Spring 2012

Interaction between source and filter in vowel identification

Christopher James Becker
James Madison University

Interaction between source and filter in vowel identification

Chris Becker

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Arts

Department of Graduate Psychology

May 2012

Acknowledgements

I owe an immense debt of gratitude to my research advisor, Dr. Michael Hall, who has guided me, mentored me, and put up with me for four years. Without him this would not have been possible. I would also like to thank my wonderful thesis committee members, Dr. Jeff Andre and Dr. Jeff Dyche (listed alphabetically, not in order of importance). Additionally, I could not have completed this project without the help and support of my friends and family. I dedicate this thesis to them.

Table of Contents

List of Tables

List of Figures

Abstract

Speech sounds can be modeled as a product of two components: the source and the filter. The effects of filter manipulations on speech perception have been studied extensively, while the effects of source manipulations have been largely overlooked. This study was an attempt to assess the impact of source manipulations on vowel identification. To this end, two source manipulations were conducted prior to filtering. First, several harmonics of the source sawtooth wave that were located near formant peaks were mistuned, either towards or away from the peaks. Mistuning towards formant peaks was expected to facilitate vowel identification by helping to convey the position of the formant more clearly; mistuning away was expected to hinder performance. Consistent with this hypothesis, a significant effect of mistuning was observed. However, follow up analyses revealed that this manipulation only had an effect in conditions where harmonics are mistuned away from formant peaks by a large degree (5%). The second manipulation consisted of adding noise to the source signal to "fill in" the acoustic spectrum. Because the addition of noise occurred before filtering, the spectral shape of the noise component was identical to that of the harmonic portion of the tone, and was expected to help convey formant peaks, especially when they were not well conveyed by harmonic information. The results reveal that the addition of noise had no effect on vowel identification. Possible stimulus based explanations for the failure to observe some of the hypothesized effects are discussed.

Introduction

There is a large body of research indicating that speech perception is critically dependent upon formants, or relatively intense regions of the acoustic spectrum (e.g. Delattre, Liberman, Cooper, & Gerstman, 1952; Peterson & Barney, 1952; Pols et al., 1969; Nearey, 1978; Miller, 1989). Formants are generated by a combination of a sound source and a filter. The effects of manipulations of filter parameters have been studied extensively (e.g. Stevens 1959; ter Keurs, Festen, & Plomp, 1991; Holt, Lotto, and Kluender, 2000). By contrast, the effects of source manipulations have been less thoroughly examined. The current investigation represents an attempt to assess the effects of source manipulations on speech perception.

Source-Filter Theory

One theory of sound production, called the source-filter (or acoustic) theory, models sound production as a two stage process (Fant, 1960; Handel, 1993). The first stage is the generation of vibrations by a source. The source typically generates a complex tone, consisting of multiple harmonics, which steadily decrease in amplitude as they go up in frequency. In the second stage, the source energy is transferred to a resonating body that radiates the energy into the air, where it can be perceived as sound. This resonating body modifies the initial source signal based on its own modes of vibration, or resonances, which are determined by its size, shape, and the material from which it is made. Thus, the relative amplitudes of different parts of the acoustic energy spectrum are selectively amplified or attenuated (i.e. filtered) based on their frequency. Importantly, the source and filter

are characterized as functioning at least quasi-independently such that filter resonances can be manipulated independently of the source, and its resonances can exist at frequencies where there is no source energy present. It is the interaction between source and filter that forms the basis for the current investigation.

The source-filter theory is best applied to systems which exhibit little coupling. Coupling refers to the degree to which the properties source and filter are non-independent (i.e. coupled), or cannot be manipulated independently. Because the theory assumes independence of the source and filter, it does not apply well to systems in which this assumption does not hold (Askenfelt, 1991). For example, woodwind instruments exhibit a high degree of coupling and the source-filter model can only be applied if a strong feedback pathway is included. Often it is better to model these systems in other ways. Other systems, such as the human voice or string instruments, typically exhibit only a minimal degree of coupling, and can be effectively modeled by the source-filter theory.

The source-filter theory was initially applied to the human voice (Fant, 1960). In the case of human speech, the source is the glottis (vocal folds) and the filter is the supralaryngeal vocal tract, which includes the pharynx and the oral and nasal airways. The vocal tract selectively amplifies and attenuates the source energy at certain frequencies based on the shape, length, and volume of the tubes (cavities) that the air is flowing through. These properties can be modified by changing the position of the articulators, allowing the speaker some degree of control over the spectral characteristics of their voice. In this way, the source signal is sculpted into a meaningful utterance by the movement of the tongue and lips.

In English, the production of different phonemes is largely determined by this talker-controlled spectral variation. Generally, in speech changing the position of articulators changes the resonances of the supralaryngeal vocal tract. These, in turn, will change the characteristics of the filter function. This largely determines the frequency location of relatively intense regions of the spectrum, or formants. The difference between the location of formant peaks is a large part of what allows listeners to differentiate between phonemes. In this way, a signal that contains meaningful phonetic information is created from the combination of the source and filter. The role of interactions between source and filter in phoneme (vowel) perception will provide the focus of the current investigation.

Although initially applied to speech, the source-filter theory also applies to a variety of other sound-producing systems to varying degrees (Handel, 1993). It has a particularly strong application to string instruments, as they share many attributes with the human voice (Askenfelt, 1991). For example, in the violin the sound source is the vibrational energy of the string, and the filter properties are determined by the resonances of the instrument body. Both a violin and the human voice have source spectra that decrease in intensity by about 6 dB per octave. Additionally, both share some degree of aperiodicity. In the voice this is caused by idiosyncrasies in the tissue of the vocal folds that affect the rate at which they open and close. In the violin, aperiodicity is caused by irregularities in the hair of the bow (for example, small differences in frictional coefficients of the hair along the length of the bow) or irregularities of the string, such as resin buildup. Furthermore, increasing the intensity of both voices and strings causes a disproportionate

increase in the number and intensity higher harmonics, leading to a "brighter" sound. Finally, pitch can be controlled continuously in both cases. In the voice, this is accomplished by changing the tension of the vocal folds, which changes the rate of glottal pulses. More tension leads to a higher pitched sound. Similarly, more tension applied to a violin string will lead to a higher pitched sound. The length of a violin string also affects the pitch, with shorter lengths leading to higher pitch.

Assumptions regarding a specialized perceptual mechanism for speech processing

The mechanism by which the listener abstracts meaningful phonetic information from the filtered speech stimulus has been the subject of much debate. One school of thought holds that speech is perceived on the basis of acoustic cues present in the signal, and that these distinctive acoustic cues are the fundamental objects of speech perception (e.g. see the TRACE model of McClelland and Elman, 1986; the Fuzzy Logical Model of Perception of Oden and Massaro, 1978; also see Stevens, 2002). In this view, the acoustic signal of a speech utterance is generally processed in the same way as any other sound. Thus, individual phonemes are identified based on the combination of their acoustic characteristics. Phoneme perception is directly dependent on the collection of acoustic cues that are present in a signal, so changing an acoustic cue (e.g. the position of a formant) can result in the perception of a different phoneme.

Alternatively, it has been suggested that a phonetic module exists to process speech separately from other types of sounds (e.g., see Liberman, 1981; Liberman and Mattingly, 1985). A module, in this sense, is a functionally distinct structure, which has acquired biological specialization for processing phonetic information

through evolution. The phonetic module is argued to be closed, meaning that it does not interact with other modules. Therefore, its processing is dependent only on its inputs, not on the inputs to, or processing of, other modules. (In contrast, the processing of open modules is not isolated from the effects of other modules and open modules do interact with one another.) Thus, inputs to this module are processed independently of general auditory information such as pitch, timbre, and loudness, and the percepts emerging from this module are strictly phonetic in nature (Liberman & Mattingly 1985). It has also been proposed that the phonetic module has priority over general auditory processing, meaning that auditory inputs are first processed by the phonetic module, and leftover energy is then passed on to the open general auditory processing modules (Whalen & Liberman, 1987).

Categorical perception (CP) of speech represents the earliest potential behavioral support for a phonetic module. CP occurs when a continuous or gradual physical change over a stimulus continuum gives rise to a discrete set of perceptual responses. In other words, a physical continuum is perceived discretely or categorically (Studdert-Kennedy, Liberman, Harris, & Cooper, 1970). It was suggested that CP was the result of a phonetic module dividing a physical continuum into phonemically meaningful categories.

Another line of evidence that supports the existence of a phonetic module comes from studies of duplex perception (DP; e.g. Rand, 1974). DP is a laboratory phenomenon in which a single auditory stimulus simultaneously contributes to two percepts. The interpretation was that the two percepts represented the outputs from two distinct processing modules.

However, the interpretation that CP and DP represent sufficient evidence to conclude in favor of the existence of a phonetic module has been called into question by related studies. For example, Kuhl and Miller (1975) discovered that the labeling functions and phonetic boundaries of several stop-consonants were very similarly caegorical for humans and chinchillas. Additionally, Miller, Wier, Pastore, Kelly, and Dooling (1976) found that noise-buzz sequences were categorically perceived. Furthermore, CP of musical intervals has been observed in several studies (e.g. Locke and Kellar, 1973; Siegel and Siegel, 1977; Burns and Ward, 1987). Similarly, a version of DP has been observed for musical chords (Hall and Pastore, 1992) and door-slam sounds (Fowler and Rosenblum, 1990)

The Motor Theory of Speech Perception is related to the study of the proposed phonetic module (Mattingly and Studdert-Kennedy, 1991). In contrast with the general auditory explanation that speech sounds are processed the same way as nonspeech sounds, this theory holds that the basis of speech perception is the listener's recreation of the articulatory gestures of the speaker. In this view, the phonetic module facilitates the abstraction of the position of the speaker's articulators. Thus, in the Motor Theory's framework, articulatory gestures (i.e. mouth movements), not acoustic cues, are the fundamental objects of speech perception. However, articulatory gestures are conveyed acoustically. Therefore, references to articulatory gestures are largely omitted from this manuscript simply for ease of communication, as the stimuli will be synthetic and the manipulations are acoustic in nature because there is no actual talker. It is important to note that the current investigation does not assume a particular theoretical perspective.

Referring to acoustic characteristics that serve as cues to vowel identification does not represent a theoretical bias, as it does not presume that a corresponding argument could not be made with reference to articulatory gestures.

Effects of the Filter on Phoneme Identification

Vowel identification is primarily determined by formant center frequency values (Nearey, 1978), which are the product of the filter. Formants are relatively intense regions of a spectrum (i.e., narrow bands of frequencies) that typically correspond to the resonances of the filter (Titze, 1994). Formants are conventionally numbered from the lowest frequency resonance to the highest such that the first formant is lowest in frequency, the second is the next lowest, and so on. In vowels formants appear as "steady-states" in that their center frequencies do not change rapidly over time. It has been widely suggested that the center frequencies of formant steady states beyond the second or third are largely irrelevant to vowel identification (e.g. see Delattre, Liberman, Cooper, & Gerstman, 1952; Peterson & Barney, 1952; Pols et al., 1969; Nearey, 1989; Miller, 1989). Many have suggested as few as two formants can be sufficient for reliable vowel identification; the Peterson and Barney (1952) data are consistent with this suggestion.

Center frequencies of the first two formants of vowels vary as a function of tongue height and position. Tongue height is defined relative to the jaw or to the roof of the mouth, and determines the center frequency of the first formant. The higher the tongue in the mouth, the lower the first formant center frequency. Tongue position, or "backness", is measured as the distance of the tongue from the back of the mouth. This controls the second formant center frequency, with higher

second formant center frequency values being associated with tongue positions closer to the front of the mouth. Since vowels can be distinguished by F1 and F2 center frequencies, they are often labeled by speech researchers with respect to both tongue height and position. For example, /a/ represents a low-back vowel and /i/ represents a high-front vowel. As a low-back vowel, /a/ has a high first formant and a low second formant, creating a compact spectrum with respect to the first two formants. On the other hand, /i/ has a low first formant and a high second formant, creating a diffuse spectrum. Thus, the position of articulators have filter consequences that result in perceptually meaningful changes to the formants.

Unlike vowel steady states, consonant formant center-frequencies show rapid change over time. They are often completed in 50-70ms, and therefore are often referred to as "transitions". Despite these rapid spectral changes, consonant perception also has repeatedly been shown to depend critically upon formant center frequencies (e.g., see Delattre, Liberman, and Cooper, 1955). However, identification of consonant categories qualitatively differs from identification of vowels in that consonants are perceived discretely, or categorically, while vowels are perceived more continuously (Fry, Abramson, Eimas, and Liberman, 1962; McMurray and Spivey, 1999).[1]

The critical dependence of speech perception upon formant (i.e., filter) information has been demonstrated several ways. It is made particularly clear by a phenomenon called sine-wave speech (e.g., see Remez, Rubin, Pisoni, and Carrell, 1981). Essentially, sine-wave speech (SWS) demonstrates the ability of listeners to understand speech given only formant center frequencies. Typically, to generate a

sine-wave analogue from a natural speech token, the center frequencies of the first three formants are extracted from the natural signal. Then the token is resynthesizeded so that it consists of only three sine waves, with each following the center frequency of one of the three formants. Remarkably, given this type of signal conveying only the most fundamental filter properties, listeners are often still able to perceive and understand speech.

However, the ability to perceive SWS as speech seems to be dependent upon the maintenance of relationships between resonances indicated by the filter. For example, if the timing of the filter movement represented by one of the sine waves is manipulated, the intelligibility of the signal is greatly reduced. Remez, Ferro, Wissig, and Landau (2008) found that listeners were able to transcribe sine-wave speech tokens without temporal manipulation at 72% accuracy. (The same tokens were transcribed at 98% accuracy in a pretest where no asynchronous tokens were present, indicating that under ideal conditions sine-wave speech is quite intelligible.) However, when the sine wave representing the position of the second formant was desynchronized by +/- 100ms, they found that transcription accuracy decreased to 7%. This indicates that speech perception is critically dependent on the temporal properties of the filter.

