Natural Correlations of Spectral Envelope and their Contribution to Auditory Scene Analysis

by

K. Jakob Patten

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved July 2017 by the
Graduate Supervisory Committee:

Michael K. McBeath, Chair
Eric L. Amazeen
Yi Zhou
Arthur Glenberg

ARIZONA STATE UNIVERSITY

December 2017

ABSTRACT

Auditory scene analysis (ASA) is the process through which listeners parse and organize their acoustic environment into relevant auditory objects. ASA functions by exploiting natural regularities in the structure of auditory information. The current study investigates spectral envelope and its contribution to the perception of changes in pitch and loudness. Experiment 1 constructs a perceptual continuum of twelve f_0- and intensity-matched vowel phonemes (i.e. a pure timbre manipulation) and reveals spectral envelope as a primary organizational dimension. The extremes of this dimension are *i* (as in “bee”) and *ʌ* (“bun”). Experiment 2 measures the strength of the relationship between produced f_0 and the previously observed phonetic-pitch continuum at three different levels of phonemic constraint. Scat performances and, to a lesser extent, recorded interviews were found to exhibit changes in accordance with the natural regularity; specifically, f_0 changes were correlated with the phoneme pitch-height continuum. The more constrained case of lyrical singing did not exhibit the natural regularity. Experiment 3 investigates participant ratings of pitch and loudness as stimuli vary in f_0, intensity, and the phonetic-pitch continuum. Psychophysical functions derived from the results reveal that moving from *i* to *ʌ* is equivalent to a .38 semitone decrease in f_0 and a .75 dB decrease in intensity. Experiment 4 examines the potentially functional aspect of the pitch, loudness, and spectral envelope relationship. Detection thresholds of stimuli in which all three dimensions change congruently (f_0 increase, intensity increase, *ʌ* to *i*) or incongruently (no f_0 change, intensity increase, *i* to *ʌ*) are compared using an objective version of the method of limits. Congruent changes did not provide a detection benefit over incongruent changes; however, when the contribution of phoneme change was removed, congruent changes did offer a slight detection benefit, as in previous research. While this relationship does not offer a detection benefit at threshold, there is a natural regularity for humans to produce phonemes at higher f_0s according to their relative position on the pitch height continuum. Likewise, humans have a bias to detect pitch and loudness changes in phoneme sweeps in accordance with the natural regularity.

TABLE OF CONTENTS

Page

CHAPTER

CHAPTER Page

APPENDIX

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Perception and parsing of an acoustic scene is the fundamental goal of the auditory system. An impressive aspect of audition is that the entire acoustic environment is sensed through only two point sources without the benefit of a continuous spatial array of information as in vision. Hermann von Helmholtz (1877/1954) likened auditory perception to identifying the sizes and distances of ships based on observation of ocean waves through a rolled-up newspaper. Despite the difficulties inherent in the task, humans and other animals are able to analyze the acoustic environment and parse individual sound sources quite well by means of auditory scene analysis (ASA).

Auditory Scene Analysis

ASA is the process through which organisms parse and organize their acoustic environment into auditory objects. This process is accomplished by exploiting natural patterns present in sounds (Bregman, 1984; Yost, 1992). Some of these patterns occur more regularly than others; some approach the ubiquity of Gibson's (1986) invariants, while others are more probabilistic. To carry von Helmholtz's metaphor further, the rate of change of wave size and the rate of change of wave direction is often related and can lead to perceptions of boat speed; slow moving boats will produce low rates of change in both domains while faster boats will produce faster changes. There could be an aquatic event that produces rapid wave size changes and no direction change, such as an earthquake, but these events are encountered far less often. Thus, the rates of wave size and direction change are natural regularities. According to ASA theory, perceptual biases arise from exposure to these natural patterns and can be used to infer the properties of sound and acoustic objects in a noisy environment. Biases arising from ubiquitous regularities are strong and difficult to overcome, while those arising from more probabilistic patterns are weak and may only be exhibited in certain contexts.

One natural pattern is harmonicity; sound emitting objects tend to produce overtones that are integer multiples of their fundamental frequency (f_0) (Johnston, 1989). Most acoustic objects produce noise through either string and node or air column architecture. Guitars and pianos exemplify the former; both produce sound from an oscillating string held in place at either end by two nodes. Flutes and the human voice function on the principle of an air column. In both instances, standing waves are produced. Based on the constraints of the system, the only possible overtones that can be produced by the initial standing (fundamental) wave must have nodes or antinodes in the same place. Figure 1 shows this physical constraint for string and node architecture. As such, a single object will produce overtones very close to integer multiples of its fundamental frequency and is incapable of producing drastically mistuned overtones. Thus, an array of harmonic frequencies is typically perceived as belonging to a single auditory object (Bregman, 1994).

Using natural regularities to parse an auditory scene and identify auditory objects can both predict and explain auditory biases and illusions, one of the main strengths of the ASA framework. If a series of several harmonics contains a single greatly mistuned harmonic, the perception changes from a single buzzing object to a buzzing object and a pure tone whistle corresponding to the mistuned harmonic (Alain, 2007). In the laboratory condition, this is a misperception; all overtones are produced by the same sound source. In the world, however, a highly harmonic object rarely – if ever – produces on greatly mistuned partial. Thus, the perception of two, independent sources is both probable and functional.

String and Node Harmonics

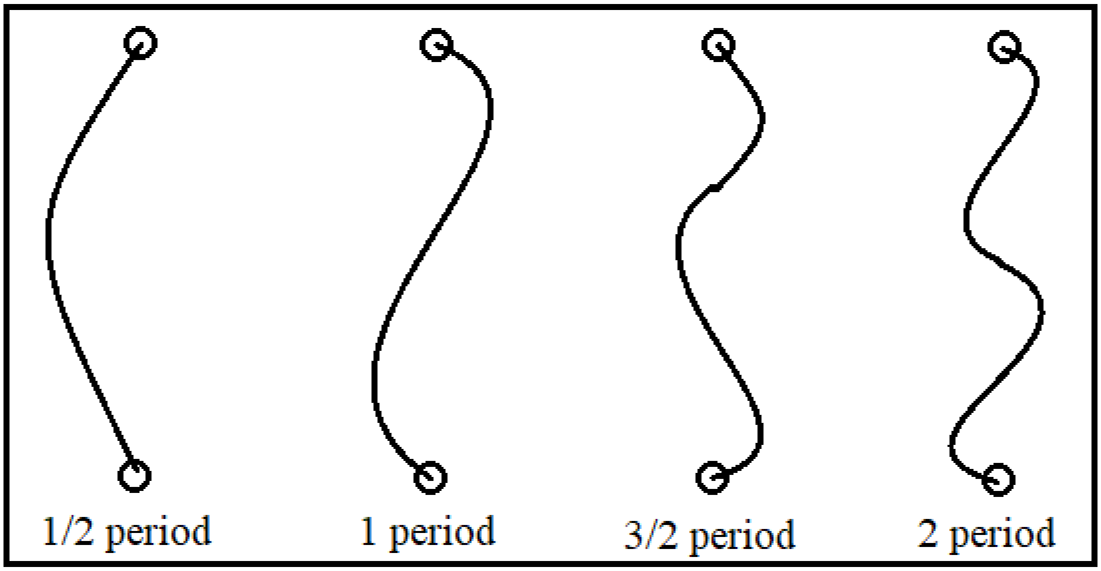


Figure 1. The lowest possible note for a string and two nodes and the three lowest overtone harmonics. Because each overtone must have a node in the wave where the physical node lies, each overtone is an integer multiple of the lowest note.

The Precedence Effect is another robust auditory bias that illustrates both the functional advantage of natural regularities and the non-veridical perception that can arise from them. The effect occurs when a resonant acoustic signal, such as speech, is emitted from two loudspeakers with one delayed from the other by less than 30 ms. The listener's percept is of a speech signal coming from only one source; the leading, undelayed speaker. If the delay is longer than 30 ms, the listener hears two identical, though delayed, messages coming from two different sources. The natural regularity of harmonicity (the delayed speech signal matches the original perfectly) is so strong that it overcomes the cue of directionality (Deutsch, 2009). While the described instance of the Precedence Effect essentially culminates in an incorrect localization or incorrect parsing of auditory objects, it often gives rise to a beneficial bias that leads to echo reduction; echoes typically reverberate in a similar acoustic pattern to the original and arrive at the listener's ear a short time later, the same stimulus pattern that gives rise to the described illusion. Thus, in typical cases – in fact, in almost every case a human might encounter outside of a laboratory – the location of the echo is ignored and the proper sound source is identified. It is important to note that the natural regularity often enhances accurate perception of the auditory scene and that ASA accounts for this illusion as part of effective perception based on known natural regularities.

ASA can account for and even predict auditory biases, but the relationship between biases and natural regularities runs both ways. Auditory biases of undetermined cause can be markers for as-yet-undiscovered natural regularities used in the process of ASA. It has been established that stimulus intensity can impact the perception of pitch. Stevens (1935) and, later, Gulick (1971) demonstrated that, for non-dynamic tones, pitch and loudness are positively correlated for tones above 2 kHz but become negatively correlated below 250 Hz. This effect is small, though that is not surprising from an ASA framework; non-dynamic tones do not occur often in a natural setting and, when they do, they do not offer natural regularities from which to form biases. As such, dynamic sounds are more ecologically valid and useful in identifying biases associated with natural regularities and ASA, such as the Doppler Illusion (Neuhoff & McBeath, 1996).

In the classic auditory Doppler Effect, a sound-emitting object travels past a stationary observer at constant velocity. The frequency of the auditory object decreases as it passes the observer. This effect is due to the constant speed of the object artificially, but consistently, compressing the sound waves in front of the object and artificially elongating the sound waves behind (Doppler, 1842). Intensity, however, increases as the object approaches the observer and decreases as it recedes. Despite the actual, physical acoustic characteristics, the human perception of pitch of such an auditory object does not match the f_0 received; participants experience an illusory increase in pitch as the auditory object draws near and an amplified decrease in pitch as the object recedes (Neuhoff & McBeath, 1996). In fact, participants were found to experience 8 semitones of illusory pitch drop when the physical change amounted to only 2 semitones; a full half-octave difference greater than the actual change (McBeath & Neuhoff, 2002). This illusion may stem from a strong correlation between intensity and frequency in the natural world. Thus, as an atypical stimulus, the Doppler Effect is perceptually distorted to correspond to more typical cases.

The correlation between pitch and loudness exists not only for the Doppler Illusion but for natural sounds, as well. When listening to animal vocalizations and making judgments of pitch change, participants often indicate a larger pitch change than the actual f_0 shift when the vocalization exhibits intensity changes. Similarly, when making judgements of loudness, participants often respond to pure f_0 changes as though they are both f_0 and intensity changes. This perceptual bias likely arises because animal vocalizations naturally exhibit simultaneous increases and decreases in f_0 and intensity (McBeath, 2014). Humans also tend to exhibit this correlation, unintentionally increasing the f_0 of their speech as they increase their vocal intensity. Without direction to control frequency, participants exhibited an approximately 8 semitone change from whispering a sentence to shouting a sentence in anger, a corresponding intensity increase of approximately 10 dB (Scharine & McBeath, 2009).

