reduced participant certainty in vowel classification. However, the differences between the original and the $F_2$-suppressed condition were not significant ($t(190) = -1.99$, $p = .15$).

We further analyzed the responses using the package *mlogit* (Croissant, 2013) in the statistical platform *R* (R Core Team, 2017) to create multinomial logit models (see, e.g., Maddox *et al.*, 2002; Nearey, 1990, 1997, for analyses of multinomial data). The (random slope and intercept) models included the standardized $F_1$ and $F_2$ values, the condition (original, $F_1$-suppressed, $F_2$-suppressed), and the interaction between the condition and the frequency values as predictors. We were primarily interested in the effects of $F_1$ variation in the $F_1$-suppressed condition, and the effects of $F_2$ variation in the $F_2$-suppressed condition.

Figure 3 presents the effects $F_1$ value has on vowel identification. More positive coefficients indicate the response is favored by higher $F_1$ values and more negative coefficients mean the response is favored more by lower $F_1$ values. The original condition, indicated by circles connected by a solid line, varies in an expected manner. For example, low $F_1$ is indicated for /u/ and /i/, while higher $F_1$ values are noted in the cases of /ɑ/ and /æ/. The other two lines represent deviation interactions from the baseline original condition. Therefore, to obtain the total effect of $F_1$ variation in one of the two suppressed conditions, its value at each vowel is added to that of the original (solid line) condition.

The triangles connected by a dotted line represent the interaction term of vowel and $F_1$ value for the $F_1$-suppressed condition. The overall effect of $F_1$ in this condition is then the sum of the original and suppressed $F_1$ lines at each vowel. We see that suppressed $F_1$ line is roughly an attenuated mirror image of the original, indicating that the perceptual effects of $F_1$ variation are substantially weakened when energy is suppressed at the $F_1$ peak. The squares connected by a dotted line indicate the effects of $F_2$ suppression. The coefficient values are always nearer to zero than for suppressed $F_1$. This shows that the effect of $F_2$ suppression on $F_1$-related vowel contrasts is smaller.

In Figure 4, which shows the coefficients for $F_2$, circles connected by a solid line again show the original condition. Not surprisingly, more negative $F_2$ coefficients are noted for back vowels and more positive $F_2$ coefficients for front vowels. The effects of formant suppression are generally quite modest and surprisingly parallel. They tend to slightly oppose the trends in the solid line (with the notable exception of /i/ where the F2 suppression actually enhances the original effect quite noticeably). The general trend indicates the effects of $F_2$ variation is

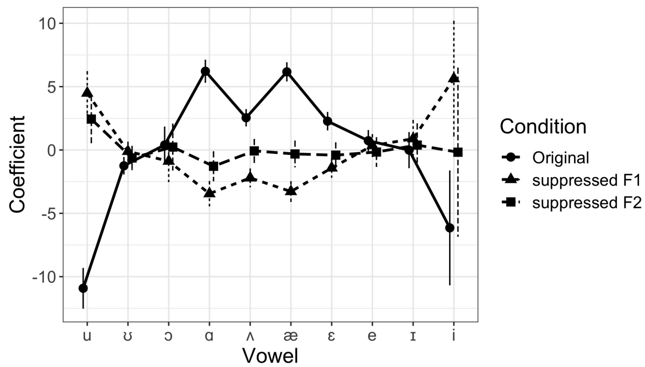



FIG. 3. Multinomial logit model coefficients per condition for $F_1$ in Experiment 1. Vertical lines indicate one standard error.

214 weaker overall in both suppressed conditions. Moreover, there is remarkably little difference in vowel identification effects with $F_2$ variation when $F_1$ (a lower formant) is suppressed in comparison to the suppression of $F_2$.

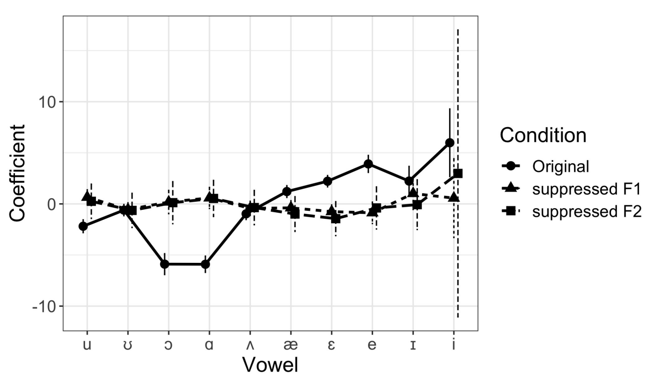



FIG. 4. Multinomial logit model coefficients per condition for $F_2$ in Experiment 1. Vertical lines indicate one standard error.