Similarly, when the tuning of one of the sine waves is manipulated, intelligibility greatly decreases. Hall (2009) found that mistuning the sine wave following the second formant reduced intelligibility, and that reversing the same sine wave also led to greatly reduced intelligibility. This indicates that listeners are

sensitive to the filter positions in frequency space, which are represented by the sine wave, in addition to the filter's temporal characteristics.

Other demonstrations of the dependence of speech perception on filter properties have used more natural speech stimuli. For example, it has been found that vowel identification is influenced by the spectral content of surrounding phonemes, or the way the filters move before and after the steady-state vowel portion of a syllable. Holt, Lotto, and Kluender (2000) continuously varied second formant center frequencies of a consonant-vowel-consonant (CVC) syllable to create a continuum of vowel sounds, the endpoints of which were perceived as [ɛ] and [ɑ].. At some point along this continuum, there was a crossover point. On one side of the crossover point the vowels were more frequently labeled as [ɛ] and the other side as [ɑ]. In addition to varying second formant center frequency, they varied the consonant sounds that came before and after the vowel portion of the CVC syllable. Their results showed that the position of the crossover point shifted when the consonants surrounding the vowel changed. The authors attribute this effect to the spectral content of the surrounding consonants. Specifically, when the initial center frequency of the consonant transition (formant) is higher, the steady state portion is perceived as being lower in frequency.

In addition to surrounding-phoneme content, it has been found that surrounding-phoneme presence (or absence) influences vowel identification (Strange, Edman, and Jenkins, 1979). This again represents an effect on identification performance of the motion (or absence of motion) of the filters before and after a vowel. The researchers found that vowel identification was facilitated by

the presence of consonants before or after a vowel, with the presence of a consonant after a vowel leading to the best performance. The poorer performance for isolated vowels was attributed to a combination of factors, including the lack of dynamic spectral information and listeners being less familiar with isolated vowels than vowels in the context of consonants.

The rate of change of filter parameters also influences vowel identification. Specifically, it has been found that vowel duration significantly impacts vowel identification. For example, Stevens (1959) generated CVC syllables with varying vowel lengths from 20 ms to 500 ms, all with the same beginning and ending consonants. The differences in duration had strong effects on listeners' ability to distinguish between certain vowel pairs. In another condition where the same stimuli were mixed with noise, the effect of duration was even more pronounced.

Gottfried, Miller, and Payton (1990) manipulated the duration of the vowel in a CVC syllable, the position of the formants in that vowel, and the speaking rate of a sentence in which the CVC syllable was embedded. They found that all three manipulations led to changes in the labels applied to vowel pairs that are typically differentiated by both temporal and spectral characteristics in natural speech (/I/-/i/ and and /ɛ/-/æ/). This suggests three different effects of filter on vowel identification. First, changing the position of formant center frequencies led to vowels being assigned to different categories, indicating that the position of the filters has an impact. Second, the length of time the filters are steadily representing the formants of a given vowel before moving to represent the next consonant (i.e. duration) impacted vowel identification. Third, the global rate of change of the

filters at other times (i.e. the rest of the sentence) influenced vowel identification. This appears to be a general influence in speech perception as rate of speaking has been shown to influence consonant identification as well (Miller and Liberman, 1979).[2]

Additionally, it has been found that when filter information is conveyed less clearly, vowel identification suffers. One way this has been demonstrated is through spectral smearing (ter Keurs, Festen, & Plomp, 1991). Spectral smearing reduces the frequency resolution (i.e. spectral detail) of a signal, thus reducing the resolution of the filter properties conveyed by the signal. The overall effect of this manipulation is to "smooth" the spectrum. Therefore, the intensities of sharp formant peaks are reduced. Spectral smearing can be accomplished in a number of ways. The researchers smeared their spectra by projecting the spectral envelope of a signal onto a log-frequency scale and then convolving it with a Gaussian-shaped filter. The convolution of a function *f(x)* with a Gaussian function is called a Weierstrass transform, an example of which can be seen in Figure 1. In this figure it can be seen that the initial (grey) function is progressively smoothed as it is further convolved with a Gaussian function. The equivalent rectangular bandwidth (ERB) of the Gaussian-shaped filter (i.e. the frequency window size over which smearing was applied) determines the degree of spectral smearing, with larger bandwidths corresponding to a greater degree of smearing. The researchers used ERBs of 1/8, 1/4, 1/3, 1/2, 2, and 4 octaves in their study.

It was found that when spectral smearing occurred over a bandwidth greater than the ear's critical bandwidth (i.e. the ear's maximum frequency resolution),

phoneme identification in noise suffered. This was especially true of vowels, which indicates that the resolution of the filters affects vowel identification. It is likely that consonants were less affected because they are dynamic in nature and the movement of the filters was preserved, although their resolution as they moved was not. Vowels, on the other hand, are relatively static in character and thus suffered more from the smearing manipulation.

The impact of the source and its interaction with the filter

While it has been well established that filter properties play a substantial role in vowel identification, considerably less work has been done on the influence of source characteristics on vowel identification. One study that did look at source characteristics found that changing the fundamental frequency affected vowel identification (Barreda and Nearey, 2012). Participants listened to vowel sounds with a variety of fundamental frequencies and formant characteristics, and were instructed to identify the vowel, determine the gender of the person who produced it, and rate the size of the person who produced it. The researchers found that vowel identification was correlated with fundamental frequency. Through partial correlation analysis of the vowel identification, talker gender, and size rating data, the researchers determined that the effect of fundamental frequency was most likely indirect. They theorize that fundamental frequency affects vowel identification through its effects on the perceived size of the vocal tract that produced the signal. This, in turn, affects talker normalization, or the process whereby phonetic tokens are abstracted from the signal through the elimination of talker-specific idiosyncrasies.

Also largely missing from the literature are studies of the interaction between source and filter. This is important because the final signal is a product of both the source and the filters. The signal can only convey filter properties at frequencies where source information is submitted to it. This can be seen in Figure 2. In this figure, the formants that the speaker intended to convey are represented by the filter function, and are not physically present in the signal. Instead, the signal (output/resultant spectrum) consists of a complex tone produced by the source (source function) and then modulated by the filter (filter/transfer function), which represents the intended formants.

However, just because a formant was intended does not necessarily mean it will be well represented in the signal. This can also be seen in Figure 2. For example, a speaker could intend to convey a formant at 500 Hz, which corresponds to the lowest-frequency peak in the filter function in Figure 2b. If the fundamental frequency of their speech signal is 200 Hz, the closest harmonics will be located at 400 Hz and 600 Hz, and the formant will not be well represented. It will have to be inferred based on the relative intensities of the harmonics that are present. This can be seen by comparing the output spectra of Figure 2a and Figure 2b. The peak in the filter function at 500 Hz is more clearly represented in Figure 2a than in Figure 2b.This sort of interaction between source information and filter properties will form the basis of the current investigation.

While the majority of the literature has focused on the effects of filter characteristics, there have been a small number studies investigating interactions between the source and filter in relation to vowel perception. One such study found

that with a given set of filter parameters, higher fundamental frequencies led to poorer vowel identification (Ryalls and Lieberman, 1982). The authors conclude that it is likely that the reduction in accuracy was due to sparser spectral sampling at higher fundamental frequencies. At lower fundamental frequencies, the transfer function is sampled at more points (i.e. there is a greater density of spectral sampling) and accuracy rates were higher. For the previous example of a formant peak at 500 Hz, if the fundamental frequency was 100 Hz (as in Figure 2a) instead of 200 Hz (as in Figure 2b), there would be a harmonic at 500 Hz, and the peak would be much more clearly represented. However, this experimental design did not enable the researchers to rule out the possibility that their results were due to differences in the density of spectral sampling near formant peaks specifically, rather than differences in overall density of spectral sampling.

A related study found that shifting the fundamental frequency (similar to moving the fundamental from 200 Hz to 100 Hz to better represent the formant peak at 500 Hz) could lead to smaller formant-change detection thresholds in synthetic vowel-like sounds (Hermansky, 1987). Although this study did not look at this phenomenon in the context of phoneme perception, it nonetheless lends strong support to the hypothesis that source characteristics could influence vowel identification.

The current investigation was designed to extend these findings and test the hypothesis that any source manipulations that reinforce filter resonances (i.e. formant peaks) will facilitate speech perception. This hypothesis was tested with two types of stimulus manipulations in a vowel identification task. First, harmonics

located near formant peaks were slightly mistuned, either towards or away from formant center frequencies, and the effects on vowel identification were observed. It has been established that harmonics mistuned by a small amount will still be perceptually integrated in a syllable (Darwin and Gardner, 1985).

This manipulation will affect the density of spectral sampling near formant peaks while holding the overall density of spectral sampling constant (i.e. the total number of harmonics is unchanged, but they will be moved closer to or farther from formant peaks). If overall density of spectral sampling is truly the driving force behind Ryalls and Lieberman's results, this manipulation should have no effect. If, on the other hand, the density of spectral sampling near formant peaks is the driving force, then this manipulation should have a substantial effect on identification accuracy.

The second manipulation involved adding noise to the source. This was expected to "fill in" the spectrum between harmonics and therefore facilitate speech perception by conveying filter resonances (i.e. formant peaks) that would otherwise not be represented by source information in cases when formant center frequency falls between adjacent harmonics. In other words, it was anticipated that noise could compensate for a lack of harmonic information near a formant peak, and facilitate vowel identification in cases where harmonics alone are insufficient to convey filter resonances. The magnitude of this effect was expected to be greatest in cases where harmonics are mistuned away from formant peaks.

Method

Participants

The participants in this study were 16 undergraduate students from James Madison University who participated in partial fulfillment of course requirements, and the researcher. It is common for researchers to participate in perceptual studies, and the researcher's data were consistent with the data from the other participants. Because the study involved listening to and identifying English vowels, participants were required to be native English speakers with self-reported normal hearing. They were all between the ages of 18 and 24.

Stimuli

The stimuli were synthesized based on the measured location of formant center frequencies and bandwidths for American English vowels reported by Klatt (1980). These values are based on measurements from one speaker (Klatt himself). Using these values instead of using modal values from multiple speakers could potentially make the task more difficult. Such increased difficulty actually should prove beneficial in helping to avoid ceiling levels of performance. Additionally, Klatt's center frequency values fall within the acceptable ranges for the various vowels reported by Peterson and Barney (1952), which were based on data from 76 talkers. Thus, Klatt's (1980) values are reasonably ecologically valid.

Five vowels were used: /i/, /I/, /ɛ/, /æ/, and /ɑ/. These vowels were selected for proximity in F1/F2 space because this was expected to reduce the chance of ceiling performance (see Figure 3). Additionally, they were selected so

that they each contained unique first and second formant center frequencies relative to the other vowels in the set (i.e. no two vowels share a center frequency value for either formant; this is also the case for the same vowels in the Peterson and Barney data). Table 1 shows the formant center frequency and bandwidth values for the five vowels being used. The actual values reported by Klatt varied slightly over the course of the syllable, but the stimuli for this study were synthesized using static filter positions based on Klatt's initial values (i.e. values from the beginning of the vowel) to ensure the validity of the mistuning manipulation.

Because the stimuli were based on observed vowel formant center frequency values, it was not possible to create them in such a way that the perceptual distance between vowels was equated. This should lead to some vowels being easier to identify than others as a result of being more perceptually distant from the other vowels in the set. The five vowels were selected with this consideration in mind, but it was not possible to eliminate the issue completely.

Thus, it was necessary to assess the degree to which the vowels' formant distances were perceptually asymmetric. The distance between formants was assessed using the mel scale, which is a perceptually weighted measure of frequency.[3] To generate a *perceptual distance* value, the average distance between the center frequencies of the first formant of a given vowel and the first formant of each of the other vowels in the set was calculated. This was repeated for the second formant, and the resulting two values were averaged. The average perceptual distance between each target vowel and the other vowels in the set is different. The

endpoint vowels had the largest perceptual distances (/ɑ/ = 252 and /i/ = 223), and the other three vowels had smaller perceptual distances (/I/ = 152, /æ/ = 140, and /ɛ/ = 129).

The stimuli were synthesized using a parallel formant synthesizer with independent source control. The source signal (i.e. the harmonics prior to filtering) was generated by a device that enables the slight mistuning of harmonics that are located near spectral peaks. This device, called *SourceBuilder* (Hall, Redpath, and Becker, 2011), was created as a virtual studio technology (VST) plugin in *MAX for Live*, which is an object-based software programming environment. The device consists of 100 sine-wave generators, and thus can generate a complex tone with up to 100 harmonics. Each sine-wave generator has independent frequency, intensity, and phase controls. The device includes the option of imposing a one- or two-pole low-pass filter on the signal to create a frequency roll-off of -6dB or -12dB/octave. It also has a 60th order low-pass filter, which serves to functionally prevent any information from passing above the frequency to which it is set. This which was set to 5.5 kHz, which is the maximum frequency observed from the speech of a typical female (Boersma and Weenink, 2010). This filter also served to eliminate aliasing, or high-frequency artifacts that are a result of digital sampling.

Another VST plug-in created in MAX for Live, called *Formant Function*, was used for filtering the source functions created by *SourceBuilder*. *Formant Function* consists of a parallel bank of six formant filters (two of which were used in this study) with independent center frequency and bandwidth controls. This device is capable of accepting a 10-second sample as a source (e.g. from *SourceBuilder*),

which enabled the creation of stimuli without looping. (The brief 250 ms stimuli for the current study were well within this upper limit). This is important because loop points with mistuned (and thus, generally out-of-phase) source material like that of the current investigation produce audible artifacts. In a purely harmonic signal this is not an issue because each harmonic completes a cycle at the period determined by the fundamental frequency, creating a point of correspondence at which the signal can be looped smoothly. However, if a harmonic is mistuned such that it is not an integer multiple of the F_0, it will lack these points of correspondence, and thus cannot be looped without introducing frequency distortion. This can be seen in Figure 4, where three in tune harmonics (shown in blue) converge on the right side of the figure at a point corresponding to one cycle of F_0, while a mistuned harmonic (shown in red) does not.