Timbre

The correlation between *f_0* and intensity has been documented as a natural auditory regularity in the areas of music, speech, animal vocalizations, and environmental sounds. Fundamental frequency and intensity are not only correlated with one another in naturally produced sound, however; dimensions of timbre also impact perception of one or both dimensions. For example, as tempo of a percussive beat increases, participants report the beat growing louder (Johnson, McBeath, & Patten, 2014). Tempo also impacts *f_0*; participants perceive voices with faster speaking rates as higher in pitch (Michalsky, 2016). Similarly, musical phrases with a quicker pace (more notes per measure) tend to be played at a higher pitch (Broze & Huron, 2013). While tempo is a dimension of timbre with well-documented interactions with both *f_0* and intensity, it is not a primary dimension of timbre.

Timbre is a messy domain; the sheer number of auditory features that fall into the realm of timbre have led to the domain being dubbed a “wastebasket category” for psychoacoustical aspects that are not pitch or loudness (McAdams & Bregman, 1979). Nevertheless, when testing 44 stimuli of different timbres, von Bismarck (1974a) noted that a majority of timbric distinctions fell along a dull-sharp continuum. Perceptions along the sharp-dull (or twangy-hollow) continuum are related to the shape of spectral envelope, specifically the average frequency or spectral centroid (Caclin, McAdams, Smith, & Winsberg, 2005; Erickson, 1975; von Bismarck, 1974b). It is known that changes in the spectral centroid impaired perception for discerning changes in *f_0*; for example, single semitone changes that are identifiable when played on a single instrument confuse participants when the instrument changes (say, from a piano to a trumpet) between notes (Allen & Oxenham, 2014; Krumhansl & Iverson, 1992; Luo & Soslowsky, 2017; Marozeau & de Cheveigne, 2007; Pitt, 1994). Interestingly, however, the specific influence of timbre changes on f_0 perception are less well documented. While some research notes that f_0 discrimination is not as diminished when the spectral centroid shifts in the same direction as the f_0, very few studies demonstrate that increases in the spectral centroid lead to perceptions of increasing f_0 (Allen & Oxenham, 2014; Luo & Soslowsky, 2017; Singh & Hirsh, 1992). Finally, the spectral centroid is not limited to influencing f_0 alone; sharper timbres are experienced as louder than hollow timbres (Melara & Marks, 1990).

When f_0 and intensity are held constant, vowel phonemes can also be characterized as changes in the spectral centroid. Both vowel phonemes and musical sounds are perceived to differ from their root sound in similar ways when adjusted in the same frequency bands (Slawson, 2005). Pitch perception of vowel phonemes is not tied to a single aspect of the spectral characteristic, but a combination of several. von Helmholtz (1877/1954) noted that the pitch of whispered vowels seemed to map largely onto the first formant of the speech spectrogram instead of the f_0, while later researchers suggest that pitch perception in speech depends on factors other than f_0 such as the second and higher formants (Carlson, Fant, & Granstrom, 1974; Higashikawa, Nakai, Sakakura, & Takahashi, 1996; Jacobsen, Schroger, & Alter, 2004; Meyer-Eppler, 1957; Thomas, 1969). Even pitch perception of non-speech stimuli is tied to other factors in addition to f_0. Perceptions of pitch are based both on height (which is roughly a function of average frequency energy) and chroma (which is largely determined by harmonic spacing that specifies diatonic notes); when participants make distinctions between musical pitch stimuli, they are often confused by changes in pitch chroma, a manipulation similar to altering the spectral centroid (Bachem, 1954; Dowling, 1978; Krumhansl, 1979; Krumhansl & Shepard, 1979). Meyer-Eppler (1957) noted that, aside from various speech formants, perceived pitch of American vowel phonemes is related to loudness, a finding that mirrors the pitch/intensity perceptions mentioned earlier.

Clearly, the literature converges on an as-yet unidentified three-way interaction between f_0, intensity, and timbre which may help explain the perception of pitch in spoken vowels. According to ASA theory, this bias must be driven by a natural regularity for speakers to produce vowels with different timbre at higher f_0s and intensities. The current study aims to further elucidate this interaction. Specifically, (1) the *i* phoneme will be perceived as higher in f_0 than the *Λ* phoneme even when both phonemes are presented at the same f_0 and intensity, and (2) the *i* phoneme will be physically produced at a higher f_0 than the *Λ* phoneme. Additionally, when using f_0 and intensity-controlled synthetic voicers, (3) phoneme sweeps from *i* to *Λ* will be perceived as falling in both pitch and loudness. Finally, (4) phoneme sweeps in which all three dimensions are congruent (i.e., *i* to *Λ*, falling f_0, and falling intensity) will be more easily identified in noise than incongruent changes.

CHAPTER 2

EXPERIMENT 1

When vowel sounds are presented at equivalent fundamental frequencies (f_0s) and intensities, they still have a differing quality of pitch to them (Carlson, Fant, & Granstrom, 1974). To discern if this aspect of pitch is a principal organizational dimension of timbre for stimuli with identical f_0s and intensities, a dimension reduction test procedure must be used. The current study uses multidimensional scaling (MDS) to discern the principal dimensions of timbre. MDS is a dimension reduction and data visualization technique wherein all stimuli are compared to one another and assessed for similarity between pairs of stimuli. The resulting matrix is then fit into an *n*-dimensional space where the distance between each stimulus is related to the rating of similarity; more similar items will cluster while disparate items are located far from one another. Dimensions that offer no substantial reduction in stress (an error-like term) are collapsed until only meaningful dimensions remain. MDS has been used to graphically represent the organizations of concepts (e.g., adjectives and animals) and perceptions (e.g., colors and tactile stimuli) (Shepard, 1980 & 1982). MDS solutions often produce functional results, such as returning a solution that resembles a geographical map or human skeleton when participants are asked to rate the similarity of US cities or body parts, respectively. This provides further support that perceptual dimensions typically reflect meaningful natural patterns.

The current experiment presents participants with pairs of North American vowel phonemes that have been equated for acoustic dimensions of f_0 and intensity and asks participants to rate the similarity of said phonemes with no guidelines for grouping. It is expected that the MDS procedure will produce a meaningful one- or two-dimensional solution, of which one of the dimensions will be pitch, as assessed by an independent rating procedure.

Method

Participants. 32 (18 female) undergraduate students (*M* age = 19.5, *SD* = 1.9) from an introductory psychology course at Arizona State University participated in this experiment in exchange for partial course credit.

Procedure. Participants are asked to gauge the similarity between 12 different North American vowel phonemes which were produced by a single human voice (International Phonetic Alphabet, 2016) and digitally adjusted to have a fundamental frequency (f_0) of 128 Hz (original f_0s ranged between 118 and 147) and an intensity of 60 dB SPL with the Audacity audio editing suite. The duration of each phoneme was normalized to 1 second.

Participants listened to all possible pairs of stimuli (66 unique pairs) separated by 1 second of silence and rated their similarity on a Likert-type scale where 10 indicated the exact same phoneme and 0 indicated an extreme difference. The resulting matrices were used to obtain a multidimensional scaling (MDS) solution.

Next, as part of a separate, independent procedure, participants were asked to arrange the same 12 phonemes from highest to lowest pitch. Using the Spatial Arrangement Method (SpAM) for MDS Java applet (Hout, Goldinger, & Ferguson, 2013), participants clicked and dragged IPA symbols of each phoneme in a two-dimensional digital space. Participants were instructed to not simply rank-order the stimuli, but to use distance to represent the magnitude of pitch differences.

Scree Plot for Experiment 1 Dimensionality

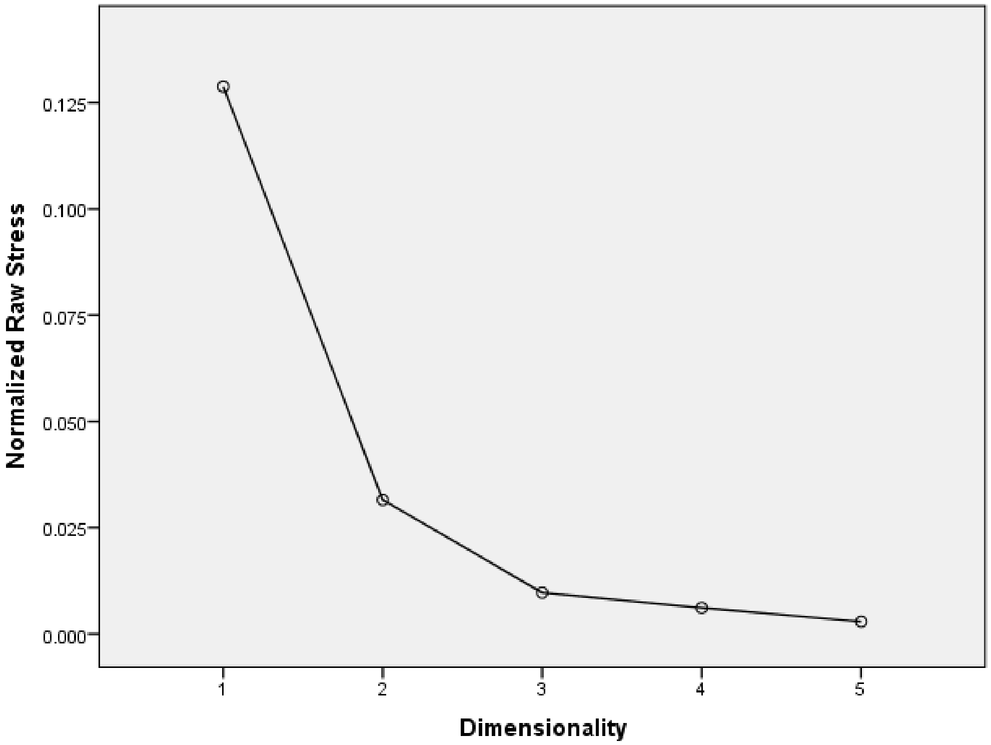


Figure 2. Scree Plot to determine the appropriate dimensionality for the solution in Experiment 1. Notice the apparent ‘elbow’ at the two-dimensional solution and a stress level below .1, indicating an appropriate reduction of S-stress.

Results

The Scree Plot of the MDS solution (see Figure 2) depicts an easily identifiable ‘elbow’ at the two-dimensional solution. Furthermore, Kruskal and Wish (1978) suggest that an optimal scaling solution should have an S-stress value less than .1; the S-stress of the two-dimensional solution is .026, while the one-dimensional solution is .129. Thus, the two-dimensional solution is in the ideal range for total stress and is the most appropriate characterization of the data.