### C. Discussion

The results of Experiment 1 show that response patterns to $F_1$- and $F_2$-suppressed vowels are similar to responses to the original full-formant vowels. Moreover, distinct classifications of vowels are observed along the suppressed formant axis even when the frequency of the non-suppressed formant is nearly constant; that is, the classification changes even when it was the suppressed formant that changed frequency. These patterns indicate that the information lost by suppressing a formant peak can largely be recovered or replaced by some other source, supporting the hypothesis that listeners effectively use other cues from the overall spectral shape instead.

However, suppressing a formant does have consequences on vowel perception, as can be seen by looking at the distribution of participant responses. Participants agree less how a certain vowel should be classified when the first formant is suppressed. We take this reduction in participant agreement as an indicator of uncertainty or loss of information. Examining how participant responses vary as $F_1$ and $F_2$ change further supports this notion. We see expected response patterns in the control condition, as $F_1$ variation distinguishes between high and low vowels, and $F_2$ variation distinguishes between front and back vowels. When $F_1$ is suppressed, $F_1$ variation has a smaller effect on vowel identification in comparison to the original condition. Suppressing $F_2$ has little effect on participant responses.

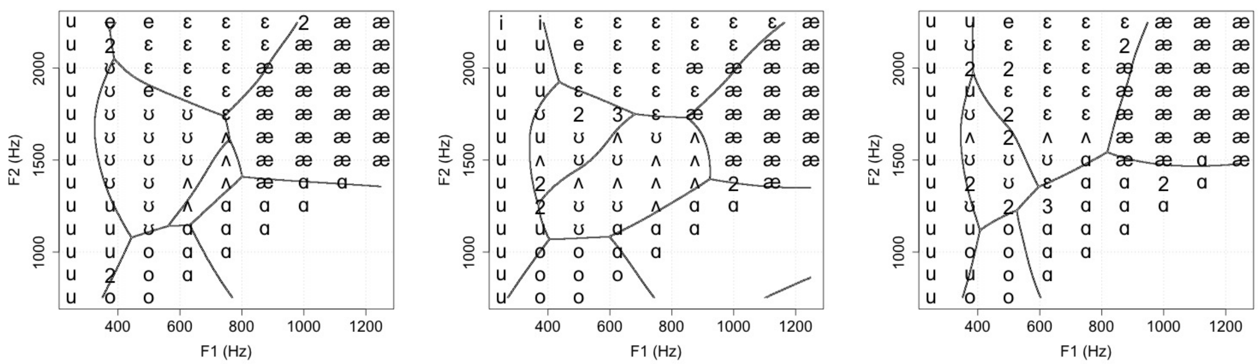
# III. EXPERIMENT 2

## A. Method

The method of the second experiment was the same as in Experiment 1, except for the following changes: 13 native speakers of Western Canadian English (18 – 27 years; M = 21.16; SD = 2.90; 2 males, 11 females) were recruited from the University of Alberta in Edmonton, Canada. These participants also completed a university phonetics course enabling them to respond using IPA vowel symbols. The stimuli were presented using a computer workstation equipped with Realtek High Definition Audio (integrated into an OptiPlex320 motherboard) over MB Quart QP 805 DEMO headphones. An image of the English language vowel quadrilateral was presented on a computer monitor, and the participants made their selection by clicking on a button that marked each vowel with an IPA symbol and an orthographic representation of an /hVd/ word. Finally, the three separate blocks of stimuli were always presented in the same order (original, $F_1$-suppressed, $F_2$-suppressed), emulating the procedure in Ito *et al.* (2001).

## B. Results

Figure 5 shows results similar to those recorded in Experiment 1. Stimuli are rarely classified as /i/ and /ɪ/. Importantly, we again note that different vowel responses are reliably given along the suppressed $F_1$ peak (e.g., /æ/, /ʌ/, and /u/ in Figure 5b), and suppressed $F_2$ peak (e.g., /æ/ and /ɑ/ in Figure 5c), much as in Experiment 1.




(a) (b) (c)

FIG. 5. Phoneme boundaries and modal responses for the (a) original synthesized vowels, (b) vowels with $F_1$ suppressed, and (c) vowels with $F_2$ suppressed in Experiment 2. The number 2 is used when two responses tied, and more rarely 3 when 3 responses tied.

However, Shannon entropy values were again different in the three conditions ($F(2, 190) = 31.16$, $p < .001$). Pairwise comparisons with Bonferroni correction indicate that the entropy of responses in the original condition is lower than in both the $F_1$-suppressed ($t(190) = -7.73$, $p < .001$) and $F_2$-suppressed condition ($t(190) = -5.26$, $p < .001$). Responses to $F_1$-suppressed vowels had slightly higher entropy than responses to $F_2$-suppressed vowels ($t(190) = 2.47$, $p = .04$).