It has been shown that mistuned single harmonics in the first formant region of vowel sounds can still contribute to the vowel percept. At large degrees of mistuning (around 8%), the harmonic begins to be segregated from the vowel percept (Darwin and Gardner, 1986). Therefore, care was taken in stimulus development to ensure that mistuned components were perceptually integrated with the syllable.

Five levels of mistuning were be used for each vowel. Mistuning will distort the conveyance of spectral information without changing the number of harmonics present. For each formant, two harmonics will be mistuned (i.e. four total mistuned harmonics) to varying degrees toward and away from formant peaks: no mistuning (0%), 2% away from peak (-2%), 5% away from peak (-5%), 2% towards the peak

(2%), and 5% towards the peak (5%). The harmonics must be mistuned in the same direction. This is because mistuning adjacent harmonics 5% in opposite directions would result in a 10% relative mistuning of the two harmonics. The two harmonics that were mistuned for each formant were the harmonic closest to the center frequency and the next closest harmonic on the same side of the center frequency. This ensured that mistuning the harmonics in the same direction in absolute terms (i.e. higher or lower in frequency) resulted in them moving in the same direction relative to the formant center frequency. Mistuning towards the formant peaks was expected to facilitate accurate vowel identification, while mistuning away from peaks was expected to reduce identification accuracy.

Figure 5 shows a conceptual diagram of the mistuning manipulation of one harmonic towards a formant peak. The top half of the figure represents the position of the second harmonic in the absence of mistuning. This results in the first formant peak not being well represented in the output (i.e., resultant) spectrum. The bottom half of Figure 4 represents the result of the mistuning manipulation. Relative to the tuned example, the first formant in the bottom half of the figure is better represented, as the peak is higher in amplitude in the output spectrum.

In addition, four noise conditions were created for each vowel. The addition of noise had the effect of "filling in" the spectrum. White noise was added as a percentage of the overall intensity of the stimulus. The noise conditions were 0% noise, 2% noise, 5% noise, and 10% noise. The addition of noise was hypothesized to facilitate identification, as noise should help convey the formant peaks. Figure 6 shows conceptual diagram of the noise manipulation. From the figure it can be seen

that the noise fills in the areas between harmonics, and thus helps to convey formant peaks that do not correspond to harmonic locations.

For each vowel, the levels of mistuning were orthogonally combined with the levels of added noise. Thus, there were a total of 100 stimuli (5 vowels x 5 mistuning levels x 4 noise levels). Stimuli were presented over circumaural earphones (Sennheiser 25SP) in a single-walled sound attenuated chamber. All stimuli were equated for loudness and presented at a peak intensity of 80 dB[A].

Procedure

Before the experiment began, informed consent was obtained from the participants. They then completed a vowel identification task. The stimuli were presented by a PC using E-Prime (v. 2). Participants were instructed to respond to each stimulus using an E-Prime Serial Response Box with 5 buttons, each corresponding to one of the 5 vowels being used. The buttons will be labeled: "b<u>ea</u>d", "b<u>i</u>d", "b<u>e</u>d", "b<u>a</u>d", and "b<u>o</u>d".

When the experiment began, participants were presented with the stimuli one at a time in random order. The task was self-paced with a 500 ms inter-trial interval after each response before the next stimulus was presented. The experiment consisted of five blocks of 200 trials each. Within each block of trials, each stimulus was presented twice. Between blocks of trials, participants had the opportunity to take a break. Blocks of trials were completed in approximately 10 minutes, and the total running time of the experiment (not including breaks) was approximately 50 minutes, depending on the participant's rate of response.

Results

It was hypothesized that manipulations that helped convey formant peaks would lead to faster and more accurate vowel identification. Optimal performance was expected in conditions where noise was present and the harmonics were mistuned toward formant peaks. Further, an interaction between mistuning and noise was expected. It was anticipated that high levels of noise would facilitate vowel identification when harmonics were mistuned away from formant peaks, but would have a negligible effect when harmonics were mistuned toward formant peaks due to perceptual masking. Each of these hypotheses was evaluated separately for accuracy and response time data, which are summarized below.

Accuracy Data

The raw data consisted of the number of incorrect responses each participant gave in response to each stimulus. These data were then converted to the probability of a correct response for a given stimulus. A listing of the means and standard errors for the accuracy of responses to each stimulus can be found in Appendix A. Two participants' data were excluded from analysis based on their overall accuracy rates, which were 22% and 23 %. There were 5 response options, so 20% accuracy represented chance performance, and these participants' accuracy rates were not significantly different from chance. This left data from a total of 15 participants for analysis.

A 5 x 5 x 4 repeated-measures factorial ANOVA was conducted to assess the impacts of vowel (/i/, /I/, /ɛ/, /æ/, /ɑ/), mistuning (-5%, -2%, 0%, +2%, +5%), and noise (0%, 2%, 5%, 10%) on response accuracy. Mean accuracy scores by vowel can

be found in Table 2, and a graph of accuracy by vowel can be found in Figure 7. As can be seen in the table and figure, there were observed differences in accuracy as a function of vowel condition. Mauchly's test indicated that the assumption of sphericity had been violated for this variable ($\chi^2(9) = 30.6$, $p < .001$), therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.595$). After this correction, a main effect of vowel condition was observed, $F(2.38, 21.81) = 16.3$, $p < .001$, $\eta_p^2 = .539$.

Post-hoc analyses using Tukey's HSD to analyze differences in accuracy by vowel condition revealed large differences between the vowels. Accuracy for /æ/ (as in "bad") was significantly lower than accuracy for each of the other vowels, all p-values ≤ .008. Conversely, accuracy for /i/ (as in "bead") was significantly higher than accuracy for any of the other vowels, all p-values < .016. Accuracy for /ɑ/ (as in "bod"), in addition to being lower than /i/ and higher than /æ/, was marginally higher than accuracy for /I/ (as in "bid"), $p = .055$, and marginally higher than accuracy for /ɛ/ (as in "bed), $p = .098$. Accuracy for /ɛ/ and /I/ did not significantly differ, $p = .193$.

Mistuning also had an impact on accuracy rates. Means by mistuning manipulation can be found in Table 3, and a graphical representation of accuracy by mistuning can be found in Figure 8. Mauchly's test indicated that sphericity also had been violated for degree of mistuning ($\chi^2(9) = 19.1$, $p = .026$), so the degrees of freedom were adjusted using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.537$). After this correction, a main effect of mistuning was observed, $F(2.15, 30.1) = 9.48$, $p = .001$, $\eta_p^2 = .404$. Post-hoc analyses using Tukey's HSD for the mistuning

manipulation revealed that accuracy in the 5% mistuned away from formant peak condition (i.e. -5%) was lower than accuracy in each of the other mistuning conditions, all *p*-values < .04. No other significant differences in accuracy were observed, all *p*-values > .068.

Additionally, a significant interaction was observed between vowel condition and level of mistuning, $F(6.05, 84.7) = 2.27$, $p = .044$, $\eta_p^2 = .140$. This interaction was further assessed with a simple main effects analysis, which consisted of five separate two-way (mistuning x noise) repeated-measures factorial ANOVAs, one for each vowel. This enabled the effects of the mistuning manipulation to be examined separately at each level of the vowel condition. A Bonferroni adjustment[4] was made to the critical alpha-level to control for familywise type I error rate inflation as a result of running multiple ANOVA analyses (Lehman, 1995). To compute the *F*-statistic for tests of simple main effects, the Mean Squared Error (MSE) term from the initial analysis (i.e. the overall three-way ANOVA) was substituted for the MSE term from each follow up ANOVA. This is because the initial analysis contains more data points from which to estimate the population error variance than does each subgroup analysis (i.e. follow up ANOVA), so using the overall MSE typically leads to a more accurate error term (e.g., see Oshima and McCarty, 2000; Keppel, 1991).

The interaction between vowel condition and degree of mistuning was due to response accuracy being affected by mistuning differentially across the different vowel conditions. For three of the vowels (/æ/, /i/, and /ɑ/) no significant effects of mistuning were observed, all $F \leq 1.65$, $p \geq .18$, $\eta_p^2 \leq .105$. For /ɛ/, the mistuning manipulation did impact accuracy rates. Mauchly's test indicated that sphericity had

been violated ($\chi^2(9) = 27.5$, $p = .001$), so the degrees of freedom were adjusted using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.530$). The corrected F-test using the adjusted degrees of freedom and MSE term revealed a main effect of mistuning, $F(2.12, 29.7) = 10.487$, $p < .001$, $\eta_p^2 = .353$. This effect was significant even when using the more stringent, Bonferroni-adjusted alpha-level (.0083).

The significant main effect of mistuning for /ɛ/ was followed up with a post-hoc analysis using Tukey's HSD. [5] This analysis revealed that accuracy in the -5% mistuned condition was significantly lower than accuracy in the -2%, 0%, and +2% mistuned conditions (all p-values ≤ .04), and marginally lower than the +5% mistuned condition, $p = .078$. In addition to being significantly higher than the -5% condition, accuracy in the no mistuning condition was significantly higher than accuracy in the +2%, and +5% mistuning conditions, p-values ≤ .028.

A marginally significant effect of mistuning was observed for /I/. Mauchly's test again indicated that sphericity had been violated ($\chi^2(9) = 17.1$, $p = .049$), so the degrees of freedom were again adjusted using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.647$). The corrected F-test using the adjusted degrees of freedom and the MSE term from the original analysis revealed a main effect of mistuning, $F(2.59, 36.2) = 3.538$, $p = .024$, $\eta_p^2 = .211$. This effect was marginally significant when using the more stringent alpha-level. A post-hoc analysis using Tukey's HSD revealed that accuracy in the -5% condition was significantly lower than in the -2%, +2%, and +5% conditions (all p-values ≤ .016), and marginally lower than accuracy in the no mistuning condition, $p = .091$. No other significant differences were observed, all p-values ≥ .159.

The predicted main effect of noise was not significant, $F(3, 42) = 1.494$, $p = .230$, $\eta_p^2 = .096$. All other main effects and interactions were non-significant, all $F \leq 1.49$, $p \geq .165$, $\eta_p^2 \leq .096$.

Response Time Data

Response time data were also analyzed. Response times that were more than three standard deviations above a given participant's mean response time were excluded from further analysis, as were response times that were less than 150 ms. Presumably, if a response time was more than three standard deviations above a participant's mean, it was due to some factor outside the task (e.g. the participant thought they responded but hadn't pressed the button hard enough to register a response). Response times of less than 150 ms were excluded because they do not allow time for processing of the signal to occur prior to responding. The mean reaction time for college-age individuals in a simple auditory detection task (i.e. "push the button as soon as you hear something") is approximately 160 ms (Galton, 1899; Welford, 1980). Thus, reactions times of less than 150 ms almost certainly are not accurate reflections of task performance, and likely represent accidental button pushing. Response time exclusions for all participants totaled 310 trials out of 15,000 (or about 2%). While individual response times for some trials were excluded, no participants were excluded from the analysis based on their response time data. Thus, the same 15 participants' response time data were analyzed. A listing of the means and standard errors of the response times for each stimulus can be found in Appendix B.

Analysis of the response time data was similar to that of the accuracy data. Initially, a 5 x 5 x 4 repeated-measures factorial ANOVA was conducted to assess the impacts of vowel, mistuning, and noise on response time. The predicted effects were not observed. The effect of mistuning was marginally significant, $F(4, 56) = 2.168$, $p = .084$, $\eta_p^2 = .134$. Similarly, there was a marginally significant effect of vowel condition, $F(4, 56) = 2.199$, $p = .081$, $\eta_p^2 = .136$. Also, the predicted effect of noise was not observed, $F(3, 42) = 1.898$, $p = 0.145$, $\eta_p^2 = .119$.

One effect that was observed was an interaction between vowel condition and level of mistuning, $F(16, 224) = 3.106$, $p < .001$, $\eta_p^2 = .182$. This interaction was further analyzed with simple effects analysis, consisting of five separate two-way (mistuning x noise) repeated measures factorial ANOVAs, one for each vowel. This enabled the effects of the mistuning manipulation to be examined separately for each vowel. As with the accuracy data, a Bonferroni adjustment was made to the critical alpha-level to control for familywise type I error rate inflation, and the Mean Squared Error (MSE) term from the initial analysis was substituted for the MSE term from each follow-up ANOVA.

The interaction between vowel condition and degree of mistuning was due to the fact that degree of mistuning exerted an effect on response time for only one of the vowels: /ɛ/. The other four vowels (/æ/, /I/, /i/, and /ɑ/) revealed no significant effects of mistuning, all $F \leq 1.71$, $p \geq .178$, $\eta_p^2 \leq .236$. For /ɛ/, Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(9) = 21.9$, $p < .01$), therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.659$). After this correction, a main effect of mistuning

was observed, $F(2.637, 36.913) = 6.93$, $p = .0028$, $\eta_p^2 = .240$. This was significant at the more stringent Bonferroni-adjusted critical alpha-level (.0083). Tukey's HSD post hoc analyses revealed that accuracy in the -5% mistuned condition was significantly lower than accuracy in the -2% and +2% mistuned conditions (all p-values ≤ .009), and marginally lower than the 0% and+5% mistuned conditions, *p*-values = .066 and .054, respectively. No other significant differences were observed, all $p \geq .101$.

Discussion

Manipulations that helped convey the position of formant peaks were expected to facilitate performance. Mistuning towards formant peaks was expected to lead to faster and more accurate responses, while mistuning away from formant peaks was expected to result in slower and less accurate responses. Similarly, adding noise was expected to facilitate performance as it would "fill in" the spectrum and reinforce the perception of formant location. It was further hypothesized that there would be an interaction effect of combining the noise and mistuning manipulations. When harmonics were mistuned away from formant peaks, it was expected that noise would compensate for the lack of harmonic information near the peak and the negative effects of the mistuning manipulation on accuracy would be reduced. When harmonics were mistuned toward formant peaks, it was expected that the contribution of the noise to formant conveyance would be reduced by perceptual masking as a result of the noise sharing frequency space with the mistuned harmonics.