The two-dimensional solution is shown in Figure 3, and has been to rotated to optimally correlate with participants' later pitch ratings of the vowel phonemes. The *y*-dimension of the rotated MDS solution is significantly correlated with rated pitch, *r* (10) = .90, *p* < .001. The *y*-coordinates of the two-dimensional solution are also correlated with the center frequency of the second formant of the phonemes, *r* (10) = .76, *p* < .05, but not the first formant. The *x*-coordinates of the MDS solution correlate marginally

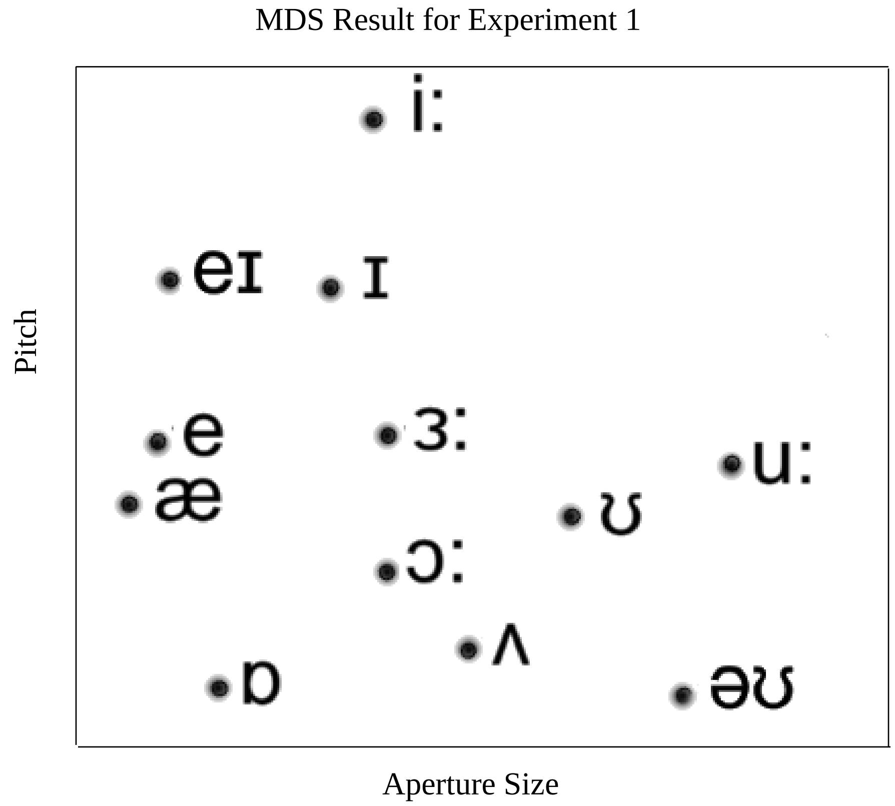


Figure 3. The two-dimensional multidimensional scaling (MDS) solution based on participants' ratings of North American vowel phoneme similarity. The *y*-coordinates of the phonemes significantly correlate with participants' ratings of perceived phoneme pitch. The *x*-coordinate marginally correlates with aperture size, as assessed from the IPA Vowel Chart (see Figure 4).

with the *y*-coordinates of the International Phonetics Association's (IPA) Vowel Chart, *r* (10) = .50, *p* = .09 (Figure 4). This chart shows how vowels are physically voiced in two production domains; aperture size and position of the tongue on the palette. Thus, the *x*-dimension of the MDS solution represents physical qualities of voicing.

Discussion

The *y*-dimension of the MDS solution correlates highly with rated pitch of the phonemes, suggesting participants not only perceive this pitch difference as salient, but use it to cognitively organize phonemes. Furthermore, the correlation between the *y*-coordinates and pitch ratings is larger than that between either the *y*-coordinates and the second formant or the *y*-coordinates and any dimension of the IPA chart. Thus, rated pitch is confirmed as a principal component used by participants to organize vowel phonemes that have been equalized for f_0 and intensity. While pitch is related to both the second speech formant and voicing position, there are other aspects of pitch that aren't captured in either of those measures. Pitch is likely related to the timbre differences (specifically, spectral envelope/pitch height) between phonemes. It is important to note that, while the second formant is derived from the spectral envelope, it does not contain all the information that contributes to pitch height.

International Phonetics Association Vowel Chart

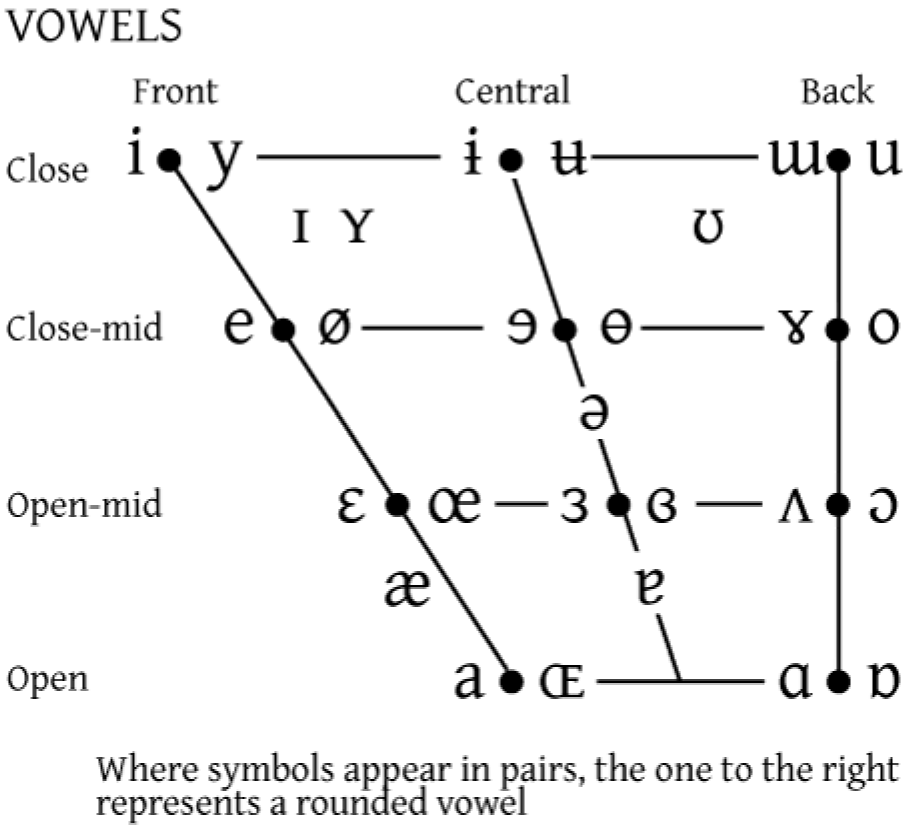


Figure 4. The International Phonetic Association Vowel Chart. Vowels are produced at different physical locations in the mouth along two dimensions, aperture size (*y*-dimension) and tongue position on palette (*x*-dimension).

CHAPTER 3

EXPERIMENT 2

The results of Experiment 1 confirm that there is a tendency to hear f_0-matched phonemes at different pitches. This perceptual bias may reflect a natural regularity of phoneme production. In addition to a perceptual interdependency between pitch and loudness, the physical acoustic dimensions of fundamental frequency and intensity have been shown to exhibit complimentary correlative changes during human production. In piano sheet music, higher frequency musical phrases are typically accompanied by more *forte* and *fortissimo* playing instructions, while lower frequency phrases are more likely to be accompanied by *piano* and *pianissimo* instructions. Similarly, notes sung or spoken at high frequencies are more often produced at more intense sound pressure levels than those voiced at low frequencies (Scharine & McBeath, 2009). Additionally, in laboratory settings, participants tend to produce vowels that are categorized as “high placement” at higher f_0s than their “low placement” counterparts. Specifically, the *i* phoneme is typically voiced at a higher f_0 than the *a* phoneme during both normal speech and singing (Fowler & Brown, 1997). Such high-placement vowels are voiced with higher f_0s across cultures, as well (Whalen & Levitt, 1995).

The results of Experiment 1 indicated that the *i* and *Λ* phonemes were the highest and lowest pitched exemplars, respectively. As such, they were chosen as stimuli in Experiment 2. The *I* phoneme (as in “bin”) is located along the center line that connects the *i* and *Λ* phonemes, and was chosen as a third exemplar. Additionally, the *I* phoneme is produced when a synthetic voicer morphs between the *i* and *Λ* phonemes (see Experiment 3). If there is a natural bias to produce phonemes at higher and lower f_0s in accordance to perceived pitch, this should be evident in unplanned human speech and unconstrained singing but not in constrained cases such as songs with phonemes and words determined by lyrics and melody.

To test this hypothesis, the first 30 utterances of each phoneme from six interview-based podcasts are analyzed. Speech in these normal settings, however, is partially constrained by social cues governing pitch and loudness. Scat singing is a type of vocal performance wherein the goal is to imitate the sounds of the natural world and other instruments, allowing the voice to be relatively unconstrained in production (Fredrickson, 2003; Miller, 2004). Therefore, any systematic differences in produced f_0s of vowel sounds should be even more pronounced in this type of vocal performance than in the unplanned speech condition. Conversely, standard lyrical music is typically composed first from either lyrics and melody. The other aspect (either melody or lyrics) is then fit to the existing component without considering the individual phonemes (Fredrickson, 2003). Because of this constrained nature, any inherent f_0 production differences between phonemes should be weak or absent entirely in songs with lyrics. The first 30 utterances of each phoneme from six scat and standard lyrical songs are also analyzed. It is expected that humans will voice the *i* phoneme with a significantly higher f_0 than the *I* and *Λ* phonemes, and that the *I* phoneme will be voiced with a higher f_0 than *Λ* in both scat performances and interview podcasts. Conversely, it is expected that there will be no f_0 differences between these phonemes in the standard lyrical songs. Finally, the scat song condition should produce the largest f_0 differences between these phonemes. Such findings would support the existence of an acoustic natural regularity in both songs and speech, confirming that phonemes that are generally experienced to be higher in pitch actually are systematically produced with higher f_0s.

Method

The first 30 instances (if available) of *i*, *I*, and *Λ* phonemes are analyzed for six recordings of scat vocal performances, performances of standard lyrical songs, and interview-based podcasts (see Appendix A for a full list), a total of 18 individual recordings. Scat and lyrical singing performances were chosen by searching YouTube and identifying performances that were 1) of high enough acoustic quality to analyze accurately, 2) mostly unaccompanied by music or accompanied by music at a low enough intensity that it would not produce inaccurate analysis results, and 3) long enough to contain approximately 30 instances of each phoneme. The interview podcast recordings were also pulled from YouTube with the only criteria being sufficient length to include approximately 30 instances of each phoneme. The Audacity (Mazzoni & Dannenberg, 2016) audio editing suite was used to analyze the fundamental frequency of each phoneme.

Results

A repeated measures ANOVA reveals an overall mean difference of fundamental frequency between *Λ* (*M* = 209.95 Hz), *I* (*M* = 250.90 Hz), and *i* (*M* = 261.45 Hz) phonemes [main effect of phoneme, F (2, 438) = 27.02, $p < .001$, $\eta^2 = .04$] and between interviews (*M* = 147.53 Hz), songs (*M* = 251.48 Hz), and scat (*M* = 323.29 Hz) [main effect of type of recording, F (2, 438) = 90.95, $p < .001$, $\eta^2 = .42$]. Additionally, there is a significant interaction between phoneme and recording type, F (4, 438) = 17.15, $p < .001$, $\eta^2 = .05$.

Pairwise comparisons reveal that phoneme f_0s in scat performances and interviews are produced in accordance with the phoneme-pitch continuum from Experiment 1. In scat performances, *i* (M = 387.6 Hz) is produced higher than *I* (M = 332.0 Hz), t (108) = 2.57, $p < .05$, $d = .49$, and *Λ* (M = 250.23 Hz), t (108) = 5.91, $p < .001$, $d = 1.18$, and *I* is produced higher than *Λ*, t (108) = 3.72, $p < .001$, $d = .71$. Similarly, the interview recordings follow the phoneme pattern from the MDS solution with *i* (M = 157.89 Hz) higher, though not significantly, than *I* (M = 153.53 Hz), t (108) = .41, p = *.n.s.*, $d = .08$, and *Λ* (M = 131.18 Hz), t (108) = 2.75, $p < .01$, $d = .53$, and *I* higher than *Λ*, t (108) = 2.50, $p < .05$, $d = .48$. There were no significant differences between phonemes in the standard lyrical song condition, as predicted. Figure 5 illustrates the differences between the f_0s produced for the three phonemes in semitones.