Multinomial logit models were numerically unstable for the full range of vowels. Therefore, we collapsed the relatively rarely selected vowel categories /i/ and /ɪ/ into a single category. The effect of $F_1$ on vowel identification (Figure 6) shows similar patterns to those of Experiment 1: suppressing $F_1$ attenuates the effect of $F_1$ variation on vowel identification, while suppressing $F_2$ again had a smaller effect on the influence $F_1$ variation has on vowel identification.

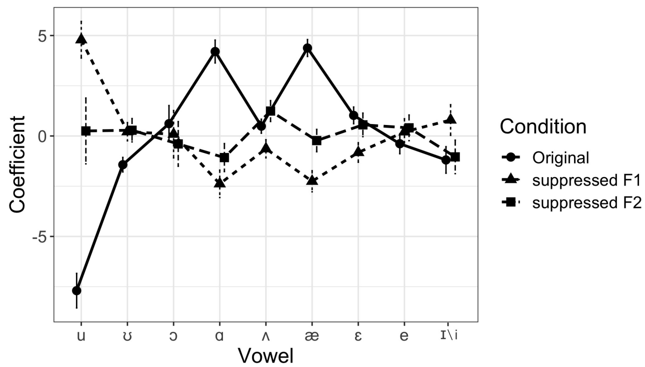



FIG. 6. Multinomial logit model coefficients per condition for $F_1$ in Experiment 2. Vertical lines indicate one standard error. The category on the far right combines responses to /i/ and /ɪ/.

In Figure 7, which shows the coefficients for $F_2$, we also see a trend similar to Experiment 1 for the original vowels (circles). As there, suppressing $F_1$ barely has any effect, and the coefficients for this condition (triangles) are all close to 0. However, we now see that suppressing $F_2$ creates the same kind of attenuated mirror image pattern shown in Experiments 1 and 2 for the suppressed $F_1$ condition: The perceptual effects of $F_2$ are weakened when $F_2$ formant peak is attenuated in Experiment 2.

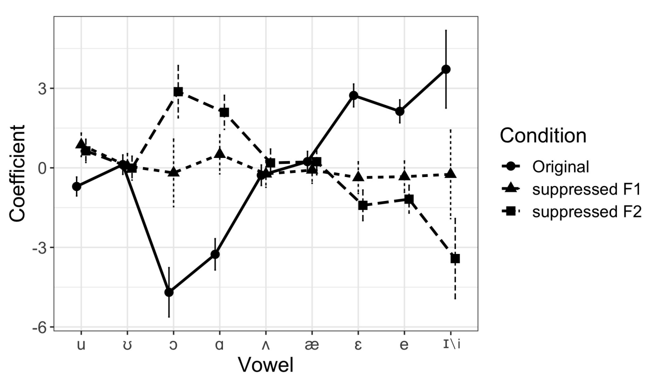



FIG. 7. Multinomial logit model coefficients per condition for $F_2$ in Experiment 2. Vertical lines indicate one standard error. The category on the far right combines responses to /i/ and /ɪ/.

### C. Discussion

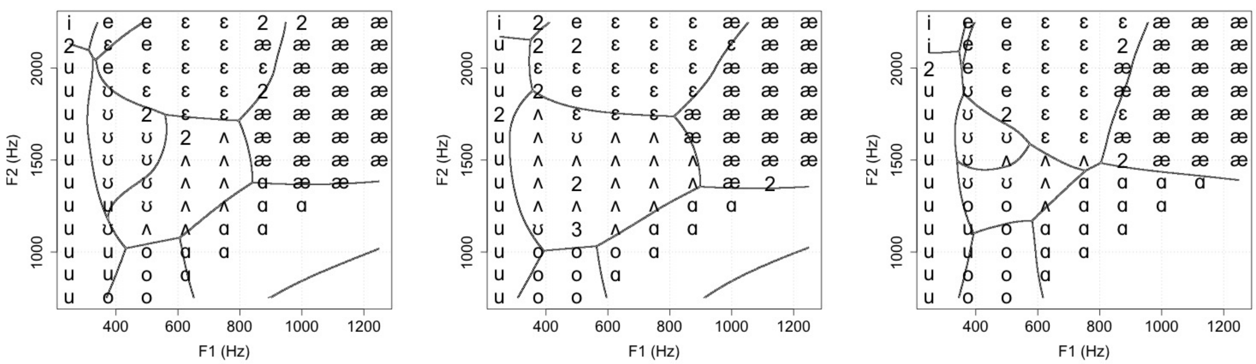
The results of Experiment 2 for the most part replicate the findings from Experiment 1, and the basic findings have now been confirmed in two dialects of English and with either fixed or randomized block order. The sole inconsistency between the experiments is the effect of suppressing $F_2$, which had little effect in Experiment 1. In Experiment 2, however, suppressing $F_2$ increased response entropy and affected how participant responses vary as $F_2$ changes. This may be due to dialect differences. Another cause may also be block order. $F_2$-suppressed vowels were always presented last in Experiment 2, when the participants could have been fatigued by the session and responded with reduced attention.