Effect of mistuning

The mistuning manipulation impacted vowel identification accuracy, although its effects on response time were less clear. The observation that responses were significantly less accurate in the -5% mistuning condition is consistent with the hypothesized effect of mistuning away from formant peaks (i.e. that it should cause formants to not be conveyed as clearly and lead to a decrease in accuracy rates). Also, consistent with the hypothesized effect, though not

statistically significant, is the observation that mean response time in the -5% mistuned was greater than in any other condition.

However, accuracy rates across the other levels of the mistuning manipulation did not follow the expected trend. Specifically, in addition to accuracy rates being lowest in the -5% mistuned condition, it was hypothesized that accuracy would increase as harmonics moved closer to formant peaks. This effect was not observed. Furthermore, the main effect of mistuning was qualified by an interaction with vowel condition. This interaction was such that the effects of mistuning were not observed across all five vowels. Reduced accuracy rates in the -5% condition were robust for /ɛ/, marginal for /I/, and were not observed for the remaining three vowels.

There is a potential formant-based explanation for the observation that there was no effect of mistuning for three of the vowels. Such an explanation is consistent with the broad hypothesis that formants are the focal points of vowel perception. It is likely that effects of mistuning were not observed for the endpoint vowels, /i/ as in “bead” and /ɑ/ as in “bod” (see Figure 3), largely because of the perceptual distance of their formant center frequencies from those of the other vowels in the set. These vowels had the highest average distances; 223 and 252 mels, respectively (see Table 4). As a result, even if their formants were not well conveyed (e.g. in the -5% mistuning condition), they were unlikely to be confused with other vowels in the set. This led to very high accuracy rates (95% and 87%, respectively), and left little room for the mistuning and noise to have an impact. Thus, due to

characteristics inherent in the vowels themselves (i.e., formant distances), /i/ and /ɑ/ were easily identified regardless of manipulation.

This interpretation is supported by anecdotal evidence that many of the errant responses to these two vowels were due to participants pressing the wrong button rather than being unable to identify the vowel. Specifically, several participants reported in their debriefing that /i/ and /ɑ/ were the easiest to identify and that they knew they had gotten some wrong because they pressed the wrong button in response to these vowels. By contrast, participants typically reported feeling less confident about identifying the other vowels.

Formant-based stimulus characteristics offer a potential explanation for the failure to observe an effect of mistuning for /æ/, as in "bad", as well. Performance in this vowel condition was much poorer with overall accuracy at only 41%. It was frequently confused with /ɛ/, as in "bed". This is likely due to the close proximity of these two vowels' second formants. It has been reported that in the F2 region, the formant center-frequency discrimination threshold is approximately 1.5% (Kewley-Port and Watson, 1994). In the case of/æ/, the second formant center frequency was 1660 Hz, 1.5% of which is roughly 25 Hz. The difference in second formant center frequencies between /ɛ/ and /æ/ is only 20 Hz (see Table 1). Thus, for most participants, the second formants of these two vowels were likely indistinguishable, leaving the first formant as the only cue by which to make an identification judgment.

The first formant center frequency used for /ɛ/ in the current study is identical to the average first formant center frequency for /ɛ/ reported in Peterson

and Barney (1952) based on analysis of the vowel production of 76 talkers. Conversely, the first formant center frequency Peterson and Barney report for /æ/ is different from the value used in the current study by 40 Hz, which is nearly three times greater than the formant discrimination threshold in the F1 region of approximately 14 Hz (Kewley-Port and Watson, 1994). Thus, it should come as no surprise that responses tended to be more accurate for /ɛ/ than /æ/, and that identifying /æ/ was very difficult regardless of the noise and mistuning manipulations. As with /i/ and /ɑ/, it appears that the accuracy rate for /æ/ was largely determined by vowel characteristics, which left little room for the mistuning and noise manipulations to exert an effect.

Consistent with the idea that response accuracy was largely determined by the inherent formant-based characteristics of each vowel, a significant correlation was found between average formant distance and accuracy rates for each stimulus, $r(98) = .674$, $p < .0001$, $R^2 = .454$. A summary of average perceptual distance and accuracy for each vowel can be found in Table 4. The raw data that were used to compute the correlation can be found in Appendix A. This correlation revealed that stimuli with larger average perceptual distances tended to have higher accuracy rates. In fact, average perceptual distance alone accounts for nearly half of the total variance in response accuracy. This implies that the main effect of vowel condition can largely be attributed to average perceptual distance.

The fact that so much of the variance is explained not by the mistuning or noise manipulations, but by the inherent qualities of the vowels, was not anticipated. However, it is broadly consistent with the idea that the formants are the

focal points of speech perception. Specifically, it indicates that those vowels with more distinctive formants (i.e. those that are more perceptually distant from the formants of other vowels) are easier to label.

However, this formant-based explanation fails to account for observed trends in the accuracy data as a function of mistuning. While it predicts the failure to observe an effect of mistuning for some vowels, it also predicts that any vowel that is impacted by mistuning should exhibit the full range of mistuning effects (i.e., effects of mistuning both toward and away from formant peaks). However, there was a failure to demonstrate the hypothesized increase in accuracy as a function of harmonics being moved closer to formant peaks.

It could be argued that the lower accuracy rates in the -5% mistuned condition were simply a result of the perceptual segregation of the mistuned harmonics. In other words, the reduction in accuracy could be explained not by harmonics being moved away from formant peaks, but by them being perceptually removed from the signal altogether. However, if the reduced identification accuracy was a result of perceptual separation of the mistuned harmonics from the remainder of the signal, one would expect that accuracy rates in the +5% and -5% mistuned conditions to be identical. The degree of mistuning in these two conditions was the same and the manipulation was applied to the same harmonics in both cases. Thus, if the mistuned harmonics perceptually segregated and listeners were applying vowel labels based only on the remaining (tuned) portion of the signal, there should be no difference in accuracy between these two conditions because the tuned portion of the signal is identical in both cases. Therefore, the mistuned

harmonics must have been contributing to the vowel percepts at least to some degree. This leaves the effects of the mistuning manipulation on formant representation as the most likely explanation.

Alternatively, the absence of an increase in accuracy as harmonics moved toward formant peaks may be a result of the fact that vowels typically contain narrow formants. The narrowness of a formant is determined by the resonance of the filter used to create it. The filtering of the current stimuli was accomplished using the *Formant Function* device, whose filters contain a *Q* ("quality") parameter that determines their resonance. This parameter is proportional to the ratio of center frequency to bandwidth, so the Q-value can alternatively be thought of as a measure of formant narrowness.[6]

As can be seen in in Table 5, the average resonance (Q-value) of the filters used to generate the current stimuli is 22.65. This means that the formants of the current stimuli were relatively narrow. In fact, they were slightly more narrow than typical formants of natural vowels. For example, recordings of natural vowel tokens reported by Fant (1972) had an average Q-value of 18.90. Synthesized vowels used in a recent study of vowel identification had an average Q-value of 9.35 (de Cheveigne, 1999). The difference in narrowness between the current stimuli and natural tokens was especially pronounced for the first formants. The first formants of the current stimuli had an average Q-value of 15.19, whereas recordings reported by other researchers had average first formant Q-values of 10.24 and 9.64 (Fant, 1972; Fujimura and Lindquist, 1964).

In general, narrow formants should be easily detected because the signal (formant peak intensity) to noise (intensity of the surrounding signal) ratio is larger for narrow formants than for wide formants. This can be seen in Figure 9, in which the narrow formants of the synthesized vowel /i/ can clearly be identified. The ease with which narrow formants should be perceived offers a possible explanation for the lack of difference in accuracy rates across the -2%, 0%, +2%, and +5% mistuning manipulations. Because the formants were so prominent, their locations were easily perceived in the absence of mistuning. Therefore, moving harmonics towards formant peaks should yield no added benefit. Even when harmonics were mistuned 2% away from formant peaks (i.e. -2%), the formants were so prominent to begin with that they were still easily detected. Thus, it took the most extreme mistuning manipulation of -5% to have any impact on accuracy rates.

Additionally, the simplified two-formant spectrum should make formants locations easier to identify. This also can be understood in terms of a signal-to-noise ratio. The two formants of the current stimuli comprise the signal, and there is no noise in the form of other formants or spectral irregularities that would make identifying formants more difficult, whereas a typical natural speech signal is full of spectral irregularities. This can be seen by comparing Figure 9, which represents one of the current stimuli (/i/), with Figure 10, which is a natural token of the same vowel. In Figure 9, the spectrum is very smooth, and harmonics steadily decrease in intensity as they move away from formant peaks; all of the harmonics in the signal "point to" the formant peaks. This is not the case in Figure 10, where there is a substantially greater degree of spectral irregularity that could potentially obscure

the location of formant peaks. Minimally, there are less data points upon which to base perception of formant location. This difference is most clear in the second formant region

Given that the first two formants are typically sufficient for vowel identification, it could be argued that observed accuracy rates should have been higher if listeners really were able to accurately perceive formant peak locations in all but the most mistuned stimuli. However, while perceiving the location of two formants may be sufficient for vowel identification, two formant stimuli do not necessarily lead to maximum accuracy in identification. In fact, even three-formant synthesized vowels are identified significantly less accurately than natural tokens (Hillenbrand and Nearey, 1999).

Furthermore, the current stimuli were unnatural in several other ways, each of which should lead to decreases in accuracy. First, the stimuli consisted of isolated vowels, which tend to be more difficult to identify than vowels embedded in CVC syllables (e.g., see Strange, Edman, and Jenkins, 1979). Second, the stimuli had completely static spectra, which would be expected to lead to lower accuracy rates because dynamic cues are typically important in vowel perception (e.g., see Nearey and Assman, 1986; Strange, Jenkins, and Johnson 1983; Nearey 1989). Additionally, the vowels were selected based on proximity in F1/F2 space, which should serve to maximize confusability even when formant location is accurately perceived.

Another potential contributing factor is the fact that the vowels were based on tokens produced by a single speaker rather than the average values of many speakers. In fact, the formant center frequency values (reported by Klatt, 1980) that

were used to generate the current stimuli differed by an average of 93 Hz from the average center frequencies reported by Peterson and Barney (1952), based on a sample of 76 speakers. It should be noted that even under ideal conditions vowel identification rates are far from perfect. For example, Peterson and Barney (1952) report that accuracy rates in response to natural recordings of the same vowels used in the present study that were as low as 86% for /ɑ/.

It is possible that an effect of mistuning would have been observed for /i/, /ɑ/, and /æ/ if the mistuning manipulation was larger in magnitude. Given that narrow formants are easily located, this would most likely occur only in the conditions where harmonics were mistuned away from formant peaks. However, there are limitations to the magnitude of the mistuning manipulation. Even with the current maximum manipulation of 5% mistuning, one participant (a highly trained musician who plays five instruments) reported hearing the mistuned components completely segregate from the syllable percept.

Effect of noise

The hypothesized effect that adding noise to the signal would lead to higher accuracy rates and faster response times was not observed. The addition of noise was expected to have a larger impact in conditions where harmonics were mistuned away from formant peaks, because the noise would help convey the formant that the harmonics were no longer conveying, and a smaller impact when harmonics were mistuned towards formant peaks due to perceptual masking. This interaction was also not observed.

One possible explanation for the failure to observe an effect of noise on response accuracy is that the noise was unnatural and distracting insofar as the current stimuli contained substantially more noise than is found in a typical speech signal. The average intensity difference between harmonic and inharmonic portions of a typical speech signal is roughly 60 dB (Lively and Emanuel, 1970).[7] In the 2% noise condition, the current stimuli exceeded this value by more than 20 dB, with a harmonic to noise intensity difference of 35.6 dB (and 21.6 dB in the 10% noise condition). These measures are consistent with anecdotal reports that the stimuli sounded "weird", "really whispery", and "ghost-like". An explanation based on noise as a distraction would predict that as more noise was added, accuracy should decrease. However, there was actually a tendency for accuracy to increase by a small amount as more noise was added (see Table 6).

Another possibility is that the amount of noise added was insufficient to produce the hypothesized effect. In the most intense noise condition, the noise was only 10% of the signal's overall intensity, with harmonic information making up the remaining 90%. Although this is significantly more noisy than a typical speech signal, it may not have been sufficient to significantly impact performance. In the 10% noise condition, the peak intensity of the noise portion of the tone was an average of 21.6 dB below the peak intensity of the harmonic portion. It is certainly possible that this simply does not represent enough energy to make a strong contribution to the vowel percept. Such an explanation is consistent with other research finding that breathiness (i.e. noisiness) does not measurably impact speech intelligibility (e.g., Javkin, Hanson, and Kaun, 1991). However, whispered speech,

represents a signal consisting almost exclusively of noise that is intelligible (Tartter, 1989), so to some degree filtered noise must be capable of conveying phonemes. Therefore, it may be possible for the addition of noise to affect vowel identification at greater levels of noise-intensity.

The narrowness of the formants in the current stimuli also may have decreased the impact of the noise manipulation. The addition of noise was expected to help convey formant location. As a result, its effects could only be observed in cases where the location of a formant was unclear. If, as argued in the *Effects of Mistuning* section of this Discussion, the narrow formants were readily perceived in all but the -5% mistuned condition, the noise manipulation could only be expected to have an effect at this level of mistuning. Furthermore, this effect of mistuning was only observed for two of the vowels, which left only 2 of the 25 vowel-by-mistuning conditions in which the noise manipulation could reasonably be expected to have an effect (-5% mistuned for /ɛ/ and /I/). Although differences in performance across levels of noise were not significant, the addition of noise did tend to increase accuracy rates in these two conditions. Thus, in conditions where the noise manipulation could be expected to have an impact, it did.

Conclusions and Future Directions

The results of the current study offer some support for the hypothesis that source manipulations that impact formant clarity through interactions with the filter should impact phoneme perception. This support comes primarily from the -5% mistuning manipulation, which led to decreased accuracy rates overall. The results in this condition also offer some support to the hypothesis that density of

spectral sampling near formant peaks is important in addition to overall density of spectral sampling. Specifically, this is supported by the obtained result that accuracy changed as a function of mistuning while the total number of harmonics was held constant. In fact, these results suggest that it is possible that the results obtained by Ryalls and Lieberman (1982) are due specifically to changes in the number of harmonics that help convey each formant peak, and not due to changes in the overall density of harmonics.

