1-8-2016

Cell Phone vs. Microphone: Judging Emotion in the Voice

Joshua Green
joshua.2.green@uconn.edu, joshua.2.green@uconn.edu

Cell Phone vs. Microphone:

Judging Emotion in the Voice

Joshua Jacob Green

B.A., University of New Hampshire, 2008

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Arts

at the

University of Connecticut

2016

APPROVAL PAGE

Master of Arts Thesis

Cell phone vs. Microphone:

Judging Emotion in the Voice

Presented by

Joshua J Green, B.A.

Major Advisor _____

Inge-Marie Eigsti, Ph.D.

Associate Advisor _____

Chi-Ming Chen, Ph.D.

Associate Advisor _____

James Magnuson, Ph.D.

University of Connecticut

2016

Acknowledgements

This project could not have been possible without the help I have had along the way. I want to thank my friends and family back home in New Hampshire - who have been so supportive - especially during times when I have been distant and engrossed in research. I also want to thank my wonderful lab mates: Christy, Brian, Allison and Anders, who have been supportive from my first day at UConn. Thanks to my research assistants who have been generous in their help testing tasks and creating stimuli. Thanks also to Marianne Barton, who has continually shown faith in my growth as a developing professional.

Of course, I am deeply grateful for our lab's amazing mentor, Inge-Marie Eigsti, who has believed in my academic abilities and guided me through the challenging maze that is the first few years of graduate school. Thanks to Chi-Ming Chen, for providing interesting methodological and theoretical ideas during the development of my research, as well as showing a genuine interest in my research pursuits. I also want to thank Jim Magnuson, who has provided me with incredibly helpful feedback on my talks and presentations, as well as help with hugely time-saving 'R' scripts.

Finally, I would like to thank my classmates and faculty affiliated with the IGERT program who have helped me through various hurdles along the way. Thanks to Rachel Theodore and Emily Myers for help with speech analysis and Garrett Smith for the thoughtful listening and help with statistics.

Table of Contents

Cell phone vs. microphone:

Judging emotion in the voice

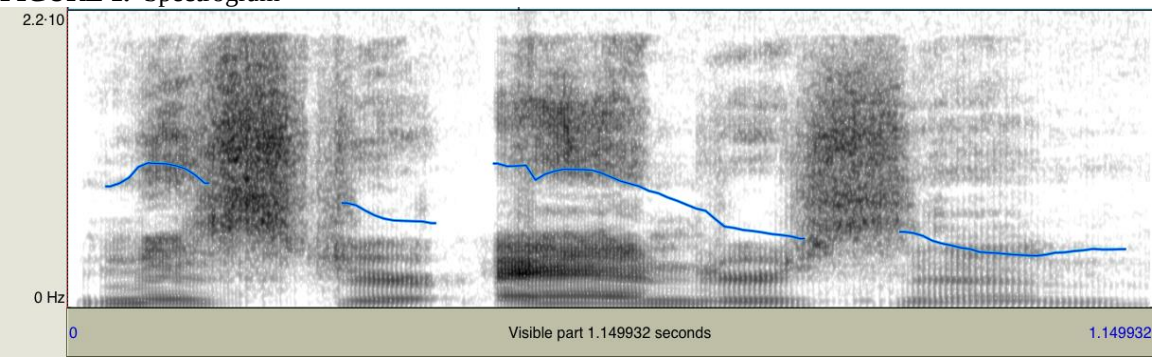
Historically, researchers have suggested that the "voice… is a mirror of the personality of the speaker" (Addington, 1968, p.492). Some of the earliest researchers of the human voice hypothesized that inflection style during reading (greater variability) is an accurate marker of intelligence (Wolf & Murray, 1937), though these early hypotheses were not replicated in later research. More recent research suggests that vocal attractiveness elicits judgments of positive personality traits, a phenomenon called "what is beautiful is good" (Babel, McGuire, & King, 2014; McAleer, Todorov, & Belin, 2014; Zuckerman & Driver, 1988). Likewise, voice disorders are associated with negative attributes (Blood, Mahan, & Hyman, 1979; McAllister & Sjölander, 2013).

Listeners can make accurate judgments of multiple physical attributes including *gender* (Zäske, Skuk, Kaufmann, & Schweinberger, 2013), *age* (Harnsberger, Shrivastav, Brown, Rothman, & Hollien, 2008; Hughes & Rhodes, 2010; Ptacek & Sander, 1966), and *body size* (Bruckert, Liénard, Lacroix, Kreutzer, & Leboucher, 2006; Evans, Neave, & Wakelin, 2006), though judgments of are not always accurate for height and weight (Bruckert et al., 2006; Lass, Barry, Reed, Walsh, & Amuso, 1979; van Dommelen & Moxness, 1995) or for race, which is confounded with dialect (Mayo, 1994; Walton & Orlikoff, 1994). The most accurate inferences involve emotional state, associated with vocal cues and intonation patterns. These suprasegmental aspects of speech are non-linguistic; that is, they contrast with the segmental aspects of speech such as vowels or consonants.