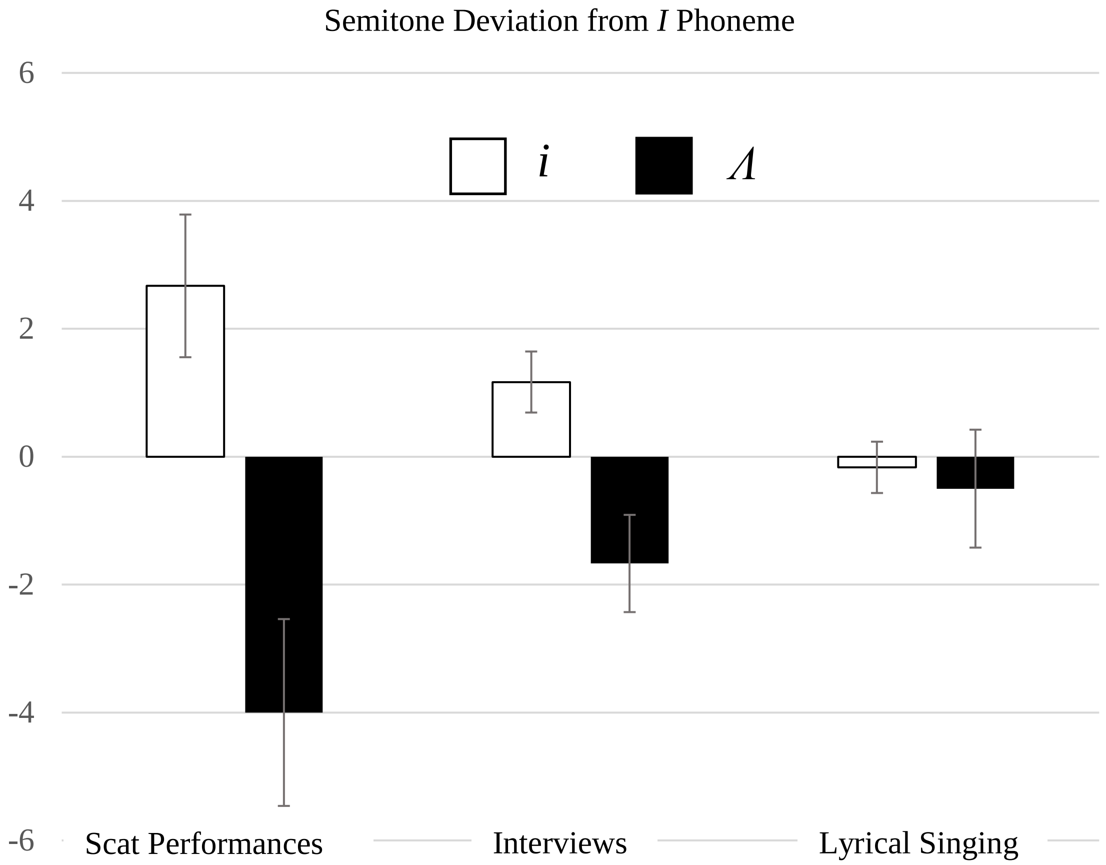


Figure 5. Deviations of *i* and *Λ* phonemes from *I* phoneme for scat performances, interviews, and standard lyrical singing. Error bars are 95% confidence intervals.

Discussion

The results confirm a bias to produce the phonemes that were previously rated as higher in pitch at higher f_0s when unconstrained. Furthermore, when unconstrained, the *i* and *I* phonemes were voiced at frequencies less disparate than *I* and *Λ*, which reflects the distance between those phonemes on the MDS solution from Experiment 1. Humans do not merely voice phoneme f_0s with rough adherence to the MDS solution; their perceptions of pitch and unconstrained phoneme production even match the ratios between their MDS ratings of the pitch height of phonemes. The bias to produce phonemes at f_0s that correspond to rated pitch is consistent with the existence of a regularity of the natural world in which timbre and fundamental frequency change in complimentary ways. This finding adds to and mirrors the documented correlation between fundamental frequency and intensity (McBeath & Neuhoff, 2002; Neuhoff & McBeath, 1996; Scharine & McBeath, 2009). When air pressure is increased in a brass or woodwind instrument, the resulting note also bends slightly to become sharper (Johnston, 1989). As demonstrated by the results of phoneme f_0s in the lyrical singing recordings, human voices do not always exhibit this correlation. It can be overcome at will when there are lyrical constraints. Thus, the appearance of a bias to perceive some f_0-matched phonemes to be higher in pitch during unplanned speaking and scat routines is a byproduct of perception that evolved to mirror natural regularities, and not merely a byproduct of the physics of resonators.

Experiment 2 confirms a correlation between the pitch height sub-dimension of timbre and fundamental frequency. While perceptual ratings from Experiment 1 provide a ranking of the pitch height of phonemes and production findings from Experiment 2 confirm the f_0, intensity, and phonetic correlations, Experiment 3 tests the strength of the coupling between pitch, loudness, and the pitch height sub-dimension of timbre, and measures the perceptual exchange rate between domains.

CHAPTER 4

EXPERIMENT 3

The perceptual relationship between timbre and fundamental frequency has the potential to be used in technologies such as sound source isolation and vocal compression. To accomplish this, the perceptual equivalencies between timbre (specifically, spectral envelope/pitch height), fundamental frequency, and intensity changes must be determined. When producing speech, humans unintentionally increase their f_0 by 3.5 semitones for every intentional 5 dB change in speech loudness (Scharine and McBeath, 2009). It is unclear, however, how much this produced change in one dimension is perceived in units of the other dimension. In perceptual measures, Neuhoff and McBeath (1996) noted a 6 semitone illusory pitch change for a coupled physical intensity change of 22 dB and 2 semitones over 1 second, a perceptual change of 4 semitones. This perceptual change is reduced to 1 semitone when the intensity change drops to 15 dB over a longer period of time. It is hypothesized that the current study will corroborate the 1 semitone illusory change noted by McBeath and Neuhoff and provide a similar psychophysical metric (within an order of magnitude) for the impact on rated pitch due to pitch height changes as a phoneme morphs from *i* to *Λ* or vice versa.

Method

Participants. 31 (17 female) undergraduate students (*M* age = 19.9, *SD* = 2.0) from introductory psychology courses at Arizona State University participated in this experiment in exchange for partial course credit. Two participants were removed from the analysis for failing to follow experimental procedure.

Apparatus. Only two phonemes – *i* and *Λ*, the two exhibiting the most extreme change and located in the same vertical plane of the MDS solution from Experiment 1 – were tested. Using the Praat audio editing suite (Boersma & Weenink, 2015), a synthetic voice was created that voiced 8 unique steps between *i* and *Λ* (one of which was a clear *I* phoneme, supporting the use of that stimulus in Experiment 2). These unique steps were then edited in Audacity (Mazzoni & Dannenberg, 2016) to create smooth glides between *i* and *Λ* that lasted 0.4 seconds. In total, there were four phonemic stimuli: a constant *i,* a constant *Λ,* an *i* to *Λ* glide, and an *Λ* to *i* glide. These stimuli were then digitally altered such that they exhibited f_0 and intensity changes. f_0 was either increased by 2 semitones, 1 semitone, remained unchanged, decreased by 1 semitone, or decreased by 2 semitones. Likewise, intensity was either increased by 20 dB, 10 dB, remained unchanged, decreased by 10 dB, or decreased by 20 dB. This procedure yielded a total of 25 changes per phonemic stimuli and 100 unique stimuli overall. All changes occurred in the same 0.4 seconds as the glide, with 0.8 seconds of constant stimuli on either side for a total stimulus duration of 2 seconds.

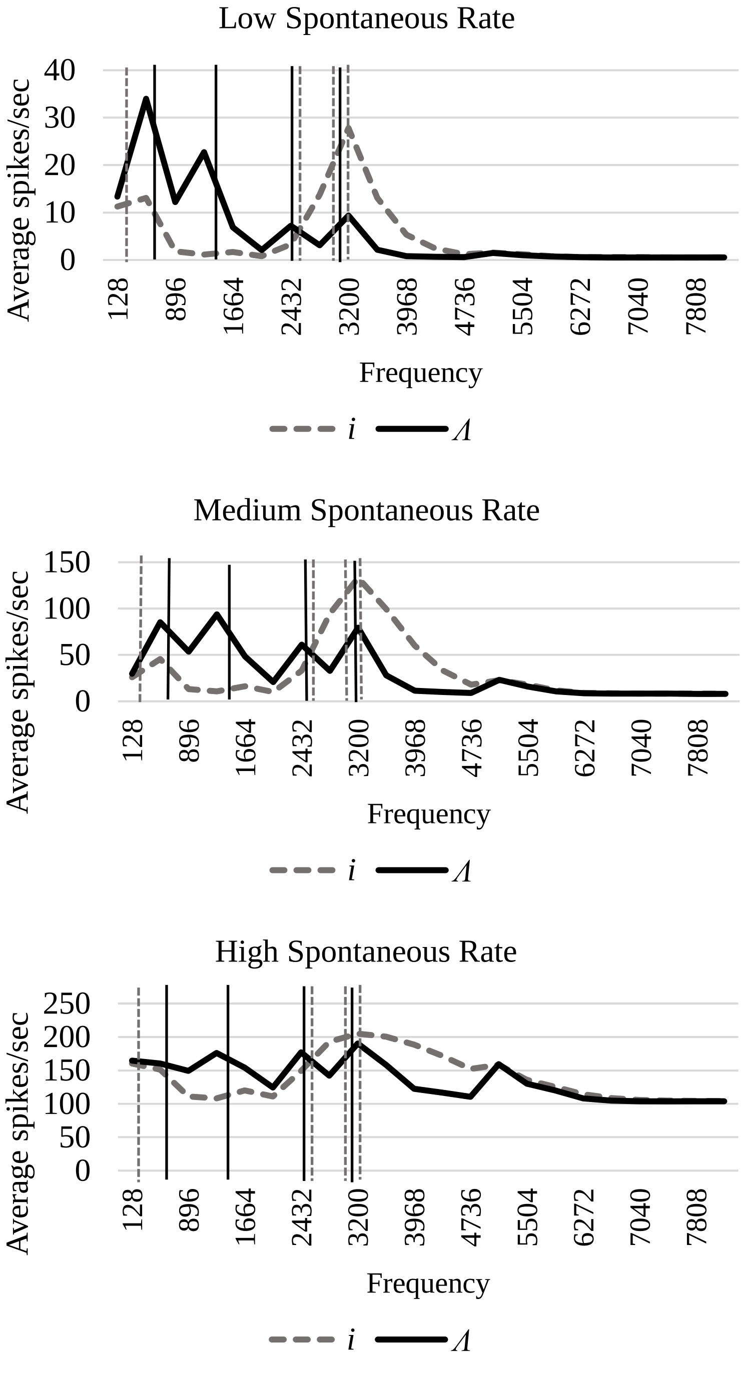
Pitch Matching. Though the *i* and *Λ* stimuli were digitally altered to have a f_0 of 128 Hz, overtones or other acoustic qualities could cause the stimuli to be perceptually matched to different frequencies. To ensure this was not the case, 11 additional participants (*M* age = 22.45, *SD* = 5.35) who were not a part of Experiment 3 were asked to match each phoneme to a pure tone sine wave in an adaptive staircase procedure. Participants were presented with a phoneme for 1 second, followed by a 1 second pure tone. Initial changes jumped by 20 Hz, followed by 10, 5, 2, and 1 Hz. Each participant matched a given phoneme twice, once from an initial frequency of 90, and once from

170. There was no significant difference between the matched frequencies of *i* (*M* = 126.23 Hz) and *ʌ* (*M* = 129.32), *t*(10) = .76, *p* = *ns*, *d'* = .26.