## IV. EXPERIMENT 3

Listening to vowels that have suppressed formant peaks may be easier if stimulus manipulation is consistent within blocks. The goal of the third experiment was to test whether identifying vowels with suppressed formants when they are presented in the same block with the original synthesized vowels impedes participants' ability to accommodate the missing information by relying on other aspects of the entire spectrum.

### A. Method

A new group of thirteen native speakers of Western Canadian English (18 – 35 years; M = 21.62; SD = 4.34; 2 male, 10 female, one participant did not wish to disclose gender information) participated in the third experiment. All participants were recruited from




(a) (b) (c)

FIG. 8. Phoneme boundaries and modal responses for the (a) original synthesized vowels, (b) vowels with $F_1$ suppressed, and (c) vowels with $F_2$ suppressed in Experiment 3. The number 2 is used when two responses tied.

the University of Alberta following the same guidelines as in the previous experiments. The same stimuli and the procedure as in Experiment 2 were used, except that the three separate blocks, each containing a single condition (original, $F_1$-suppressed, $F_2$-suppressed), were replaced by three blocks each containing an equal number of randomly selected vowels from each of the three conditions (the blocks were balanced). In other words, the experiment switched among the three stimuli types from trial to trial.

### B. Results

Contour plots of participant responses in Experiment 3 (Figure 8) resemble those of Experiment 1 and 2. The distribution of responses between conditions is similar, and the differences in responses persist along the suppressed formant axis (e.g., /ɛ/ and /æ/ for $F_1$-suppressed, and /ɛ/ and /ɑ/ for $F_2$-suppressed).

Shannon entropy values were again different between conditions ($F(2, 190) = 22.27$, $p < .01$): responses in the control condition had lower entropy than responses in both the $F_1$-suppressed ($t(190) = -6.63$, $p < .001$) and the $F_2$-suppressed condition ($t(190) = -3.98$, $p < .001$), while responses in the $F_1$-suppressed condition had slightly higher entropy than responses in the $F_2$-suppressed condition ($t(190) = 2.65$, $p = .03$).

We also ran multinomial logit models for responses collected in Experiment 3. Although the magnitudes of the effects are somewhat smaller, the coefficient patterns for $F_1$ are similar overall to those from Experiment 1. Suppressing $F_1$ led to reduction of $F_1$ coefficients, indicating its limited importance in vowel selection (Figure 9). One noticeable difference between Experiment 1 and Experiment 3 are the smaller coefficients for vowel /i/.

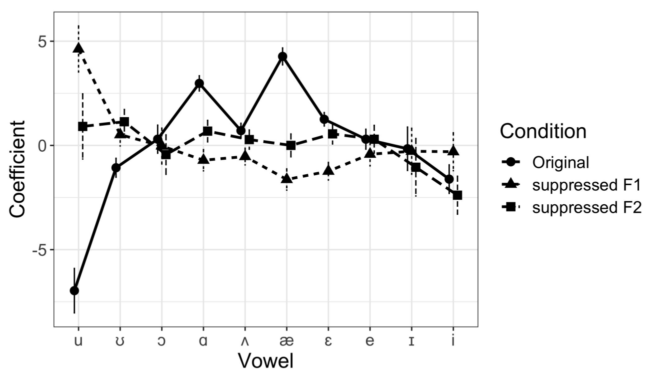



FIG. 9. Multinomial logit model coefficients per condition for $F_1$ in Experiment 3. Vertical lines indicate one standard error.

Considering $F_2$ coefficients, $F_2$ peak suppression had a more noticeable effect on $F_2$ coefficient change in Experiment 3 than in Experiment 1, although these effects were still not particularly large. The dashed line connecting squares in Figure 10 ($F_2$-suppressed) appears to be an attenuated mirror image of the solid line (original vowels), particularly in vowels such as /u/, /o/, /ɪ/, and /i/. Not surprisingly, suppressing $F_1$ had little impact on $F_2$ coefficients, except in the case of /i/ where we note a small effect.

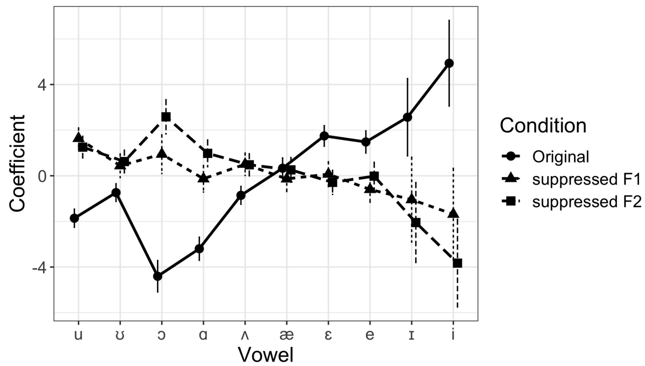



FIG. 10. Multinomial logit model coefficients per condition for $F_2$ in Experiment 3. Vertical lines indicate one standard error.