Suprasegmental qualities include intonation, stress and pauses, as well as pitch, rhythm, and duration (Lahey & Bloom, 1988). Taken together, these speech variables comprise prosody, sometimes called the "melody" of speech. *Grammatical prosody* refers to pauses, stresses, and intonation changes that change the form and meaning of a sentence. For example, rising pitch at the end of a sentence indicates a question, while a pause can communicate clausal groups (e.g. chocolate cake and ice cream vs. chocolate, cake, and ice cream). *Affective prosody* refers to emotional qualities, such as the pitch and rhythm changes that correspond to anger, fear or other emotions. The current study included a manipulation of affective prosody variables.

Acoustic qualities of speech. There are many acoustic correlates, or cues, in the speech signal, that convey emotions and physical attributes, including both basic and more complex measures. The speech signal as commonly analyzed contains three dimensions: time (x-axis), frequency (y-axis), and intensity/amplitude (z-axis, represented as color or saturation levels); see Figure 1. In a musical note, the frequency dimension contains multiple repeating patterns of mathematically predictable periods; the lowest of these is the fundamental frequency, abbreviated F0. F0 and amplitude are associated with perceptions of pitch and loudness, respectively. F0 in a voice is largely determined by the size of the vocal folds and thus is typically lower in men than in women and in adults than in children. Amplitude is largely determined by the volume of air passing over the vocal folds; thus, physical size is associated with greater volume. Note that these two parameters interact. For example, a high pitch that slowly increases in intensity (overall amplitude) is often perceived as increasing in pitch (Rossing, 1982).

FIGURE 1. Spectrogram



This figure is a spectrogram of the phrase "Mr. Anderson" spoken in an 'angry' tone of voice by Talker 'A'. The blue line represents the fundamental frequency contour. The y-axis represents frequency, the x-axis represents time, and the shades of gray represents intensity or amplitude, where darker colors indicate more energy.

The average adult male has an F0 of 120 Hz; average F0 is 210 Hz for females (Traunmüller & Eriksson, 1994). When constant pressure is applied to the vocal chords, they open and close at a specific frequency that determines F0. As air exits the larynx, it passes through the openings of the oral cavity, each structure of which serves to amplify or dampen specific frequencies of the vibrating vocal folds. The frequencies at which energy is most concentrated, called the formants (abbreviated F1, F2, F3, etc.), are at intervals above F0. In the "source-filter" model of speech production, the vocal folds are the source, while the rest of the oral cavity serves as the filter. As the air moves through this vocal filter, it is shaped into the bits of speech that we call phonemes, altering the signal in subtle and complex ways (for a detailed review of the anatomy of speech, see McAllister & Sjölander, 2013). Speech perception is complex and multi-determined.

Speech rate or duration, often measured in syllables per second, is also an important voice quality. Speech rate can be measured in multiple ways. Jacewicz, Fox, O'Neill, and Salmons (2009) discussed a distinction between speech rate and articulation rate: articulation rate excludes suprasegmentals, while speech rate does not. Additional important voice cues include F0 or amplitude *perturbation*, which refer to the consistency of the vibration of the vocal folds. Perturbation measures include jitter, shimmer, and harmonic-to-noise ratio (HNR). Jitter refers to

the degree to which the cycle from one opening and closing of the vocal folds is consistent over time. Thus, tighter control of the vocal chords results in less jitter. Many vocal pathologies include increased jitter (Teixeira & Fernandes, 2014), and one common cue to aging is a high percentage of jitter due to a loosening of the vocal folds (Wilcox & Horii, 1980). Similar to jitter, shimmer is a measure of amplitude perturbation. Visually, shimmer appears as a series of peaks in a mountain range; low shimmer would be visualized as peaks of equal height. Although shimmer is amplitude variability, both jitter and shimmer are usually described perceptually as breathiness or hoarseness (Eskenazi et al., 1990). HNR is a measure of "noisiness" and refers to the ratio between the periodic and aperiodic aspects of the speech signal, sometimes referred to as tonal and noisy (e.g. Hammerschmidt and Jürgens, 2006). Perceptually, a listener might report more "roughness" in a voice with a low HNR. There are myriad ways to analyze a speech single depending on the methodological goals. For example, Texiera and Fernandes (2014) sought to extensively characterize healthy voices, thus they focused exclusively on jitter, shimmer and HNR, and identified four separate parameters of jitter (i.e., absolute, relative, relative average perturbation, and period perturbation quotient) as well as four parameters of shimmer. Such specific and detailed acoustic parameters are less commonly studied, however, in the majority of research on the human voice.