However, while an overall effect of mistuning was observed, the full range of hypothesized effects was not. The failure to obtain all of the expected results may have been a result of unanticipated issues with the stimuli. In particular, it is possible that the formant-based characteristics of the set of vowels used and the narrowness of the formants prevented some of the effects of the experimental manipulations from being observed. Because these manipulations had not previously been conducted in a similar context, it was not possible to anticipate these issues beforehand. However, it should be possible to easily overcome these issues in the future.

The first issue was that the inherent formant-based characteristics of the vowels (formant center frequencies, average perceptual distance of formants) played a large role in determining accuracy rates, and left little room for the effects of the experimental manipulations. Two vowels were too easy to identify (/i/ and /ɑ/), and one vowel was too hard to identify (/æ/). Those vowels that fell somewhere between these extremes (/I/ and /ɛ/) tended to show effects of mistuning. Therefore, it should prove useful to replicate the current study with a

different set of vowels, selected to be of intermediate difficulty. It may only be possible to identify these vowels through pilot testing. Also, it is possible that including a larger set of vowels would lead to better results as a given vowel's difficulty may be dependent on the other vowels in the set. Careful selection of a larger set of vowels should make it possible to increase the proportion of the vowels for which the effects of source manipulations can be observed.

The second issue was that the formants themselves were too easily perceived, which left little room for the mistuning and noise manipulations to impact formant perception. This is likely a result of using a signal that consisted only of two narrow formants. One possible remedy would be to use a source with a higher fundamental frequency. This would result in there being fewer harmonics conveying each formant, which would make formant peaks more difficult to perceive. Additionally, having fewer harmonics conveying each formant peak would mean that the perceptual impact of mistuning a given number of them (in this case 2) should be larger. Furthermore, this increased difficulty in perceiving formants could enable the effects of the noise manipulation to be observed. Also, the larger gaps between harmonics that would result from the use of a higher fundamental frequency would leave more space for the noise manipulation to fill in and potentially impact performance.

A replication of the current procedure with stimuli modified as outlined above should allow a better understanding of the impact of source manipulations on speech perception. This would provide a stronger basis for drawing conclusions about the full range of mistuning and noise effects. Minimally, the current study

provides evidence that it is possible for speech perception to be affected by source manipulations, as the results suggest that mistuning harmonics in the source can impact vowel identification. Furthermore, the results reveal that mistuning can have this effect without the mistuned harmonics being perceptually segregated from the vowel percept, which had not been previously established.

Tables

Table 1

Formant center frequencies and bandwidths for the five vowels used in the current study, measured in Hertz (from Klatt, 1980)

Vowel	Formant Center Frequency (Hz)		Formant Bandwidth (Hz)	
	First Formant	Second Formant	First Formant	Second Formant
/i/ (“bead”)	310	2020	45	200
/I/ (“bid”)	400	1800	50	100
/ɛ/ (“bed”)	530	1680	60	90
/æ/ (“bad”)	620	1660	70	150
/ɑ/ (“bod”)	700	1220	130	70

Table 2

Mean probability of correct response (and standard error) by vowel, collapsed across mistuning and noise conditions

Vowel	Mean Accuracy (Standard Error)	95% Confidence Interval	
		Lower Bound	Upper Bound
/i/ ("bead")	.950 (.014)	.920	.981
/I/ ("bid")	.673 (.069)	.525	.822
/ɛ/ ("bed")	.771 (.036)	.694	.849
/æ/ ("bad")	.412 (.080)	.241	.583
/ɑ/ ("bod")	.866 (.042)	.775	.957

Table 3

Mean probability of correct response (with standard error) by mistuning manipulation, collapsed across vowel and noise conditions, along with 95% confidence intervals

Degree of Mistuning	Mean Accuracy (Standard Error)	95% Confidence Interval	
		Lower Bound	Upper Bound
-5%	.691 (.030)	.627	.756
-2%	.744 (.029)	.683	.806
0%	.756 (.030)	.692	.820
2%	.749 (.026)	.693	.804
5%	.733 (.030)	.668	.798

Table 4

Calculated "perceptual distance" of the formants of a given vowel from all other vowels in the set, with mean accuracy rate (and standard error) and mean response time (and standard error) for each vowel

Vowel		Average Formant Perceptual Distance	Mean Accuracy (Standard Error)	Mean Response Time (Standard Error)
/i/	("bead")	223	.950 (.014)	1257.21 (24.25)
/I/	("bid")	152	.673 (.069)	1316.73 (30.98)
/ɛ/	("bed")	129	.771 (.036)	1350.11 (29.68)
/æ/	("bad")	140	.412 (.080)	1280.62 (26.16)
/ɑ/	("bod")	252	.866 (.042)	1323.05 (26.63)

Table 5

Q ("quality") or resonance values resulting from the formant center frequencies and bandwidths of each formant for each vowel

Vowel		*Q* Value	
		First Formant	Second Formant
/i/	("bead")	13.78	20.20
/I/	("bid")	16.00	36.00
/ɛ/	("bed")	17.67	37.33
/æ/	("bad")	17.71	22.13
/ɑ/	("bod")	10.77	34.86
Mean value		15.19	30.10

Table 6

Mean probability of correct response (with standard error) by amount of noise, collapsed across vowel and mistuning manipulations, along with 95% confidence intervals

Amount of Noise	Mean Accuracy (Standard Error)	95% Confidence Interval	
		Lower Bound	Upper Bound
0%	.725 (.028)	.664	.786
2%	.732 (.031)	.666	.798
5%	.739 (.027)	.682	.797
10%	.742 (.028)	.682	.802

Figures

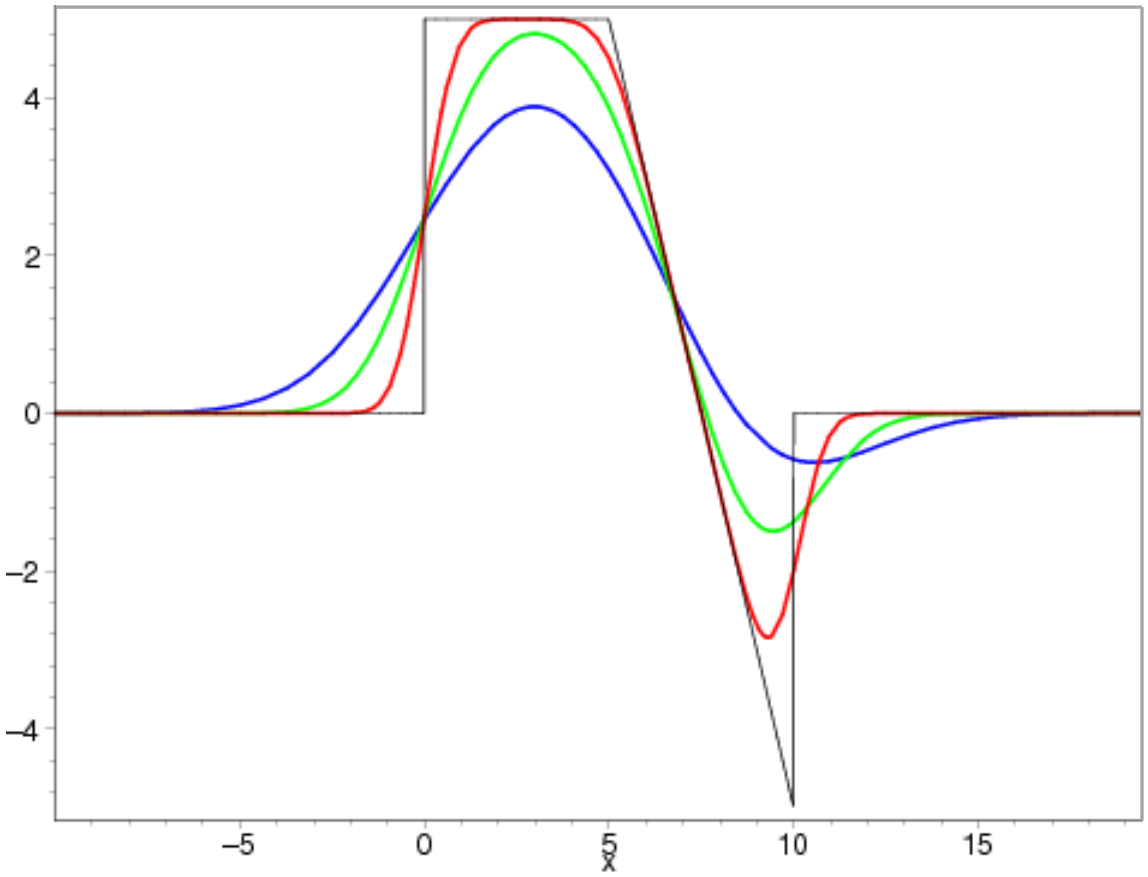

Figure 1. The graph of a function *f*(x) (grey) and its generalized Weierstrass transforms for t = 0.2 (red), t = 1 (green) and t = 3 (blue). The standard Weierstrass transform *f*(x) is given by the case t = 1, the green graph.

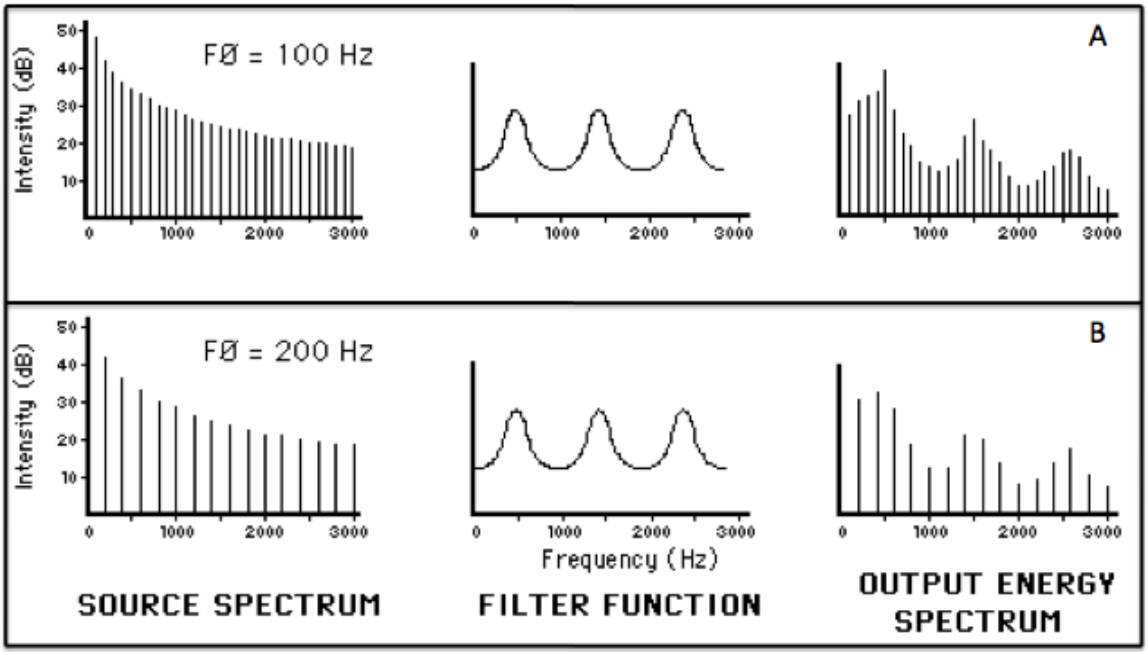

Figure 2: The source-filter model of speech production. The source spectrum represents the spectrum of typical glottal air flow with a fundamental frequency of 100 Hz. The filter, or transfer, function is for an idealized neutral vowel, with formant frequencies at approximately 500 Hz, 1500 Hz and 2500 Hz. The output energy spectrum shows the signal that would result if the filter function shown here was excited by the source spectrum shown at the left.

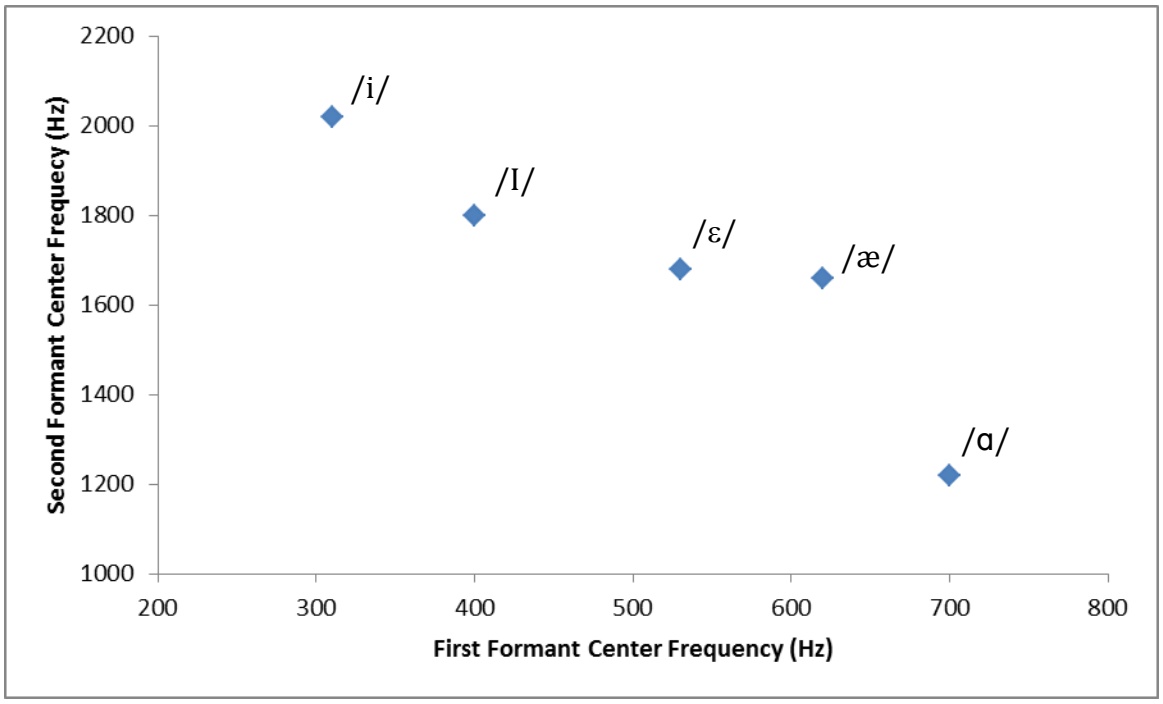

Figure 3. The relative position of the five vowels used in the current study in F1/F2 space.