### C. Discussion

The contour plots of condition-randomized Experiment 3 for the most part mimic those obtained in the condition-blocked Experiment 2, indicating that the participants' are able to deal with variable missing formant information on a stimulus-by-stimulus basis; that is, it does not require a stable change in the stimuli over longer periods of time as in the case when the conditions are placed in separate blocks.

Taken together, results from Experiments 1-3 all point to the same conclusions. On the one hand, suppressing either $F_1$ or $F_2$ does not have an overwhelming effect on vowel identification — contour maps of responses resemble the original pattern; that is, suppressing a formant does not consistently lead to perception of a different vowel for any stimulus. Furthermore, differences in vowel identification along the axis of the suppressed formant peak are noted for both $F_1$-suppressed and $F_2$-suppressed stimuli, indicating that the missing local information can to an extent be replaced or recovered from the rest of the spectrum. On the other hand, quantitative analyses show significantly lower agreement in participant responses if a formant is suppressed.

## V. EXPERIMENT 4

Although the uncertainty in which vowel to select as a response increases when a formant is suppressed, we still noticed that there is considerable participant disagreement in responses to original stimuli as well. Some of the stimuli were probably unusual and difficult for participants to place as they are relatively remote from typical spectral patterns of any English vowel, given that they were synthetic monophthongs. Additionally, vowels /ɪ/ and /i/ were rarely chosen by listeners. In Experiment 4 we wanted to present our participants with a set of stimuli with formant patterns based on averages measured in a dialect of Canadian English and to investigate how attenuating formants of such stimuli influences their identification.

### A. Method

A new group of 11 native speakers of Western Canadian English (19 – 33 years; M = 22.45; SD = 3.77; 3 male, 8 female) participated in the fourth experiment. All participants were recruited from the University of Alberta following the same guidelines as in the previous experiments.

We synthesized 10 Canadian English vowels as described by Nearey and Assmann (1986) in terms of both formant frequency values and formant frequency changes (see also Hillenbrand *et al.*, 1995). This did not alter the choice set used in Experiments 1-3 as response options, except that we decided to mark /e/ as /eɪ/ and /o/ as /oʊ/ in the response choices to better represent the formant value change in these now-diphthongs. All the formant frequencies in Nearey and Assmann (1986) were scaled down by 1.06 to make the voice more male as original values were averages of both male and female speakers. We then used the formula from Nearey (1989) to calculate $F_3$ values. The formula for front vowels is given in Equation 2a and the formula for the back vowels is given in Equation 2b. These formulae were applied separately to the target values of the first and last frames of the vowel. The vowels were synthesized at each 8 ms frame with 4 ms overlap using the same procedure as Ito *et al.* (2001) to suppress either $F_1$ or $F_2$. Each window was combined with an overlap add procedure after applying a 8-ms Hamming window (again with 4-ms overlap). In a few frames the procedure was unable to locate two harmonics that met the criteria for removing a formant peak and those frames were created as repetitions of the previous frame. The duration of all synthesized vowels was 400 ms.

$$\text{front}F_3 = 0.522F_1 + 1.197F_2 + 57 \tag{2a}$$

$$\text{back}F_3 = 0.7866F_1 - 0.365F_2 + 2341 \tag{2b}$$

A total of 30 stimuli (10 stimuli in each of the three conditions) were created in this manner. The stimuli and a table specifying their formant values are included in the supplementary material. Note that all stimuli were now in the realm of realistic vowel formant values for a listener of Western Canadian English, and that they all included varying degrees of $F_1$ and/or $F_2$ change, as well as correlated $F_3$ change. The same procedure as in Experiment 3 was used except for the number of unique stimuli. Since only 10 vowels in three conditions were synthesized in Experiment 4 (30 different stimuli), each of the vowels was presented to the participants three times for a total of 90 stimulus presentations excluding practice.

## B. Results

Cochran-Mantel-Haenszel tests (presented in supplementary materials along with confusion matrices) show highly significant effects of formant suppression in Experiment 4. For brevity, we focus on general patterns of change (or lack thereof) in response patterns associated with changes in condition. Responses to /ɑ/, /æ/, /ʌ/ are for the most part unaffected, while responses to /i/, /ʊ/, and /u/ are only slightly affected by formant peak attenuation.