Neural Response Modeling. While the phonemes used in the experiment were equated for both intensity and f_0, it is possible that differences in the structure of the spectral envelope may lead to physical changes in the inner ear and auditory nerve that could alter perceptions of the sounds, especially in the domains of pitch and loudness. If present, these changes could mean that the perceptual bias to hear phonemes at different pitches when equated for the same f_0 is not a perceptual bias, but a physical change caused by the structure of the inner ear. To identify if this type of change was present, models of the auditory nerve response for the *i* and *ʌ* phoneme stimuli used throughout Experiments 1, 3, and 4 were constructed using a procedure published by Zilany and colleagues (Zilany M. S., Bruce, Nelson, & Carney, 2009; Zilany, Bruce, & Carney, 2014).

Neural models for *i* and *ʌ* were constructed such that they spanned the first four formants of both phonemes. The Zilany procedure models the response of a single characteristic frequency (CF). The current modeling procedure began at the shared *f0* for both phonemes (CF = 128 Hz) and rose in third octave increments to the final CF of 8192 Hz. This procedure was repeated at low, medium, and high rates of spontaneous neural firing. The resulting models (depicted in Figure 6) revealed excitation patterns that may explain some of the effects observed in Experiment 1. The *i* phoneme consistently produces neural spike peaks at higher frequencies than the *ʌ* phoneme. In all cases, these peaks occur close to the second formant of the *i* phoneme (2577 Hz) and coincide closely with the third and fourth formants (3062 and 3196 Hz). Because the auditory nerve

Auditory Nerve Response Models for *i* and *ʌ* at Different Spontaneous Rates



Low Spontaneous Rate

Medium Spontaneous Rate

High Spontaneous Rate

Figure 6. Auditory nerve response models for *i* and *ʌ* phonemes at three different levels of spontaneous nerve firing rate. Models span the first and second formants for both phonemes. Note the large increase in firing rate for the *i* phoneme around 3200 Hz. This bump is near the second formant for the *i* phoneme and also coincides with the third and fourth formants.

tuning curve produces higher gain for these frequencies, it is possible pitch ratings are correspondingly influenced. The *Λ* phoneme, conversely, has more dispersed formants (F2 = 1314 Hz, F3 = 2465 Hz, F4 = 3172 Hz) that do not receive the same degree of neural gain. The effects of the auditory nerve tuning curve are not clear in regard to the perception of intensity. The model indicates medium spontaneous rate fibers do produce a notably more intense peak for the *i* phoneme, though low and high fibers do not. Similarly, the *Λ* phoneme produces more intense spike rates throughout the frequency range.

Second formant contributions to perceptions of frequency change and vowel quality have been noted. The first two formants and the distance between them are extremely useful in both the identification and synthesis of vowels (Carlson, Fant, & Granstrom, 1974; Delattre, Liberman, Cooper, & Gerstman, 1952; Fujisaki & Kawashima, 1968). The second formant is also the sole determinant of pitch in whispered speech (Thomas, 1969). Higher formants, however, contribute to perception of both phoneme and frequency when the distance between the first two formants are approximately equal (Fujisaki & Kawashima, 1968). Furthermore, while it is established the second formant is involved in pitch ratings of vowel phonmes, the reason behind this remains undetermined. Vowels are often able to be pitch matched as the vowels used in the current study were (in the above subsection), so the major frequency component is not perceptually altered. Perhaps the link between the increased neural gain and the production bias observed in Experiment 2 explains this perception.

Procedure. Stimuli were presented to participants using the PsychoPy experimentation software (Peirce, 2007; 2009). Participants were randomly assigned to complete either the loudness or pitch portion first. The loudness portion began with a training session in which participants used a mouse cursor to represent changes in loudness. Red lines halfway between the center of the screen (where the cursor began each trial) and the top and bottom of the screen indicated a 20 dB change in a 128 Hz tone. Participants experienced seven trials of a 20 dB increase, seven of a 20 dB decrease, and six trials that varied randomly between 5 dB and 25 dB to ensure subsequent instances of the 20 dB change corresponded adequately with a cursor movement to the red line. Following training, participants were presented with two randomized blocks of the 100 unique stimuli and asked to indicate their perception of loudness with the mouse cursor, using the 20 dB lines as benchmarks. The pitch portion followed a similar paradigm, wherein training included a 128 Hz tone changing by 2 semitones that corresponded to red lines on the screen. Participants were also tested on six tones varying between .5 and 2.5 semitones to ensure 2 semitone changes approximately corresponded to the established benchmarks.

Results

Pitch. Physical changes in f_0, intensity, and phoneme all significantly predicted ratingss of pitch. As with loudness, participants' ratings of pitch approximated the trained amount; mouse movements corresponded to .84 semitones for an intended movement of 1 semitone, a degree by which mouse movements were recalibrated. The contribution of f_0 change to participants' ratings of pitch was significant, t (99) = 32.64, $p < .001$, Cohen’s d = 4.62. A 10 dB change in intensity corresponded to a .42 semitone pitch change, t (99) = 13.41, p < .001, Cohen’s d = 1.90. Lastly, a change along the phoneme continuum contributed to a .38 semitone change, t (99) = 12.52, p < .001, Cohen’s d = 1.77. The entire model accounts for 94% of the variance in pitch ratings, R^2 = .94, F (3, 96) = 467.41, p < .001. Figure 7 illustrates pitch ratings at all $f0$, intensity, and phoneme levels.

Loudness. Physical changes in intensity, f_0, and phoneme all significantly predicted ratings of loudness. Participants closely aligned their ratings of loudness with the trained amount. Using the 20 dB benchmark lines, a 10 dB change in intensity corresponded to a mouse movement equal to a 9.44 dB change in loudness. Mouse movements were recalibrated such that the actual moved amount (9.4) corresponded to the intended amount (10). The contribution of intensity change to participants’ loudness change ratings was significant, t (99) = 35.19, p < .001, Cohen’s d = 4.98. A one semitone change in f_0 corresponded to a 1.96 dB change, t (99) = 6.90, p < .001, Cohen’s d = .98. Lastly, a change along the phoneme continuum contributed to an approximate .75 dB change, t (99) = 2.65, p < .01, Cohen’s d = .38. The entire model accounts for 93% of the variance in loudness ratings, R^2 = .93, F (3, 96) = 430.97, p < .001. Figure 8 illustrates loudness ratings at all $f0$, intensity, and phoneme levels.

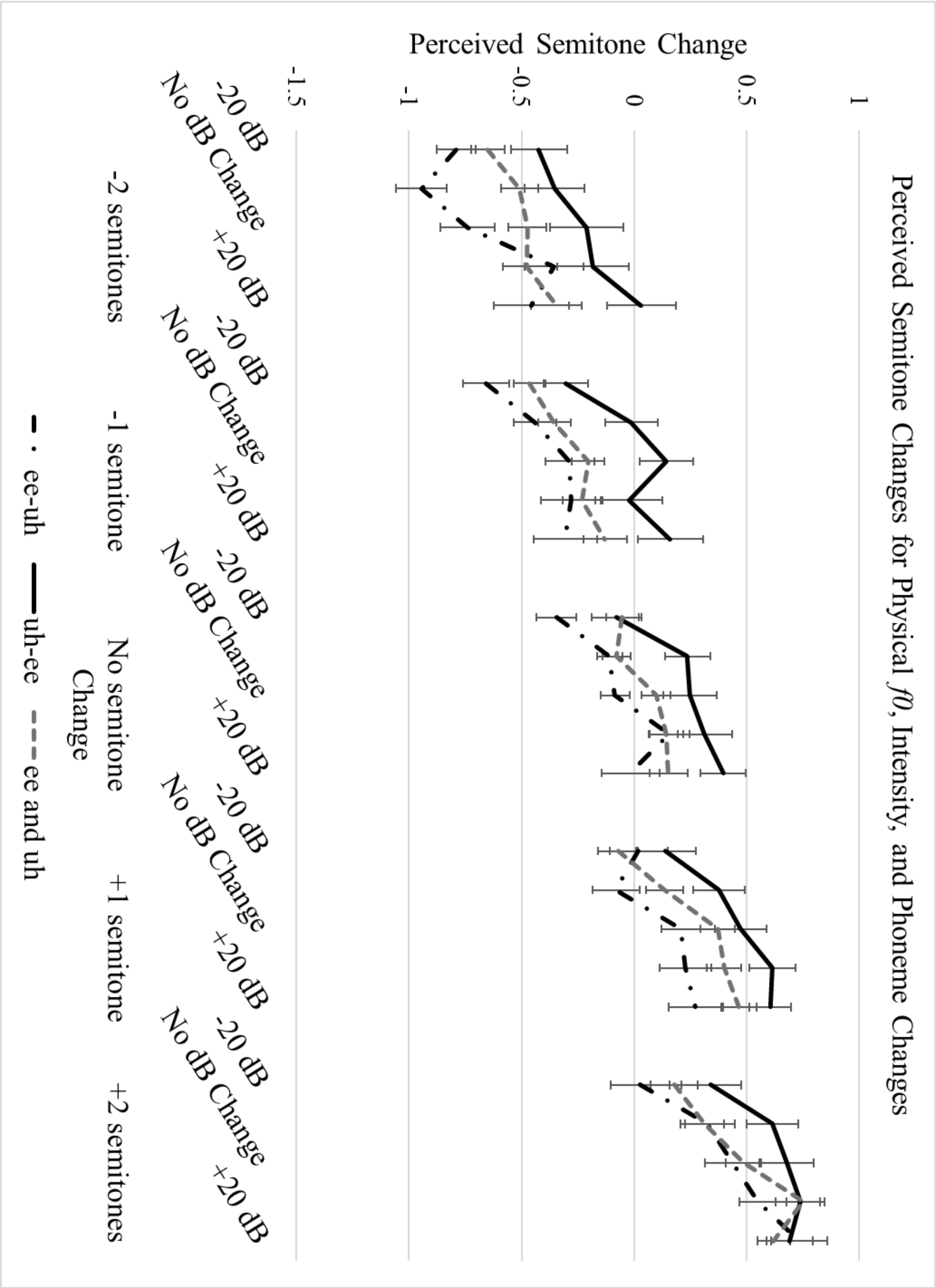


Figure 7. Pitch ratings in semitone units for all combinations of *f0*, intensity, and phoneme changes. Note that the predicted pattern (i.e., the *i* to *Λ* sweep is experienced as a decrease in frequency relative to the *Λ* to *i* sweep) is maintained at all levels.