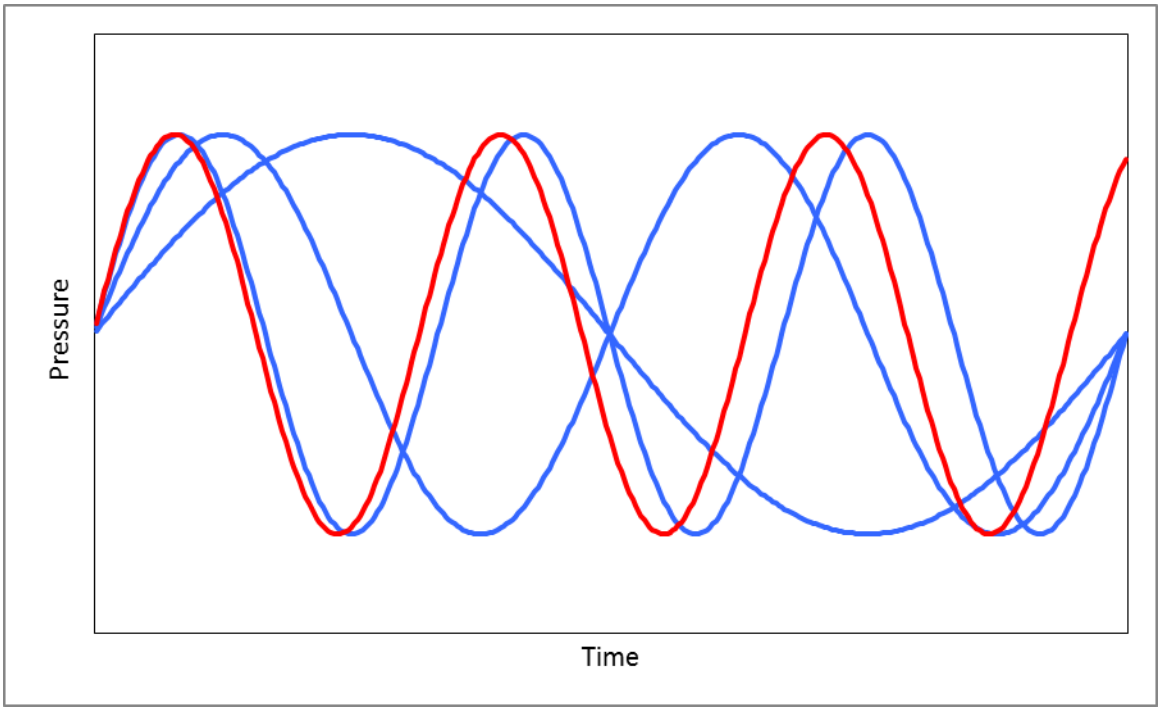

Figure 4. Graphic representation of four harmonics. The red line represents a mistuned harmonic, while the blue lines represent three tuned harmonics. The mistuned harmonic does not share the point of correspondence (i.e. it does not converge with the other harmonics at the far right of the figure). The frequencies shown are: $F_0 = x$; $H_2 = 2x$; $H_3 = 3x$; $H_4 = 3.17x$.

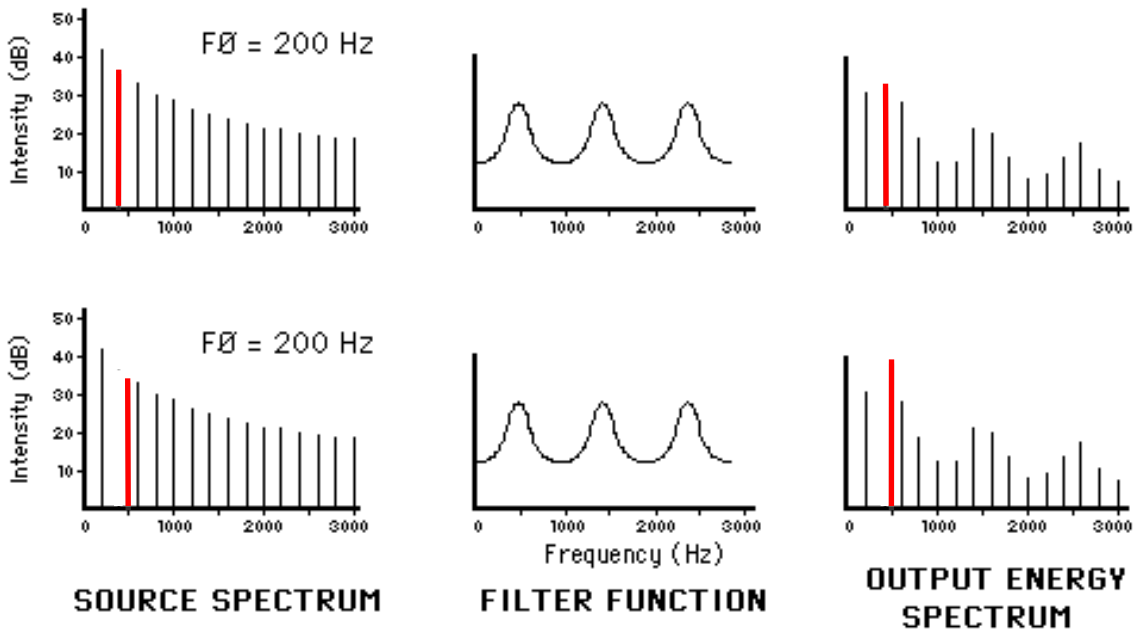

Figure 5. Conceptual representation of the mistuning manipulation. The red harmonic is the mistuned harmonic. The top half of the figure represents the spectrum before mistuning, and the bottom half represents the results of the mistuning manipulation.

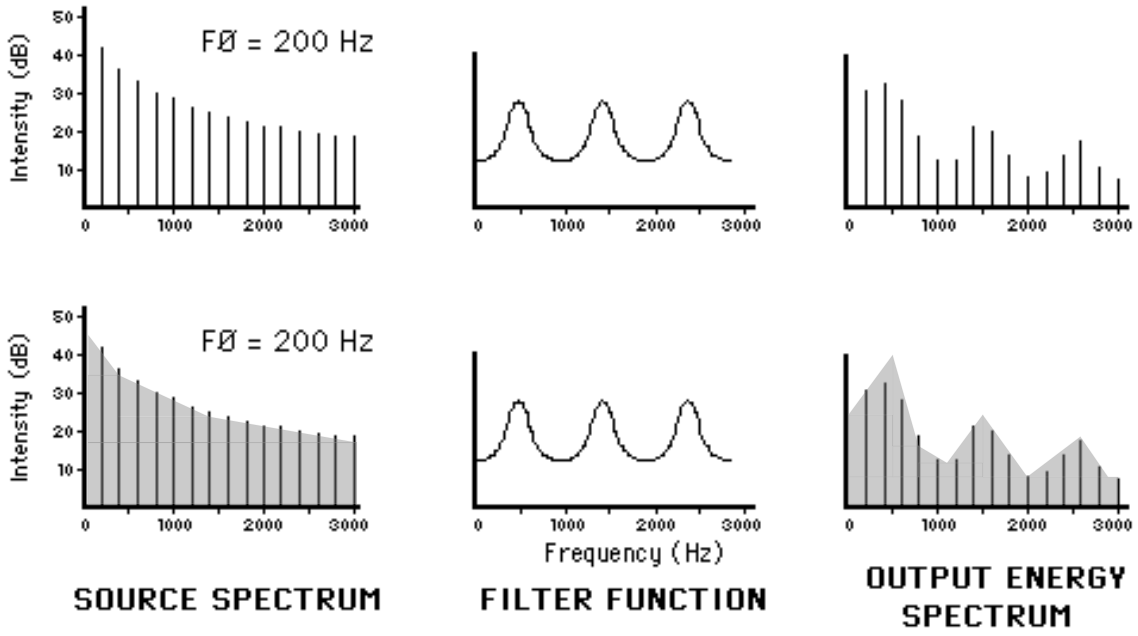

Figure 6. Conceptual representation of the noise manipulation. The noise is represented by grey shading. The top half of the figure represents the stimulus before the noise manipulation, and the bottom half after.

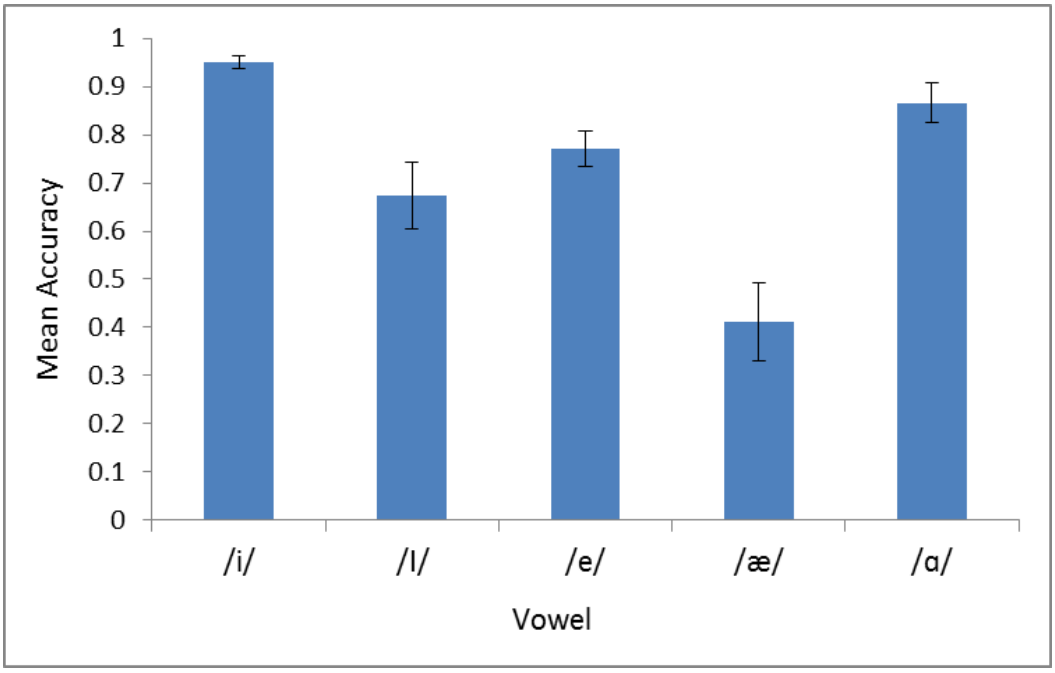

Figure 7. Response accuracy (with standard error bars) as a function of vowel condition.

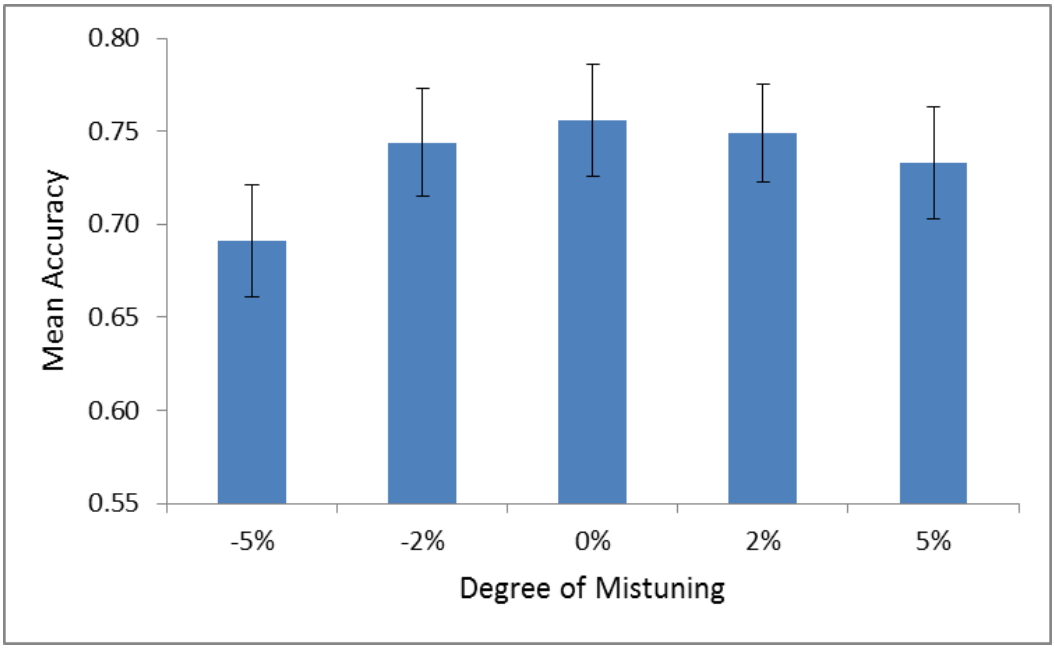

Figure 8. Response accuracy (with standard error bars) as a function of degree of mistuning.