The bulk of the change in vowel identification occurs in four base stimuli due to suppression of $F_1$. For the vowel stimulus /ɪ/, the responses are identical in the original and $F_2$-suppressed condition: two thirds of the responses are correct, there are 24.24% /ɛ/, and 9.09% /eɪ/ responses. When $F_1$ is suppressed, however, only 39.39% of the responses are correct, 24.24% are /ɛ/, and other responses are spread across most other remaining options. In the case of vowel /ɛ/, 84.85% of the responses are accurate in the original and 87.88% in the $F_2$-suppressed condition. In the $F_1$-suppressed condition, however, only 33.33% of the responses are correct, with /ɪ/ receiving 33.33% and /æ/ receiving 21.21% of the responses. Virtually all responses to the /oʊ/ vowel are correct except when $F_1$ is attenuated, where only 57.58% are correct and a third of the responses becomes /ɑ/. The most notable difference, however, occurs for the diphthong /eɪ/. Again virtually all responses to this vowel in the original and the $F_2$-suppressed condition are correct, but in the $F_1$-suppressed condition only one is correct, 72.73% of responses become /i/ instead, and others are spread across remaining options.

### C. Discussion

The results of Experiment 4 yielded two important findings: first, we see that the listeners mostly agree on which vowel they are presented with if its $F_1 \times F_2$ combination and formant change fit the ordinarily encountered values. In our previous experiments, such high agreement in responses, even to control stimuli, was rare.

Second, suppressing a formant may or may not lead to changes in perception, depending on the original vowel. Large changes were noted for vowels /ɪ/, /ɛ/, /eɪ/, and /oʊ/, but smaller changes were noted for /i/, /ʊ/, and /u/, and especially for /ɑ/, /ae/, and /ʌ/. At first glance, there are no vowel features exclusive to the vowels which were affected by the experimental manipulation of formants. However, if we take note that the changes in vowel identification were registered in the suppressed $F_1$ condition, a pattern emerges: according to Nearey and Assmann (1986), /ɪ/, /ɛ/, /eɪ/, and /oʊ/ have magnitudes of $F_1$ change throughout their production larger that 100 Hz, while other vowels never reach an $F_1$ change of more than 50 Hz.

These results do show that information loss from attenuating a formant, when that formant is not changing appreciably, can be compensated for by using other information in the signal. For many vowels, we recorded no changes despite suppressing $F_1$ or $F_2$, and even in those vowels where we did, listeners were still somewhat successful in responding correctly. On the other hand, vowels with substantial movement in $F_1$ showed substantial information loss when the $F_1$ peak is suppressed. By contrast, $F_2$ suppression has little effect even for vowels like /eɪ/ and /oʊ/ that have substantial F2 movement. Indeed $F_2$ suppression has very little effect in any of the four experiments reported here.

## VI. GENERAL DISCUSSION

The dominant "formant hypothesis" of vowel identification was challenged by findings of the study by Ito *et al.* (2001) in which suppressing formant peaks did not radically change vowel identification. The authors instead argued in favor of the "whole-spectrum hypothesis" in which the gross spectral shape is used as a cue by listeners when deciding which vowel was heard. In the present paper, we attempted to replicate this finding in two dialects of English, which both include more vowel categories than the Japanese vowel system. We also subjected the data to detailed quantitative analyses, which yielded insights beyond simply observing vowel plots. Finally, we also took a step towards assessing the usefulness or reliability of the gross spectral shape when vowels are presented under more ecologically valid circumstances.

Visual inspection of vowel plots in Experiments 1-3 leads to conclusions that at least partly match those of Ito *et al.* (2001). It appears that suppressing $F_1$ or $F_2$ peak does not prevent listeners from making vowel distinctions along that formant's frequency axis. In other words, suppressing a formant peak does not cause that formant to perceptually "disappear" or be reassigned perceptually to the next preserved formant peak. Instead, listeners appear to be able to either compensate for the missing formant with some other spectral property or to estimate its frequency value using other available cues in the acoustic signal.

Visual inspection may not reveal differences between experimental conditions that are evident when quantitative analysis is performed. However, comparing the entropy of participant responses showed that participants diverge more in their selection if the first formant was suppressed. We take this lack of agreement as an indication of uncertainty of vowel categorization. Similarly, varying $F_1$ in $F_1$-suppressed stimuli has a smaller effect on vowel selection than when the original unmodified vowels are presented. In Experiment 1, these results did not extend to $F_2$-suppressed vowels. However, in Experiments 2, although less salient, and 3 these effects appear for $F_2$-suppressed vowels as well.