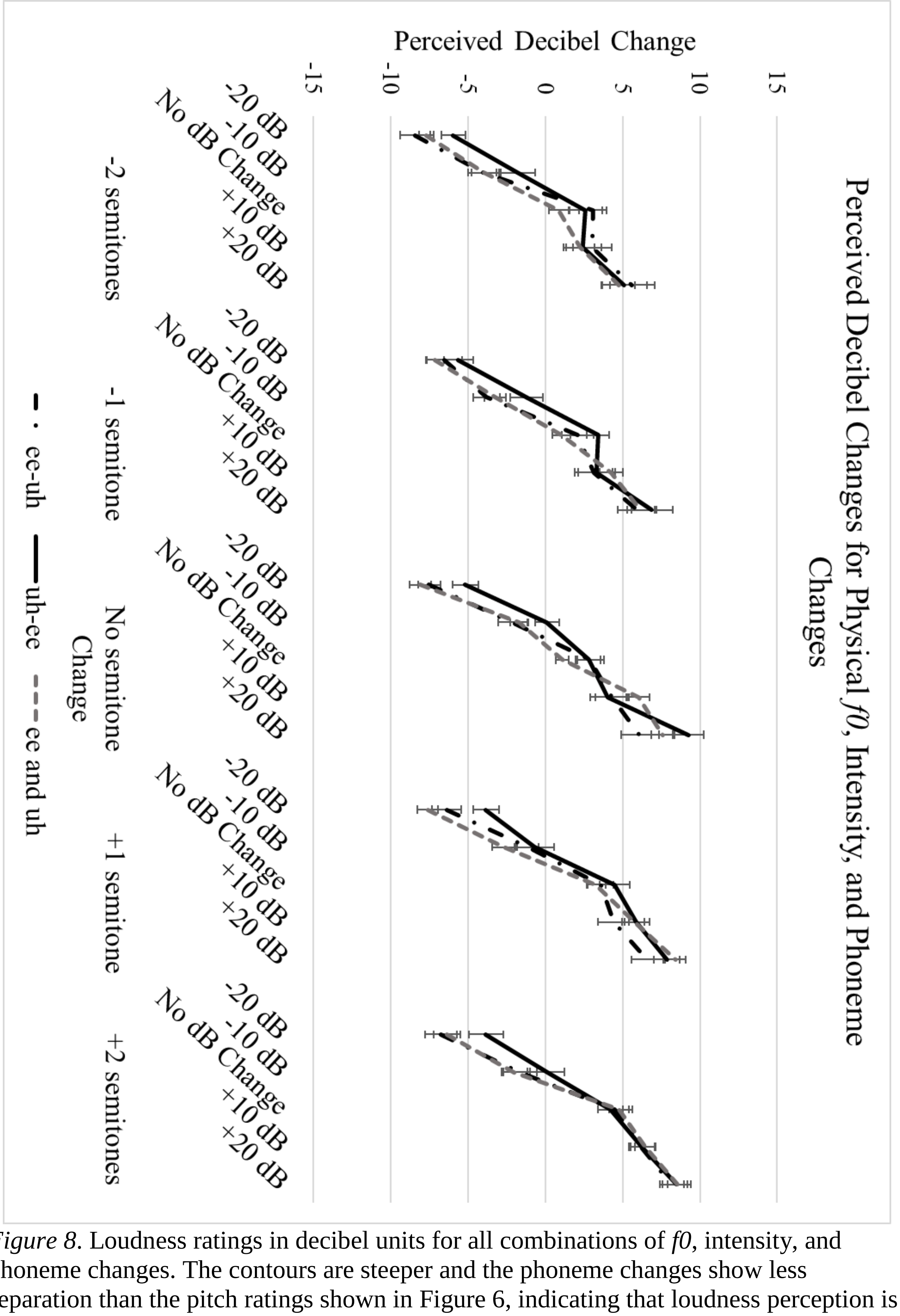


Figure 8. Loudness ratings in decibel units for all combinations of *f0*, intensity, and phoneme changes. The contours are steeper and the phoneme changes show less separation than the pitch ratings shown in Figure 6, indicating that loudness perception is less dependent on *f0* and phoneme change. This is corroborated in the numerical results.

Discussion

The results of the current experiment, as well as Experiments 1 and 2, confirm that there is a distinct perceptual illusion concerning timbre color. Furthermore, the .42 semitone perceptual change when the intensity of a sound modulates by 10 dB equates to a .63 semitone change at 15 dB. This figure is similar to the 1 semitone perceptual change at a 15 dB change noted by Neuhoff and McBeath (1996). As the spectral envelope changes from an *i* phoneme to an *Λ* phoneme, listeners perceive both loudness and pitch as falling. Though this confirms a perceptual bias, it remains unproven if there is a functional aspect to these biases such that they aid in acoustic source parsing and tracking.

CHAPTER 5

EXPERIMENT 4

The previous three experiments confirmed that there are robust perceptual correlations between pitch, loudness, and timbre. It was also demonstrated that these perceptual correlations mirror physical natural regularities between f_0, intensity, and timbre that are prevalent in nature. Though these types of natural regularities have been shown in several studies, the potential utility of this correlation has rarely been demonstrated, and has not been demonstrated for timbre contributions. Scharine and McBeath (2009) showed that detection thresholds for acoustic stimuli against a background of noise are lower when the stimuli change in accordance with natural regularities like the correlation between f_0 and intensity. Congruent changes (e.g., falling f_0 with falling intensity) were characterized by significantly lower detection thresholds than incongruent changes.

A shift in detection thresholds for sounds that act like typical auditory objects imply that the perceptual bias is not simply an imitation of the natural regularity, but one that aids in sound source detection and parsing. It is hypothesized that detection thresholds will be reduced when f_0, intensity, and timbre color change together consistent with natural regularities confirmed in Experiment 2 compared to when changes are incongruent.

Method

Participants. 28 (16 female) undergraduate students (mean age = 20.7, standard deviation = 2.8) from introductory psychology courses at Arizona State University participated in this experiment in exchange for partial course credit.

Procedure. A subset of 36 stimuli from Experiment 3 were used in this experiment. These included all four phoneme transitions (constant *i*, constant *Λ*, *i* to *Λ*, and *Λ* to *i*), three of the five f_0 sweeps (down 2 semitones, constant, and up 2 semitones), and three of the five intensity sweeps (down 2 dB, constant, and up 2 dB). Of these stimuli, 4 are constructed such that each auditory cue is congruent (e.g., *Λ* to *i* phoneme transition, a rising f_0, and a rising intensity). A further 8 stimuli are constructed such that each auditory cue is incongruent (e.g., *Λ* to *i* phoneme transition, a falling f_0, and a constant intensity). The remaining 24 stimuli are partially incongruent (e.g., *Λ* to *i* phoneme transition, a rising f_0, and a falling intensity).

Stimuli were presented using the method of constant stimuli within white noise characterized by a constant 65 dB intensity. The initial intensity of stimuli ranged from -38.1 to -7.2 decibel signal to noise ratio with step sizes corresponding to changes of .05 voltage gain (corresponding to slightly variable intensity changes averaging 4.4 decibels). Participants were asked to indicate which of the four phonemes/phoneme transitions they detected in each presentation. This subjective method was chosen over the more common objective method of simply indicating the presence of an auditory signal for two reasons. First, pilot studies indicated that there were no differences between congruency conditions, now were there any main effects of other cues. It is possible that the auditory system is able to detect a portion of the auditory signal at such a low intensity that the cues are too degraded to have a noticeable effect. Second, the current method allows for the computation of *d'* statistics for phoneme cues.

Results and Discussion

After partialing out the contribution of the f_0, intensity, and phoneme changes, as well as the intensity level of the cue and individual differences using an ANCOVA model, congruent cues did not offer a benefit to detection threshold as measured by overall proportion correct (OPC) for all intensity levels, F (2, 282) = .566, p = *n.s.*, η^2 = .004. Curves of proportion correct for congruent and incongruent cues at different initial intensity levels can be seen in Figure 9. Intensity and phoneme change both proved to be significant covariates, so an omnibus ANOVA was run to discern the effects of these variables. As in the ANCOVA, the contribution of f_0 change was not significant; however, intensity change accounted for a small effect, F (2, 252) = 8.92, $p < .001$, η^2 = .05. Sounds with constant intensity contours (*M* OPC = .61, *SD* = .28) were detected more readily than sounds with dynamic contours. Of the changing stimuli, sounds that began at higher intensities and fell (*M* OPC = .51, *SD* = .25) were identified correctly more often than sounds beginning at low intensities and rising (*M* OPC = .40, *SD* = .24). As the maximum intensity of all sounds at a given presentation level were equal, this finding is expected. The fact that falling intensities are identified more often than rising intensities may be an effect of cuing; the presence of an identifiable cue allows the target stream to be followed below uncued thresholds (Kidd Jr., Mason, Richards, Gallun, & Durlach, 2008).

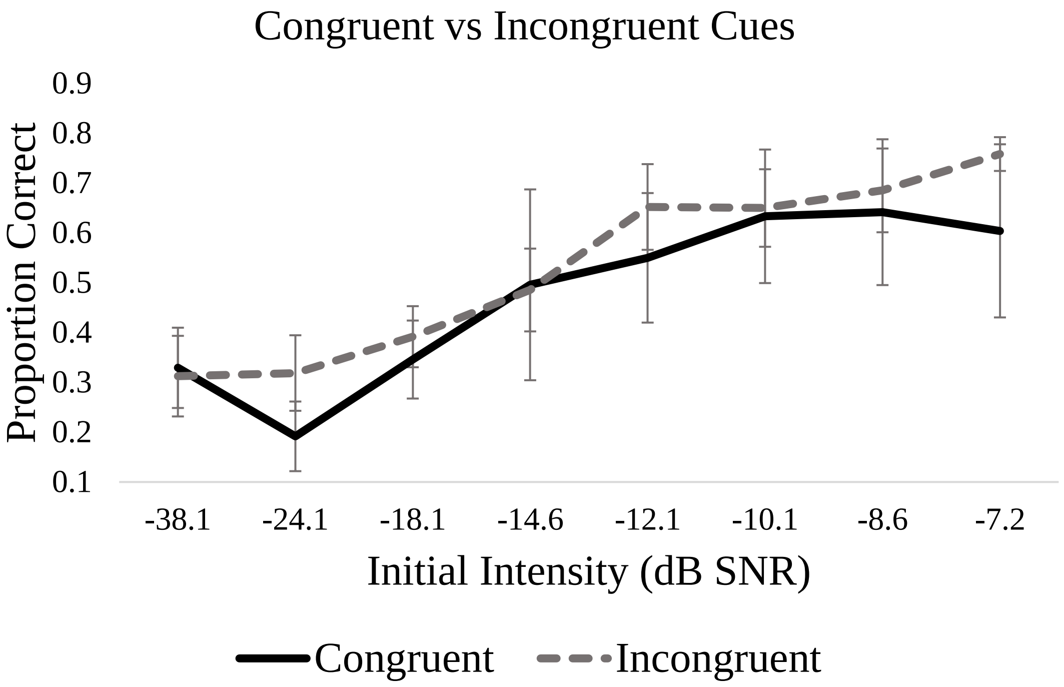


Figure 9. Proportion of phonemes correctly identified for congruent and incongruent changes of intensity, f_0, and phoneme at different initial intensity levels. Contrary to the hypothesis, there was no difference between cue congruencies.