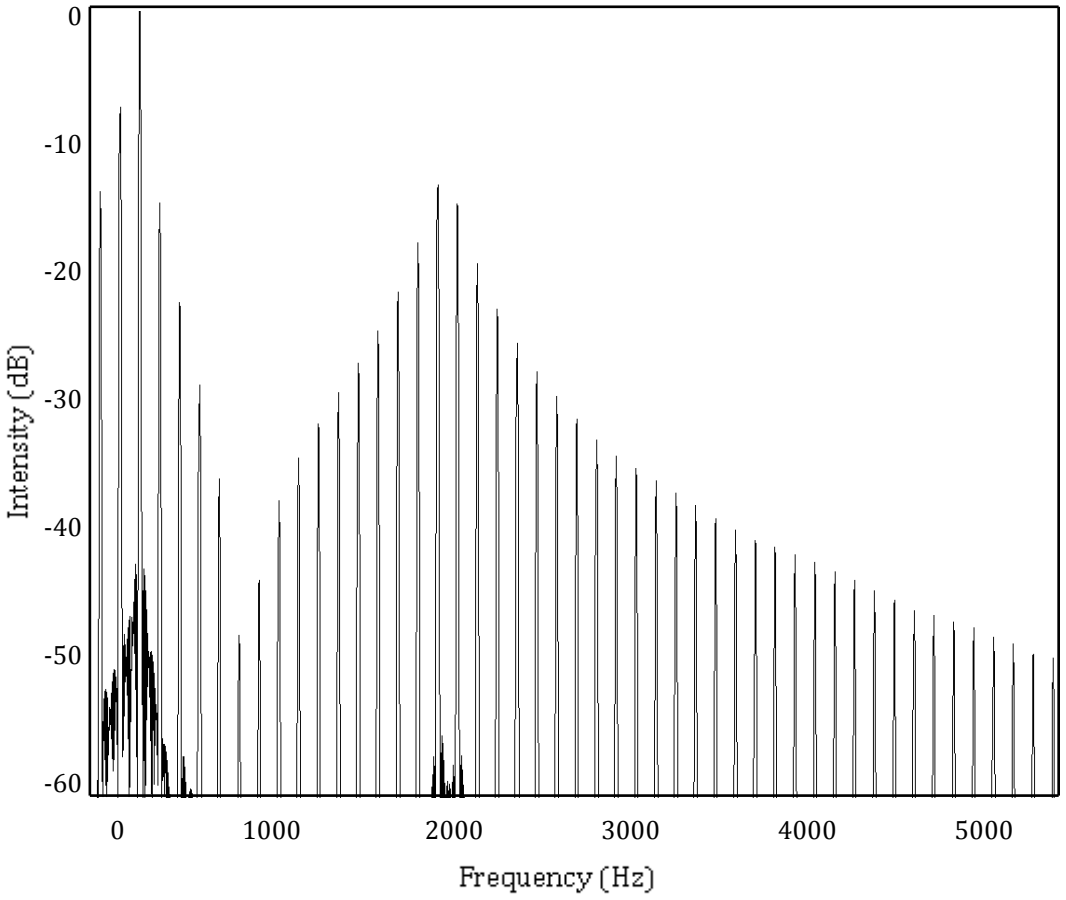

Figure 9. Spectral analysis of /i/ with no mistuning or noise. As a result of large Q values for both formants, they are relatively narrow. Spectral rolloff outside the bandwidth of the filters is nearly 30 dB/octave. Frequency is measured in Hertz, and intensity is measured in dB down from maximum.

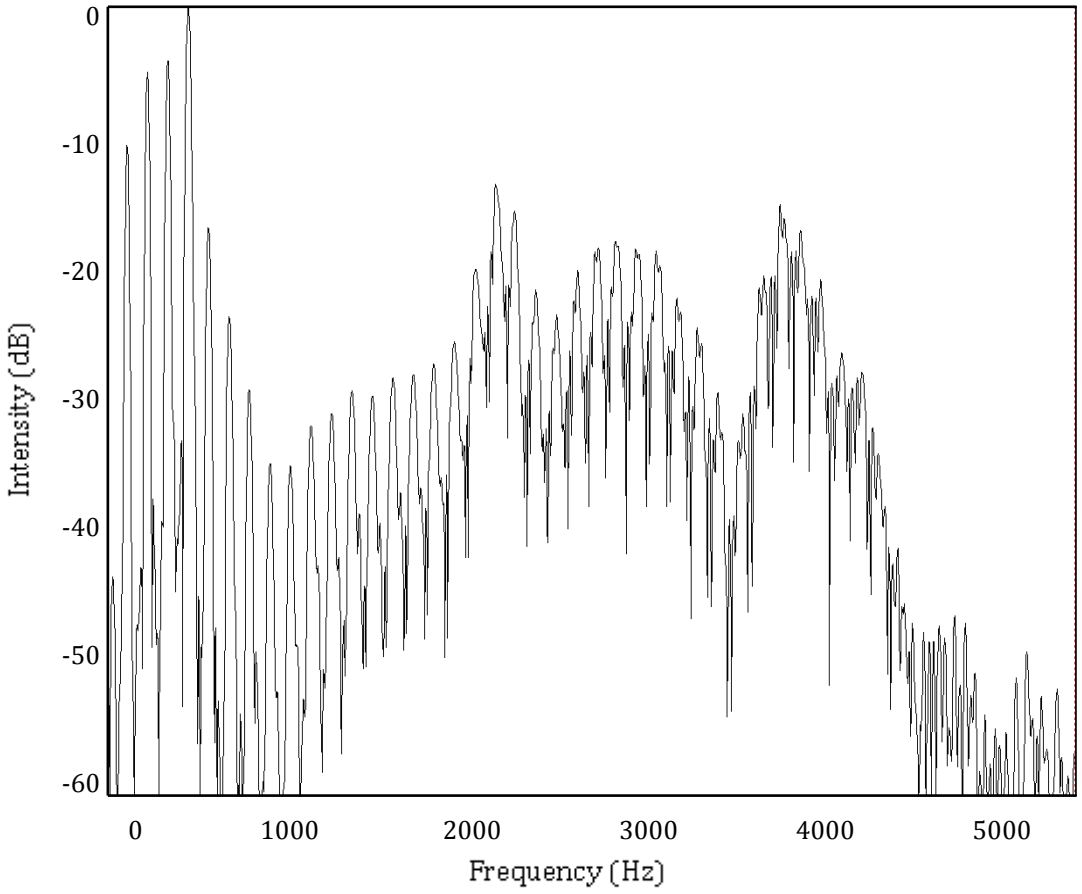

Figure 10. Spectral analysis of a natural /i/ token. Compared with the spectral shape shown in Figure 9, this represents a much more complex signal, with more spectral irregularities, and less clearly represented formants. Frequency is measured in Hertz, and intensity is measured in dB down from maximum.

Footnotes

1. Given the similarities in production between musical instruments and speech, it follows that there should be similarities in their perception. Like phoneme identification, instrument identification should be largely a product of filter properties as well. In fact, Brown (1999) found that a computer programmed to identify musical instruments using only spectral (i.e. filter) characteristics was able to identify a given instrument as well as musical expert human listeners. This finding is remarkable because the computer, in using only spectral information, was ignoring a great deal of information that was present in the signal. Also remarkable is the fact that the computer was given only one-minute samples of each instrument on which to base its identification.

Additionally, it has been shown that the primary acoustic dimension that relates to instrument identification in human listeners is spectral envelope shape (Hall and Beauchamp, 2009). Importantly, instrument identification was dependent on the location of individual formant peaks, or the position of individual filters, and was not dependent on overall spectral qualities such as spectral centroid. Spectral centroid is a measure of the center of the energy spectrum, calculated as the mean of all the harmonics present in the signal weighted by their intensities, and corresponds to the perceptual dimension of brightness. The position of individual filters should be important in speech as well, given the similarities between the two types of signal.

2. The researchers found that in a CV syllable, manipulating the length of the vowel had a significant impact on the perception of the consonant. The same

transition was perceived as either [ba] or [wa] depending on the length of the vowel. The authors conclude that the duration of the vowel serves as a rate normalization cue, and that consonant identification is dependent on the relative lengths of the two components. Furthermore, they found that manipulating the duration of a vowel in a second syllable (i.e. [bada] or [wada]) also had a significant effect on perception of the initial consonant.

3. The formula that was used to convert Hertz into mels was (O'Shaughnessy, 1987):

$$\text{Mels} = 2595 * \log_{10}\left(1 + \frac{\text{Hz}}{700}\right)$$

4. The Bonferroni correction consists of dividing the critical alpha level by the number of tests conducted; α/n. In this case there were a total of six ANOVAs run, so the critical α-level was adjusted by:

$$\alpha = \frac{.05}{6} = .00833$$

5. The formula for Tukey HSD pairwise comparisons contains a mean squared error (MSE) term that in this case is based on the subgroup. Thus, it could be argued that the values should be recalculated using the more appropriate error term from the original analysis (e.g., see Lehman, 1995). However, the assumption of sphericity was violated, so the benefits of using an error term other than the one associated with the particular analysis are unclear, and could potentially lead to a dramatic decrease in statistical power.

The formula for Tukey's HSD (honestly significant difference) test is:

$$\text{HSD} = q\sqrt{\left(\frac{\text{MSE}}{n}\right)}$$

where q is the studentized range statistic value for a given alpha and degrees of freedom, *MSE* is the mean squared error term for the factor being analyzed, and n is the number of observations per cell. The MSE value used here was that associated with each subgroup analysis (i.e. not the MSE term from the overall analysis).

6. The equation for calculating the resonance, or Q, is center frequency divided by bandwidth:

$$\text{Q} = \frac{\text{Center Frequency}}{\text{Bandwidth}}$$

A Q-value of $\text{Q} = \frac{1}{\sqrt{2}}$ (approximately .707) imposes the standard 6 dB per octave rolloff of a one-pole filter. Typical first formants of speech signals have Q-values on the order of 10, and second formants on the order of 25.

7. This value specifically applies to female speech. Females tend to have breathier, or noisier, voices than males, so the difference between the amount of noise in the current stimuli and that in a natural token would be larger for typical male speech. (The fundamental frequency and formant center frequencies of the current stimuli are characteristic of male speech.)

Appendix A

Mean accuracy (and standard error) for each stimulus. Marginal means for each condition are displayed along the bottom and right side of each table.

Table A1

Mean accuracy (and standard error) by degree of mistuning and amount of noise for all stimuli based on the vowel /i/, as in "bead"

Noise	Degree of Mistuning					Marginal Means
	-5%	-2%	0%	2%	5%	
0%	.95 (.03)	.93 (.03)	.93 (.04)	.95 (.03)	.96 (.02)	.94 (.01)
2%	.94 (.03)	.96 (.02)	.94 (.02)	.96 (.02)	.97 (.01)	.95 (.01)
5%	.97 (.02)	.94 (.03)	.96 (.02)	.96 (.02)	.94 (.02)	.95 (.01)
10%	.95 (.02)	.97 (.02)	.96 (.02)	.95 (.02)	.93 (.03)	.95 (.01)
Marginal Means	.95 (.1)	.94 (.1)	.94 (.01)	.96 (.01)	.96 (.01)	.95 (.01)

Table A2

Mean accuracy (and standard error) by degree of mistuning and amount of noise for all stimuli based on the vowel /I/, as in "bid"

Noise	Degree of Mistuning					Marginal Means
	-5%	-2%	0%	2%	5%	
0%	.59 (.08)	.69 (.07)	.64 (.08)	.70 (.06)	.65 (.07)	.65 (.03)
2%	.58 (.08)	.60 (.07)	.69 (.08)	.67 (.08)	.70 (.07)	.65 (.03)
5%	.62 (.08)	.72 (.08)	.68 (.08)	.71 (.07)	.72 (.08)	.69 (.03)
10%	.66 (.08)	.68 (.09)	.69 (.08)	.78 (.06)	.69 (.08)	.70 (.03)
Marginal Means	.60 (.04)	.67 (.04)	.67 (.04)	.70 (03)	.69 (.04)	.67 (.02)

Table A3

Mean accuracy (and standard error) by degree of mistuning and amount of noise for all stimuli based on the vowel /ɛ/ ("bed")

Noise	Degree of Mistuning					Marginal Means
	-5%	-2%	0%	2%	5%	
0%	.64 (.09)	.78 (.05)	.84 (.04)	.75 (.05)	.72 (.04)	.75 (.03)
2%	.63 (.07)	.85 (.04)	.87 (.04)	.74 (.05)	.81 (.04)	.78 (.02)
5%	.72 (.06)	.76 (.05)	.80 (.04)	.83 (.04)	.80 (.04)	.78 (.02)
10%	.71 (.05)	.83 (.04)	.84 (.04)	.77 (.05)	.72 (.05)	.78 (.02)
Marginal Means	.66 (.03)	.80 (.02)	.84 (.02)	.78 (.02)	.78 (.02)	.77 (.01)

Table A4

Mean accuracy (and standard error) by degree of mistuning and amount of noise for all stimuli based on the vowel /æ/, as in ("bad")

Noise	Degree of Mistuning					Marginal Means
	-5%	-2%	0%	2%	5%	
0%	.37 (.08)	.43(.09)	.45 (.09)	.45 (.09)	.45 (.08)	.43 (.04)
2%	.39 (.08)	.46 (.09)	.45 (.09)	.39 (.08)	.37 (.08)	.41 (.04)
5%	.36 (.08)	.38 (.09)	.45 (.09)	.47 (.08)	.35 (.08)	.40 (.04)
10%	.39 (.09)	.40 (.09)	.41 (.09)	.38 (.08)	.43 (.09)	.40 (.04)
Marginal Means	.37 (.04)	.42 (.04)	.45 (.04)	.44 (.04)	.39 (.04)	.41 (.02)

Table A5

Mean accuracy (and standard error) by degree of mistuning and amount of noise for all stimuli based on the vowel /a/, as in ("bod")

Noise	Degree of Mistuning					Marginal Means
	-5%	-2%	0%	2%	5%	
0%	.76 (.07)	.88 (.04)	.88 (.05)	.85 (.05)	.87 (.06)	.85 (.02)
2%	.87 (.05)	.87 (.05)	.89 (.04)	.87 (.05)	.83 (.06)	.87 (.02)
5%	.83 (.06)	.87 (.04)	.87 (.04)	.89 (.04)	.88 (.06)	.87 (.02)
10%	.91 (.04)	.89 (.04)	.88 (.05)	.88 (.04)	.86 (.05)	.88 (.02)
Marginal Means	.82 (.03)	.87 (.02)	.88 (.02)	.87 (.02)	.86 (.03)	.87 (.01)

Appendix B

Mean response time (and standard error) for each stimulus. Marginal means for each condition are displayed along the bottom and right side of each table.