These results point to two main conclusions. First, even when formant peaks are missing, listeners can use other cues to identify vowels in a way that does not deviate as much as would be expected if information near formant peaks formed the sole basis for vowel identification. Second, formants may still provide the most important cues, as they cannot be suppressed and then fully and faithfully replaced with some other source. Neither the "formant hypothesis" nor the "whole-spectrum hypothesis" fully correspond to these findings. We acknowledge that listeners do not rely solely on frequencies near peak formant amplitudes and that they can use additional information about general spectral shape in choosing among vowel categories. However, loss of information near formant peaks often distorts vowel perception considerably. Some of this may simply be because such local modifications distort part of the overall spectral shape. Nevertheless, there is good evidence in the literature (see introduction) and in our experiments that high amplitude components near formant peaks have greatest weight in perception in many circumstances.

In Experiment 4 we presented participants with synthesized vowels with changing formants that better match vowels from actual speech (Hillenbrand *et al.*, 1995; Nearey and Assmann, 1986), and the results were markedly different. Participant agreement was higher in Experiment 4 as they were presented with vowels that (1) had formant frequency values closer to their dialect, (2) some degree of formant frequency change rather than steady formant frequencies, and (3) multiple presentations of the same stimulus.

Crucially, formant suppression barely affected certain vowels, whereas it lead to a reliable change in responses in others. We suggested that the source of this distinction could be in the extent $F_1$ changes throughout the vowel, with larger vowel-specific patterns of change being associated with difficulty in recognizing the vowel if $F_1$ is suppressed: vowels for which the listener needs to account for the extent and speed of change in formant frequency are affected by disruption caused by the formant peak being flattened (this may mean that a suppressed $F_2$ is easier to estimate when dynamic formant values are used as well; see our supplementary materials for an analysis predicting missing formant frequency from other nearby formants showing better results for $F_2$). If this claim is true, then the gross spectral shape (which will retain some evidence of the suppressed formant's movement and changes in, e.g., the levels of the upper spectral components) is insufficient to fully replace or recover formant information, at least for the range of stimuli used in our experiments. In other words, participants can use the gross spectral shape to remedy losses in the most important regions (formants), but only if the gross spectral shape of the vowel (formants of course included) is steady (see also Kiefte and Kluender, 2005, where spectral tilt effects were greatly diminished in diphthongal stimuli).

These findings come from experiments which tested University students that completed an introductory course of phonetics. We cannot guarantee that these results would not differ somewhat if participants were naive listeners. However, we wanted to avoid artifacts associated with orthographic ambiguity of English vowels (Assmann *et al.*, 1982) and we have no reason to believe that vowel perception in listeners with relatively modest training in the use of phonetic symbols is different from the general population.

In the present study, we regarded two extreme positions on the role formant peaks (versus the gross spectral shape) have in vowel perception. Other approaches may assume that slightly more than just formant peaks, i.e., additional yet still local features such as "shoulders" of the formant peaks may be relevant and guide vowel identification. This notion merits investigation, but was not the focus of the current study. However, we include a "peak-and-shoulder" analysis in the supplementary material.

Finally, it is only fair to note that in Experiment 4 we artificially suppressed formant peaks as they shifted along the formant axis, not particular frequency bands. Hearing loss or background noise usually cover a particular frequency band, meaning that a formant peak may be obscured only for a portion of the vowel signal, not its entirety. Therefore, future studies could investigate vowel identification using stimuli that have an attenuated stop band that partly coincides with the changing formant values. This kind of manipulation is only one way to increase ecological validity of the experiments. Experiment 4 introduces synthesized stimuli that are clearly at least a step closer to naturally spoken English vowels than are pure steady-state stimuli. However, more could be done to better represent everyday listening/speech perception conditions and we see three avenues to explore. The first is to investigate synthesized vowel identification in carrier or precursor sentences (see also Kiefte and Kluender, 2008) with varying degrees of formant/spectral shape attenuation. The second option is to present manipulated vowels in background noise or with some other kind of interference, matching the noisy environment in which we usually listen to speech. The third is to present listeners with actual vowel recordings (made in or out of word/sentence context), where some would have attenuated formant peaks or noise bands coinciding with formant peak frequency.

[1]For brevity, we will use the term "gross spectral shape" to not only mean very long range spectral properties like spectral balance or overall tilt across the spectrum, but also to include possibly more focused local features such as the amplitudes of those formant peaks that are not suppressed in the stimulus. That is, from the perspective of a suppressed formant peak, "gross spectral shape" will be a shorthand for any aspect of the spectrum other than the frequency (and amplitude) of the formant peak itself.

[2]See Supplementary materials at [URL will be inserted by AIP] for additional analyses and figures.

Assmann, P. F., Nearey, T. M., and Hogan, J. T. (**1982**). "Vowel identification: Orthographic, perceptual, and acoustic aspects," The Journal of the Acoustical Society of America **71**(4), 975–989.

Bladon, A. (**1982**). "Arguments against formants in the auditory representation of speech," The representation of speech in the peripheral auditory system 95–102.