Phoneme change also accounted for a small effect according to conventional standards (Cohen, 1988), *F* (3, 252) = 5.13, *p* < .01, η^2 = .042. Unlike the effect of intensity change, this difference was not expected. Post hoc comparisons revealed no significant difference in the detection rate of unchanging phonemes (*M* OPC *i* phoneme = .61, *SD* = .22; *M* OPC *Λ* phoneme = .58, *SD* = .24), but that they were detected more often than changing phonemes. Detection differences between stimuli that swept from *i* to *Λ* (*M* OPC = .43, *SD* = .24) and from *Λ* to *i* (*M* OPC = .39, *SD* = .31) were not significantly different, though the trend of the difference corroborates that of intensity. The results of Experiment 3 demonstrated that the *i* phoneme is experienced louder than the *Λ* phoneme even when both are presented at the same intensity. Thus, the differences observed for phoneme in this experiment mirror the aforementioned results of intensity; sounds that begin louder and fall are identified more often than sounds that rise in loudness. Additionally, phoneme perception is often categorical or considered part of a dual-process model in which one aspect is categorical (Chang, et al., 2010; Dehaene-Lambertz, 1997; Mottonen & Watkins, 2009; Repp, 1984). If this is true, the listener may have no signal information for the phoneme when the sound crosses the boundary between *i* and *Λ*, or vice versa. Similarly, the seemingly abrupt change of phoneme in noise may be jarring to the listener and disjoin the phoneme change from its accompanying *f0* and intensity changes.

When disregarding phoneme and assessing cue congruency only on the basis of f_0 and intensity, the omnibus ANCOVA model – again controlling for the direction of the f_0, intensity, and phoneme changes, as well as initial presentation intensity and individual differences – showed a significant effect of cue congruency, F (2, 137) = 3.87, $p < .05$, $\eta^2 = .054$. According to Cohen's guidelines for effect size using ANCOVA (Cohen, Cohen, West, & Aiken, 2003), the effect of congruent cues, while not statistically significant, is a small but detectable effect. Moreover, pairwise comparisons revealed that the difference between fully congruent cues (*M* OPC = .53, *SD* = .27) and fully incongruent cues (*M* OPC = .46, *SD* = .25) was also significant, $p < .05$. Curves of proportion correct for congruent and incongruent cues regardless of phoneme change at different initial intensity levels can be seen in Figure 10. These results replicate previous findings that show sounds in which changes along the f_0 and intensity dimension have similar fates (both falling or both rising) are detected at lower intensities than sounds in which changes along those dimensions are not related (Scharine & McBeath, 2009). In other words, sounds that change in accordance with the natural regularity provide a detection benefit when presented in noise.

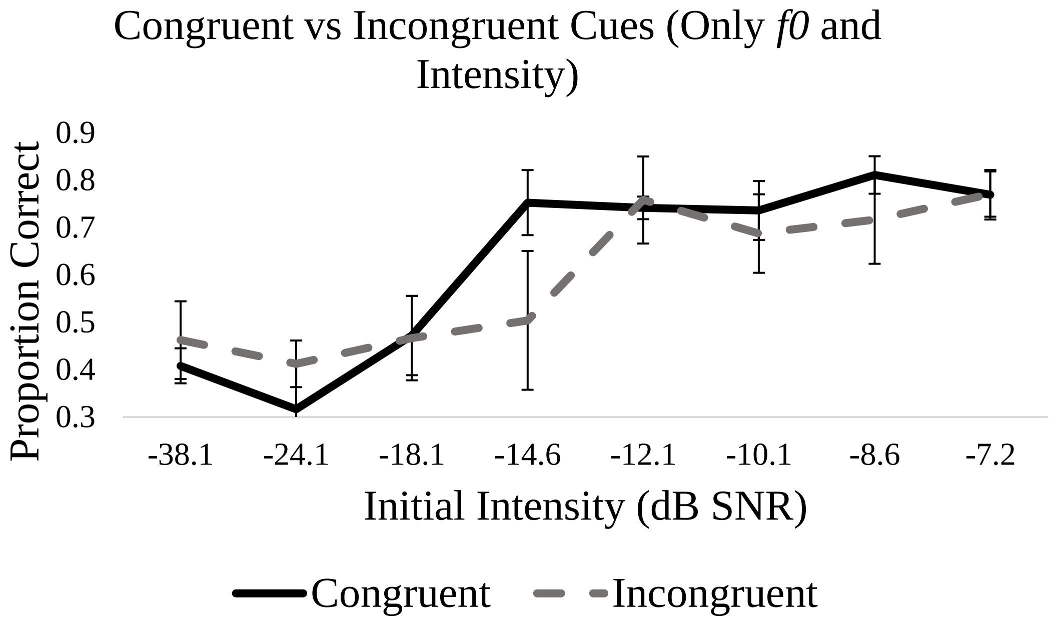


Figure 10. Proportion correct for congruent and incongruent changes of only *f_0* and intensity at different initial intensity levels. When averaged over all initial intensities, congruent cues were detected more often than incongruent cues. The depicted curves illustrate that congruent changes reached a near ceiling level at a lower signal to noise ratio than incongruent changes.

The results from Experiment 3 show that pitch and loudness changes are most heavily influenced by the physical change often associated with that perception (*f_0* and intensity changes, respectively), followed by the other physical change (intensity and *f_0*, respectively). In both domains, phoneme change was a significant, though small, contribution to the overall perception. It is possible that the failure to find an effect of congruent cues in the current experiment stems from this small effect of phoneme change; congruent changes in the *f_0* and intensity domain may overpower the small effect of phoneme change.

The unequal detection of different intensity and phoneme changes likely also overpowers the expect effect of phoneme change. In both cases, an unchanging stimulus was detected more often than a changing stimulus. Similarly, a stimulus that begins high (more intense or an *i* phoneme) were detected more often than a stimulus that began low (less intense or an *Λ* phoneme). This could be due to an informational release from masking (Plack & White, 2000). When participants listen to a broadband white noise masker and a narrowband white noise signal with simultaneous onsets, the narrowband signal is not detected. However, if the narrowband signal begins slightly before the masker, it can be detected throughout the duration of the masker. The unchanging intensity and phoneme stimuli are more readily detected because they are above threshold for the entire presentation. Rising and falling stimuli, however, have the same average frequency. Falling stimuli begin above threshold, which provides information that may allow participants to follow the stimuli into the noise. Rising stimuli, on the other hand, begin below threshold and provide no information. That the phoneme stimuli behaves in the same way as the intensity stimuli does strengthen the findings of Experiment 3; a sweep from *Λ* to *i* is perceived as falling in intensity. The possible release from masking is represented visually in Figure 11. To accurately gauge the difference between congruent and incongruent cues, intensity and phoneme changes would have to be equated for detection rates, likely for each individual participant. If the changes were equalized, the presence of a significantly higher detection rates for congruent cues in only the unchanging phonemes suggests that this finding would be replicated.

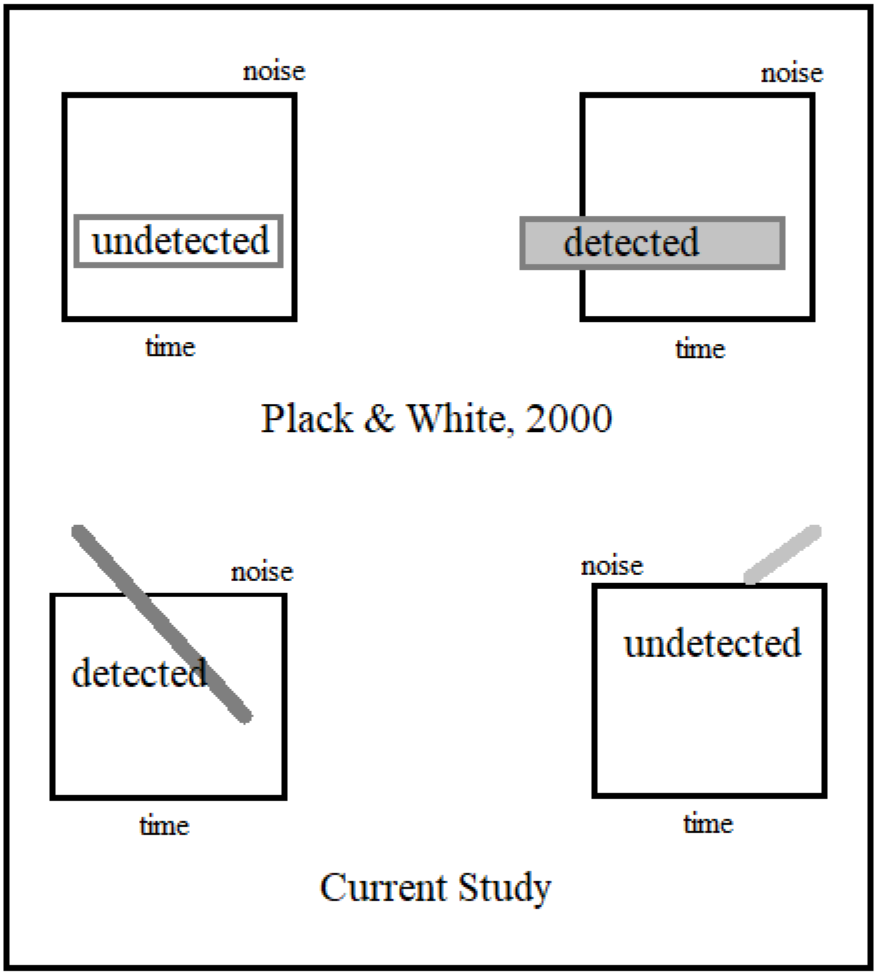


Figure 11. A representation of informational release from masking. A narrowband of white noise is detected within a broadband white noise masker when onset of the signal is slightly before that of the masker (White & Plack, 2000). This may be the case with falling intensity and phoneme stimuli in the current experiment; the small amount of unmasked stimuli allows participants to follow the stimulus into the noise. Rising stimuli begin masked and offer no information until the signal rises above threshold.

CHAPTER 6

GENERAL DISCUSSION

The current study is the first to examine the dimensions on which vowel phoneme stimuli are organized when controlled for f_0 and intensity. Experiment 1 demonstrated that vowels are arranged two dimensionally; by aperture size (or openness) and pitch. While the paradoxical perception of pitch differences in f_0-controlled vowels has been researched, Experiment 1 confirms that this perception is not adequately explained by physical characteristics of sound such as first or second formants or spectral centroid. The correlation between the pitch dimension of the MDS solution and rated pitch accounts for more variance than any physical characteristic. According to the ASA framework, perceptual biases such as this can flag corresponding natural regularities.

Indeed, Experiment 2 confirmed the existence of just such a regularity. In both unconstrained conversational speech and the very unconstrained condition of scat singing, f_0s of phonemes were voiced in accordance with the pitch dimension of the MDS solution in experiment 1. For all observations, *i* phonemes were voiced at significantly higher f_0s than *Λ* phonemes. In standard lyrical singing, however, the reguarity disappeared. This was predicted, however, as lyrics and melody dictate the f_0 of phonemes. Experiments 1 and 2 both illustrate contributions of the pitch height dimension (in the form of spectral differences between vowel phonemes) in the perception of pitch.

Experiment 3 identified the exact contributions of f_0 and vowel changes on loudness and the contributions of intensity and vowel changes on pitch. Experiment 4 failed to demonstrate a detection threshold benefit of congruent changes (e.g., *i* to *Λ*, falling f_0, and falling intensity) over incongruent changes. However, when analyzing only unchanging *i* and *Λ* phonemes, congruent changes (e.g. falling f_0 and intensity) were detected at quieter intensities than incongruent changes. Taken together, these results can be used in several areas related to speech detection and identification, noise reduction, speech production, and alarm system design.