Table B1

Mean response time (and standard error) by degree of mistuning and amount of noise for all stimuli based on the vowel /i/, as in "bead"

Noise	Degree of Mistuning					Marginal Means
	-5%	-2%	0%	2%	5%	
0%	1300.36 (140.71)	1342.82 (134.46)	1268.92 (114.03)	1286.35 (116.21)	1274.22 (100.54)	1294.53 (52.51)
2%	1230.59 (102.85)	1280.31 (126.08)	1211.45 (99.49)	1276.13 (112.34)	1238.17 (122.43)	1247.33 (47.98)
5%	1406.20 (155.26)	1236.05 (110.34)	1099.27 (77.00)	1237.30 (117.98)	1209.28 (113.20)	1237.62 (49.35)
10%	1293.59 (132.20)	1225.10 (114.80)	1292.56 (136.19)	1243.53 (92.42)	1191.99 (97.59)	1249.35 47.55)
Marginal Means	1312.38 (63.28)	1286.39 (57.66)	1193.22 (53.60)	1266.59 (49.71)	1240.56 (52.46)	1257.21 (24.25)

Table B2

Mean response time (and standard error) by degree of mistuning and amount of noise for all stimuli based on the vowel /I/, as in "bid"

Noise	Degree of Mistuning					Marginal Means
	-5%	-2%	0%	2%	5%	
0%	1473.20 (132.65)	1384.35 (162.84)	1329.55 (139.32)	1387.62 (169.97)	1308.96 (142.96)	1376.73 (64.40)
2%	1280.07 (129.11)	1332.19 (142.93)	1215.72 (107.05)	1285.30 (131.25)	1312.67 (116.74)	1285.19 (55.68)
5%	1309.67 (146.40)	1371.07 (177.88)	1219.63 (135.05)	1301.52 (134.91)	1328.39 (143.79)	1306.05 (65.98)
10%	1375.75 (147.50)	1193.13 (112.19)	1239.87 (113.52)	1314.27 (148.62)	1371.59 (161.20)	1298.92 (60.24)
Marginal Means	1354.31 (74.40)	1362.54 (75.93)	1254.97 (62.18)	1324.82 (72.14)	1316.67 (70.11)	1316.73 (30.98)

Table B3

Mean response time (and standard error) by degree of mistuning and amount of noise for all stimuli based on the vowel /ɛ/, as in "bed"

Noise	Degree of Mistuning					Marginal Means
	-5%	-2%	0%	2%	5%	
0%	1615.25 (168.67)	1404.72 (161.15)	1338.74 (125.80)	1283.90 (153.27)	1421.36 (145.86)	1412.79 (67.40)
2%	1424.57 (151.34)	1223.35 (95.69)	1275.50 (104.28)	1274.08 (130.11)	1391.94 (167.36)	1317.89 (58.89)
5%	1430.79 (122.99)	1250.99 (97.52)	1399.74 (162.56)	1276.67 (113.41)	1342.37 (141.47)	1340.11 (54.79)
10%	1500.08 (148.78)	1214.66 (107.48)	1339.95 (128.54)	1248.51 (120.77)	1345.04 (123.65)	1329.65 (56.37)
Marginal Means	1490.20 (75.22)	1293.02 (57.65)	1337.99 (65.51)	1278.22 (64.63)	1385.23 (70.30)	1350.11 (29.68)

Table B4

Mean response time (and standard error) by degree of mistuning and amount of noise for all stimuli based on the vowel /æ/, as in "bad"

Noise	Degree of Mistuning					Marginal Means
	-5%	-2%	0%	2%	5%	
0%	1291.42 (126.36)	1276.63 (150.19)	1294.40 (109.78)	1260.69 (145.70)	1275.17 (132.57)	1279.66 (60.46)
2%	1293.39 (131.79)	1364.88 (157.54)	1204.75 (94.22)	1294.01 (128.58)	1331.51 (113.21)	1297.71 (56.47)
5%	1242.91 (101.13)	1172.98 (95.40)	1259.71 (105.11)	1318.22 (128.84)	1293.87 (73.53)	1257.54 (43.99)
10%	1211.40 (109.59)	1326.79 (120.48)	1365.11 (126.56)	1317.06 (108.45)	1217.59 (105.38)	1287.59 (49.69)
Marginal Means	1275.91 (58.23)	1271.50 (67.26)	1252.95 (54.95)	1290.97 (62.11)	1300.18 (53.71)	1280.62 (26.16)

Table B5

Mean response time (and standard error) by degree of mistuning and amount of noise for all stimuli based on the vowel /a/, as in "bod"

Noise	Degree of Mistuning					Marginal Means
	-5%	-2%	0%	2%	5%	
0%	1316.10 (119.48)	1271.72 (115.87)	1442.83 (138.38)	1257.36 (88.72)	1208.82 (101.94)	1299.37 (45.20)
2%	1372.20 (130.62)	1280.07 (124.28)	1390.41 (121.94)	1329.45 (145.77)	1338.29 (118.54)	1342.08 (55.11)
5%	1260.38 (125.68)	1332.89 (123.77)	1398.87 (139.59)	1354.84 (115.24)	1321.69 (128.00)	1333.73 (54.72)
10%	1392.73 (132.84)	1344.98 (160.26)	1294.01 (106.41)	1247.44 (114.35)	1305.86 (110.75)	1317.01 (54.57)
Marginal Means	1316.23 (63.32)	1294.89 (64.81)	1410.70 (59.95)	1313.88 (56.85)	1289.60 (55.00)	1323.05 (26.63)

References

Askenfelt , A. (1991). Voices and strings: Close cousins or not? In J. Sundberg, L. Nord, & R. Carlson (Eds.), Proceedings of Music, Language, Speech and Brain: An International Symposium. Wenner-Gren Center, Stockholm, 5-8 September, 1990.

Barreda, S. and Nearey, T. M. (2012). The direct an indirect roles of fundamental frequency in vowel perception. *Journal of the Acoustical Society of America,* 131(1), 466-477.

Boersma, P. and Weenink, D. (2010). Praat Program Manual. Retrieved from: http://www.fon.hum.uva.nl/praat/manual/Sound__To_Formant__burg____.html

Burns, E. M., & Ward, W. D. (1978). Categorical perception-phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals. *Journal of the Acoustical Society of America,* 63(2), 456-468.

Darwin, C. J. and Gardner, R. B. (1986). Mistuning a harmonic of a vowel: Grouping and phase effects on vowel quality. *Journal of the Acoustical Society of America,* 79(3), 838-845.

de Cheveigne, A. (1999). Formant bandwidth affects the identification of competing vowels. *International Conference on Phonetic Sciences, 1 ,* 2093-2096.

Delattre, Pierre C.; Liberman, Alvin M.; Cooper, Franklin S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27, 769-773

Delattre, P., Liberman, A. M., Cooper, F. S., and Gerstman, L. J. (1952). An

experimental study of the acoustical determinants of vowel colour. *Word*, 8, 195-210.

Dubno, J. R., and Dorman, M. F. (1987). Effects of spectral flattening on vowel identification. *Journal of the Acousticsl Society of America*, 82(5), 1503-1511.

Fant, G. (1960). Acoustic Theory of Speech Production. Mouton & Co, The Hague, Netherlands.

Fant, G. (1972). Vocal tract wall effects, losses, and bandwidths. *STL-QPSR 2-3/1972*, 28-52.

Fowler, C. A., & Rosenblum, L. D. (1990). Duplex perception: a comparison of monosyllables and slamming doors. *Journal of Experimental Psychology: Human Perception and Performance*, 16(4), 742-754.

Fry, D. B., Abramson, A. S., Eimas, P. D., and Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language and Speech*, 5, 171-189.

Fujimura, O., and Lindquist, J. (1964). Experiments on vocal tract transfer. *STL-QPSR 3/1964*, 1-7.

Galton, F. (1899). On instruments for (1) testing perception of differences of tint and for (2) determining reaction time. *Journal of the Anthropological Institute* ,19, 27-29.

Gottfried, T. L., Miller, J. L., and Payton, P. E. (1990). Effect of speaking rate on perception of vowels. *Phonetica* 47, 155-172.

Hall M. D. (2009). *Event perception based on minimal spectral information*. Presented

at the 2009 Auditory Perception, Cognition, and Action Meeting (APCAM 2009) Boston, MA.

Hall, M. D., & Pastore, R. E. (1992). Musical duplex perception: perception of figurally good chords with subliminal distinguishing tones. *Journal of Experimental Psychology: Human Perception and Performance*, 18(3), 752-762.

Hall, M. D., Redpath, T., and Becker, C. (2011). A formant-based synthesizer for psychoacoustic research within Max for Live. Presented at the 2011 Auditory Perception, Cognition, and Action Meeting (APCAM 2011) Seattle, WA.

Handel, S. (1993). *Listening: An Introduction to the Perception of Auditory Events.* MIT Press. Cambridge, MA.

Hermansky, H. (1987). Why is the formant frequency difference limen asymmetric? *Journal of the Acoustical Society of America,* 81(S1).

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

Holt, L. L., Lotto, A. J., and Kluender, K. R. (2000). Neighboring spectral content influences vowel identification. *Journal of the Acoustical Society of America*, 108(2), 710-722.

Javkin, H., Hanson, B., and Kaun, A. (1991). The effects of breathy voice on intelligibility. *Speech Communication*(10), 5-6, 539-543.

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

Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: voiced-

voiceless distinction in alveolar plosive consonants. *Science*, 190, 69-72.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd. ed.). Englewood Cliffs, NJ: Prentice Hall.

Kewley-Port, D., and Watson, C. S. (1994). Formant-frequency discrimination for isolated English vowels. *Journal of the Acoustical Society of America*, 95(1), 485-496.

Lehman, R. S. (1995). *Statistics in the behavioral sciences: A conceptual Introduction*. Pacific Grove, CA: Brooks/Cole Publishing Company.

Liberman, A. M. (1981). Duplex perception cues for stop consonants: Evidence for a Phonetic mode. *Perception and Psychophysics*, 30(2), 133-143.

Liberman, A. M., & Mattingly I. G. (1985). Motor theory of speech perception revised. *Cognition*, 21, 1-36.

Lively, M. A., Emanuel, F. W. (1970). Spectral noise levels and roughness severity ratings for normal and simulated rough vowels produced by adult females. *Journal of Speech and Hearing Research,* 13, 503-517.

Locke, S. & Kellar, L. (1973). Categorical perception in a nonlinguistic mode. *Cortex*, 9(4), 355-69.

Mattingly, I. G., and Studdert-Kennedy, M. (1991). Modularity and the Motor Theory of Speech Perception: Proceedings of a Conference to Honor Alvin M. Liberman. Psychology Press.

McClelland, J. L., and Elman J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1–86.

McMurray, B. and Spivey, M. (1999). The categorical perception of consonants: The

interaction of learning and processing. *Proceedings of the Chicago Linguistics Society* 35, 205-219

Miller, J. L., and Liberman, A. L. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception and Psychophysics* 25(6), 457-465.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America,* 85(5), 2114-2134.

Miller, J. D., Wier, C. C., Pastore, R. E., Kelly, W. J., and Dooling, R. J. (1976). Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America*, 60(2), 410-417.

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

Nearey, T. M. and Assman, P. (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, 80, 1297-1308.

Oden, G. C., and Massaro D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, 85 (1978), pp. 172–191.

Oshima, T. C. and McCarty, F (2000). *How should we teach follow-up tests after significant interaction in factorial analysis of variance?* Presented at American Educational Research Association, New Orleans, LA.

Peterson, G. E., and Barney, H. L. (1952). Control methods used in the study of vowels. *Journal of the Acoustical Society of America,* 24, 175-184.

Pols, L. C., van der Kamp, L. J., and Plomp, R. (1969). Perceptual and physical space of vowel sounds. *Journal of the Acoustical Society of America,* 46(2), 458-467.

Pols, L. C. W., van der Kamp, L. J. Th., and Plomp, R. (1969). Perceptual and physical space of vowel sounds. *Journal of the Acoustical Society of America, 46, 458-467.*

Rand, T. C. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, 55(3), 678-680.

Remez, R. E., Ferro, D. F., Wissig, S. C., & Landau, C. A. (2008). Asynchrony tolerance in the perceptual organization of speech. Psychonomic Bulletin & Review.

Remez, R. E., Pardo, J. S., Piorkowski, R. L., & Rubin, P. E. (2001). On the bistability of sine-wave analogues of speech. Psychological Science, 12, 24-29.

Remez, R., Rubin, P., Pisoni, D., and Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science* 212(4497), 947-950.

Ryalls. J. H. and Lieberman, P. (1982). Fundamental frequency and vowel perception. *Journal of the Acoustical Society of America,* 72(5), 1631-1633.

Siegel & Siegel (1977). Categorical perception of tonal intervals: Musicians can't tell *sharp* from *flat*. *Perception and Psychophysics*, 21(5), 399-407.

Strange, W., Edman, T. R., and Jenkins, J. L. (1979). Acoustic and phonological factors in vowel identification. *Journal of Experimental Psychology: Human Perception and Performance,* 5(4), 643-656.

Strange, W., Jenkins J. J., and Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, 74, 695-705.

Stevens, K. (1959). Effect of Duration upon Vowel Identification. *Journal of the Acoustical Society of America*, 31(1), 109-109.

Stevens, K. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America* 111(4), 1872-1891.

Studdert-Kennedy, M., Liberman, A. M., Harris, K. S., & Cooper, F. S. (1970). Motor theory of speech perception: A reply to Land's critical review. *Psychological Review*, 77(3), 234-249.

Tartter, V. C. (1989). What's in a whisper? *Journal of the Acoustical Society of America, 86(5), 1678-1683.*

ter Keurs, M., Festen, J. M., and Plomp, R. (1991). Effect of spectral envelope smearing on speech reception. *Journal of the Acoustical Society of America*, 91(5), 2872-2880.

Titze, I. R. (1994). Principles of voice production. Englewood Cliffs, N.J.: Prentice Hall.

Welford, A. T. (1980). Choice reaction time: Basic concepts. In A. T. Welford (Ed.), *Reaction Times.* Academic Press, New York, pp. 73-128.

Whalen, D. H., & Liberman, A. M. (1987). Speech perception takes precedence over nonspeech perception. *Science*, 237, 169-171.