Bladon, A. (**1983**). "Two-formant models of vowel perception: Shortcomings and enhancement," Speech Communication **2**(4), 305–313.

Bladon, R., and Lindblom, B. (**1981**). "Modeling the judgment of vowel quality differences," The Journal of the Acoustical Society of America **69**(5), 1414–1422.

Chistovich, L. A., and Lublinskaya, 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**(3), 185–195.

Croissant, Y. (**2013**). *mlogit: multinomial logit model*, https://CRAN.R-project.org/package=mlogit, r package version 0.2-4.

Delattre, P., Liberman, A. M., Cooper, F. S., and Gerstman, L. J. (**1952**). "An experimental study of the acoustic determinants of vowel color; observations on one-and two-formant vowels synthesized from spectrographic patterns," Word **8**(3), 195–210.

Fox, R. A., Jacewicz, E., and Chang, C.-Y. (**2010**). "Auditory spectral integration in the perception of diphthongal vowels," The Journal of the Acoustical Society of America **128**(4), 2070–2074.

Glasberg, B. R., and Moore, B. C. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hearing Research **47**(1-2), 103–138.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of american english vowels," The Journal of the Acoustical society of America **97**(5), 3099–3111.

Hillenbrand, J. M., and Houde, R. A. (**2003**). "A narrow band pattern-matching model of vowel perception," The Journal of the Acoustical Society of America **113**(2), 1044–1055.

Hillenbrand, J. M., Houde, R. A., and Gayvert, R. T. (**2006**). "Speech perception based on spectral peaks versus spectral shape," The Journal of the Acoustical Society of America **119**(6), 4041–4054.

Hillenbrand, J. M., and Nearey, T. M. (**1999**). "Identification of resynthesized/hvd/utterances: Effects of formant contour," The Journal of the Acoustical Society of America **105**(6), 3509–3523.

Ito, M., Tsuchida, J., and Yano, M. (**2001**). "On the effectiveness of whole spectral shape for vowel perception," The Journal of the Acoustical Society of America **110**(2), 1141–1149.

Kakusho, O., Hirato, H., Kato, K., and Kobayashi, T. (**1971**). "Some experiments of vowel perception by harmonic synthesizer," Acta Acustica united with Acustica **24**(4), 179–190.

Kiefte, M., Enright, T., and Marshall, L. (**2010**). "The role of formant amplitude in the perception of/i/and/u," The Journal of the Acoustical Society of America **127**(4), 2611–2621.

Kiefte, M., and Kluender, K. R. (**2005**). "The relative importance of spectral tilt in monophthongs and diphthongs," The Journal of the Acoustical Society of America **117**(3), 1395–1404.

Kiefte, M., and Kluender, K. R. (**2008**). "Absorption of reliable spectral characteristics in auditory perception," The Journal of the Acoustical Society of America **123**(1), 366–376.

Kiefte, M., Nearey, T. M., and Assmann, P. F. (**2013**). "Vowel perception in normal speakers," Handbook of vowels and vowel disorders **2**, 160.

Klatt, D. (**1982**). "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*

*ICASSP'82.*, IEEE, Vol. 7, pp. 1278–1281.

Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," the Journal of the Acoustical Society of America **67**(3), 971–995.

Maddox, W. T., Molis, M. R., and Diehl, R. L. (**2002**). "Generalizing a neuropsychological model of visual categorization to auditory categorization of vowels," Perception & Psychophysics **64**(4), 584–597.

Molis, M. R. (**2005**). "Evaluating models of vowel perception," The Journal of the Acoustical Society of America **111**(2), 2433–2434.

Nearey, T. M. (**1989**). "Static, dynamic, and relational properties in vowel perception," The Journal of the Acoustical Society of America **85**(5), 2088–2113.

Nearey, T. M. (**1990**). "The segment as a unit of speech perception.," Journal of Phonetics .

Nearey, T. M. (**1997**). "Speech perception as pattern recognition," The Journal of the Acoustical Society of America **101**(6), 3241–3254.

Nearey, T. M., and Assmann, P. F. (**1986**). "Modeling the role of inherent spectral change in vowel identification," The Journal of the Acoustical Society of America **80**(5), 1297–1308.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," The Journal of the acoustical society of America **24**(2), 175–184.

R Core Team (**2017**). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Rosner, B. S., and Pickering, J. B. (**1994**). *Vowel perception and production.* (Oxford University Press).

Shannon, C. E. (**1948**). "A mathematical theory of communication," Bell System Technical Journal **27**, 379–423.

Yu, D., and Deng, L. (**2014**). *Automatic Speech Recognition: A Deep Learning Approach* (Springer).

Zahorian, S. A., and Jagharghi, A. J. (**1993**). "Spectral-shape features versus formants as acoustic correlates for vowels," The Journal of the Acoustical Society of America **94**(4), 1966–1982.