Modern hearing aids are built such that frequency bands outside the range of human speech (the upper and lower extremes of frequencies humans can detect) are either not amplified or not amplified to the extent of the speech frequencies (Boymans & Dreschler, 2000; Lunner, Hellgren, Arlinger, & Elberling, 1997). Though this aids in reduction of non-target stimuli, there are many sources of noise that fall inside this speech window and targets that fall outside (Edwards, 2007). The biases and parsing strategies used by the normally functioning auditory system are often poorly understood and, thus, poorly modeled by modern assistive technology (Alain, Arnott, & Picton, 2001). Using the auditory parsing bias described in the current paper, auditory assistive technology can identify and amplify target frequency ranges by monitoring for congruent f_0, intensity, and pitch height changes (perhaps by using models of vowel phonemes or the spectral centroid). Similarly, noise bands with incongruent cues can be limited or eliminated entirely.

Not only can the intelligibility of real speech be enhanced to individuals with assistive devices but, using the results of the current paper, intelligibility of synthetic speech to listeners with normal hearing can also be enhanced. Current synthetic speech has a correct perception rate of approximately 85% when presented in low to moderate noise (Cooke, et al., 2013). Using congruent sweeps, such as raising the speech f_0 and intensity slightly for *i* phonemes relative to *Λ* phonemes, could boost the intelligibility rating in noise as it would more closely replicate the natural regularities of speech. Similarly, synthetic speech is especially susceptible to degredations when presented over telephone bands as the signal is already constrained (Pocta & Beerends, 2015). Adding congruent sweeps could allow listeners to exploit the biases already inherent in their perception of speech and natural sounds to infer the missing information.

The current study not ony confirms that there is a natural regularity in the physical world for changes in spectral envelope (such as phonemes), f_0, and intensity to be correlated, but also demonstrates a perceptual bias to experience sounds in accordance with that regularity even when the actual acoustic information does not. This perceptual bias likely stems from exposure to the natural regularity in the form of both natural sounds and human vocalizations which, in turn, reinforces the production of sounds in accordance with bias. The bias to perceive a relationship between f_0, intensity, and pitch height aids in the ability of listeners to parse and identify objects in their acoustic environment, the fundamental goal of Auditory Scene Analysis.

REFERENCES

Alain, C. (2007). Breaking the wave: Effects of attention and learning on concurrent sound perception. *Hearing Research, 229*(1-2), 225-236.

Alain, C., Arnott, S. R., & Picton, T. W. (2001). Bottom-up and top-down influences on auditory scene analysis: Evidence from event-related brain potentials. *Journal of Experimental Psychology: Human Perception and Performance, 27*(5), 1072-1089.

Allen, E. J., & Oxenham, A. J. (2014). Symmettric interaction and interference between pitch and timbre. *Journal of the Acoustical Society of America, 135*(3), 1371-1379.

Boersma, P., & Weenink, D. (2015). Praat ver. 6.0.

Boymans, M., & Dreschler, W. A. (2000). Field trials using a digital hearing aid with active noise reduction and dual-microphone directionality. *Audiology, 39*, 260-268.

Bregman, A. S. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge, MA: MIT Press.

Broze, Y., & Huron, D. (2013). Is higher music faster? Pitch-speed relationships in Western compositions. *Music Perception, 31*(1), 32-45.

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *Journal of the Acoustical Society of America, 118*, 471-482.

Carlson, R., Fant, G., & Granstrom, B. (1974). Two-formant models, pitch, and vowel perception. *Acta Acoustica, 3*, 360-362.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience, 13*(11), 1428-1432.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd ed.* Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, Third Edition.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., & Tang, Y. (2013). Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication*, *55*(4), 572-585.

Dehaene-Lambertz, G. (1997). Electrophysical correlates of categorical phoneme perception in adults. *NeuroReport, 8*, 919-924.

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

Deutsch, D. (2009). Auditory Illusions. In E. B. Goldstein, *Encyclopedia of Perception, Volume I* (pp. 160-164). Thousand Oaks, CA: Sage.

Doppler, J. C. (1842). Uber das farbige Licht der Doppelsterne und einiger anderer Gestirne des Himmels. In *Versuch einer das Bradley'sche aberrations-theorem als integrirrenden Theil in sich schliessenden allgemeineren Theorie* (p. 465). Prague: K. Bohm Gesellschaft der Wissenschaften.

Edwards, B. (2007). The future of hearing aid technology. *Trends in Amplification, 11*, 31-45.

Erickson, R. (1975). *Sound Structure in Music.* Los Angeles: University of California Press.

Fowler, C. A., & Brown, J. E. (1997). Intrinsic f0 differences in spoken and sung vowels and their perception by listeners. *Perception and Psychophysics, 59*(5), 729-738.

Fredrickson, S. (2003). *Scat Singing Method.* New York: Scott Music Publications.

Fujisaki, H., & Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics, 16*(1), 73-77.

Gibson, J. J. (1986). *The Ecological Approach to Visual Perception.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Hout, M. C., Goldinger, S. D., & Ferguson, R. W. (2013). The versatility of SpAM: A fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology (General), 142*, 256-281.

International Phonetic Alphabet. (2016). *Phonetic Vowels.* Retrieved from International Phonetic Alphabet: internationalphoneticalphabet.org

Johnson, A., McBeath, M. K., & Patten, K. J. (2014). Drumming and tempo: The effects of loudness change on tempo perception and action. *Paper presented the 2014 Auditory Perception, Cognition, and Action Meeting, Long Beach, California*.

Johnston, I. (1989). *Measured Tones: The Interplay of Physics and Music.* New York: Adam Hilger.

Kidd Jr., G., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. L. (2008). Informational Masking. In W. A. Yost, A. N. Popper, & R. R. Fay, *Auditory Perception of Sound Sources* (pp. 143-190). New York: Springer.

Krumhansl, C. L., & Iverson, P. (1992). Perceptual interactions between musical pitch and timbre. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 739-751.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling (Vol. 11).* New York: Sage.

Lunner, T., Hellgren, J., Arlinger, S., & Elberling, C. (1997). A digital filterbank hearing aid: Three digital signal processing algorithms - User preference and performance. *Ear and Hearing, 18*(5), 373-387.

Luo, X., & Soslowsky, S. (2017). Interactions between pitch and timbre perception in normal-hearing listeners and cochlear impant users.

Marozeau, J., & de Cheveigne, A. (2007). The effect of fundamental frequency on the brightness dimension of timbre. *Journal of the Acoustical Society of America, 114*, 383-387.

Mazzoni, D., & Dannenberg, R. (2016). Audacity, ver. 2.1.

McAdams, S., & Bregman, A. (1979). Hearing musical streams. *Computer Music Journal, 3*(4), 26-43.

McBeath, M. K. (2014). The Fundamental Illusion. *Paper presented at the 55th annual meeting of the Psychonomic Society, Long Beach, California*.

McBeath, M. K., & Neuhoff, J. G. (2002). The Doppler effect is not what you think it is: Dramatic pitch change due to dynamic intensithy change. *Psychonomic Bulletin and Review, 9*(2), 306-313.

Melara, R. D., & Marks, L. E. (1990). Interaction among auditory dimensions: Timbre, pitch, and loudness. *Perception and Psychophysics, 48*(2), 169-178.

Michalsky, J. (2016). Perception of pitch scaling in rising intonation on the relevance of f0 median and speaking rate in German.

Miller, M. (2004). *The Complete Idiots Guide to Solos and Improvisation.* Indianapolis, IN: Alpha Books.

Mottonen, R., & Watkins, K. E. (2009). Motor representations of articulators contribute tto categorical percption of speech sounds. *The Journal of Neruoscience, 29*(31), 9819-9825.

Neuhoff, J. G., & McBeath, M. K. (1996). The Doppler illusion: The influence of dynamic intensity change on perceived pitch. *Journal of Experimental Psychology: Human Perception and Performance, 22*(4), 970-985.

Pitt, M. (1994). Perception of pitch and timbre by musically trained and untrained listeners. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 976-986.

Plack, C. J., & White, L. J. (2000). Perceived continuity and pitch perception. *The Journal of the Acoustical Society of America, 108*(3), 1162-1169.

Pocta, P., & Beerends, J. G. (2015). Subjective and objective measurement of synthesized speech intelligibility in modern telephone conditions. *Speech Communication, 71*, 1-9.

Repp, B. H. (1984). Categorical perception: Issues, methods, and findings. In N. J. Lass, *Speech and Language: Advances in Basic Research and Practice* (pp. 243-323). Cambridge, Massechusetts: Academic Press, Inc.

Scharine, A. A., & McBeath, M. K. (2009). The integral perception of sound intensity and frequency: From perceptual regularity to auditory scene analysis.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science, 210*(4468), 390-398.

Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. *Psychological Review, 89*(4), 305.

Singh, P. G., & Hirsh, I. J. (1992). Influence of spectral locus and F0 changes on the pitch and timbre of complex tones. *Journal os the Acoustical Society of America, 92*, 2650-2661.

Slawson, A. W. (2005). Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. *The Journal of the Acoustical Society of America, 43*(1).

Thomas, I. B. (1969). Perceived pitch of whispered vowels. *The Journal of the Acoustical Society of America, 46*, 468-470.

von Bismarck, G. (1974a). Timbre of steady sounds: A factorial investigation of its verbal attributes. *Acta Acustica united with Acustica, 30*(3), 146-159.

von Bismarck, G. (1974b). Sharpness as an attribute of the timbre of steady sounds. *Acta Acustica united with Acustica, 30*(3), 159-172.

von Helmholtz, H. L. (1877/1954). *On the Sensations of Tone.* (A. J. Ellis, Trans.) New York: Dover Publications.

Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic f0 of vowels. *Journal of Phonetics, 23,* 349-366.

Zilany, M. S., Bruce, I. C., Nelson, P., & Carney, L. H. (2009). A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics. *Journal of the Acoustical Society of America, 126*(5), 2390-2412.

Zilany, M. S., Bruce, I., & Carney, L. H. (2014). Updated parameters and expanded simmulation options for a model of the auditory periphery. *Journal of the Acoustical Society of America, 135*(1), 283-286.

APPENDIX I
SOUND RECORDINGS USED IN EXPERIMENT 2

SOUND RECORDINGS USED IN EXPERIMENT 2

Table 1

Recordings and Songs Used to Examine Phoneme/f_0 Correlation

Title	Voicer	Year
How to Scat	Indra Aziz	2014
Scat This	Scatman	2012
The Scat Song	Cab Calloway	1932
What's Jazz?	Ella Fitzgerald	1976
What's Jazz?	Mel Torme	1976
Summertime	Maya Bensalem	2012
I've Got a Crush on You	Reina Lam	2008
You Raise Me Up	Josh Groban	2003
Don't Stop Me Now	Freddie Mercury	1978
Starman	David Bowie	1972
Big Iron	Marty Robbins	1959
Radiolab: Remembering Oliver Sacks	Robert Kurlwich	2015
Science Friday: Coffee's Natural Creamer	Harold McGee	2013
Fresh Air	Terry Gross	2015
Fresh Air	Louis C K	2015
Numberphile: British Numbers Confuse Americans	Lynne Murphy	2013
Numberphile: British Numbers Confuse Americans	CGP Grey	2013

🟦 Scat Singing   🟧 Lyrical Singing   🟩 Interview