F0 and amplitude have been the primary acoustic values of interest in most voice research. Protopapas and Lieberman (1997) examined cues of emotional stress in recordings of a panicked helicopter pilot, and found that mean and maximum F0 were associated with perceived stress, whereas jitter was not (it was associated with hoarseness). Yet, a high F0 seems to signal many emotions. Belyk and Brown (2014) reported high F0 in several emotional categories. They organized emotional expression into three groups: motivational (e.g., joy, distress), moral (e.g.,

gratitude, anger), and aesthetic (e.g., sensual pleasure, disgust). They found that high mean F0 and mean amplitude were associated with positive valence in the 'motivational' emotions and vice-versa for negative valence. The 'moral' emotions were more nuanced: while positively and negatively valenced emotional expressions had similar amplitudes, they differed in F0. Anger was signaled via low F0 and high amplitude, a finding consistent with other research (McAleer et al., 2014). In contrast, gratitude was signaled by high F0 and low amplitude. In the 'aesthetic' emotion family, sensual pleasure was signaled by low amplitude and the lowest F0 of any emotion, while disgust was signaled by high F0 and high amplitude. Taken together their results suggest that the F0 and amplitude of a speech signal seem to be necessary, but not sufficient cues for detecting emotional valence. Indeed, both happy and panicked speech may show high F0 and high amplitude, but clearly convey very different emotional meaning. Other cues, such as formant frequencies, duration, and F0 contour may also be important.

Hammerschmidt and Jürgens (2006) also compared acoustic cues with emotional categories. They analyzed utterances of the name 'Anna' spoken by drama students and split them into two categories: aversive (e.g., 'rage/hot anger') or hedonistic (e.g., affection, tenderness) emotions. Using stepwise discriminant function analysis, they decomposed 94 original acoustic parameters down to fifteen parameters that were uncorrelated with each other. Of these fifteen parameters, six correctly classified 75% of the utterances into aversive or hedonistic categories. These parameters included: 1) amplitude (relative amplitude based on the logarithmic root mean square method); 2) duration; 3) local modulations of the first dominant frequency band; 4) mean harmonic-to-noise ratio; 5) mean range (difference between highest and lowest frequency within a segment averaged across all time segments); and 6) the mean distribution of frequency amplitudes (frequency at which the 50% of the sum of all amplitudes is

reached). This method of using uncorrelated cues is useful because it sheds light on what cues a listener might use above and beyond that of mean F0, which likely failed to predict emotional category because some emotions in both categories had similar mean F0 (e.g., rage/hot anger and joyful surprise). These findings are similar to those found in Belyk and Brown (2014); mean and max F0 did not differentiate between hedonistic and aversive emotions, confirming the nuanced nature of affect transmission in speech. Despite the failure of mean F0 as a predictor of emotional category in these two studies, it is clear that F0 is important in communicating emotion in the voice. F0 may be best thought of as a cue that is interactive with other cues (such as amplitude and HNR) to signal individual emotions.

Signal quality. Given the complexities of the speech signal as described above, it is clear that we know only some of the cues critical to identifying speaker qualities, including affect, from the voice. Moreover, it is unclear to what extent these differing cues interact with each other to carry affect in the voice. Psycholinguists and other researchers may often be motivated to include recordings of speech in studies of language acquisition and language perception. For example, a study of infant language acquisition might include a comparison of maternal and paternal speech. While researchers would ideally make recordings in sound booths using high-quality microphones, for practical reasons, it would be useful to collect data from a speaker who is not physically present. Of note, this approach may also be more ecologically valid. For example, our lab is conducting a study of affective prosody comprehension in children with autism, comparing perception of affect in a *stranger's* to affect in the subject's *mother's* voice. To make the stimuli in advance, it is critical to record the mother over the phone. It is thus essential to know whether cell phone recordings of affective speech contain the relevant cues necessary for emotion perception on the part of the listener.

The quality of speech diminishes significantly over a cell phone as compared to a high-quality microphone. For example, cell phone transmission filters all signals above ~4kHz-4.5kHz, while a typical microphone has a range of up to 22kHz. Several studies have used cell phone recordings, including investigations testing the effect of transmission on speaker recognition (Byrne & Foulkes, 2004; Jong, Hudson, Nolan, & Mcdougall, 1995; Künzel, 2001). To date, one study of affective prosody has included a comparison of cell phone and microphone recordings (Leemann, Kolly, & Dellwo, 2014). This study examined prosodic features that contribute to speaker identity in German speakers who produced both spontaneous and read speech. Results revealed high between-speaker variability (different people have different speaking styles), but low within-speaker variability for read and spontaneous speech (the same people have similar styles whether reading or speaking freely). The use of a mobile phone had no significant effects on speech rhythm or voicing. This study provided evidence that some features of speech are robust to the distortions introduced by a cell phone. It is not known whether other manipulations of the speech signal, such as F0, are also robust to recording variability, though at least one study suggests that a phone transmission may cause a slight increase in F0 (French & Howard, 1995).

The reader might ask: why not simply compare cell phone versus microphone recordings using acoustic processing software? Such an approach would be compromised by a well-known phenomenon in speech research called the "lack of invariance" problem. This refers to the finding that a speech signal can change while a listener's perception remains the same, and vice-versa. For this reason, one cannot look at a spectrogram and "read" it like a musician reads music. Although speech is organized mentally and on paper into separable units (i.e., letters, words, phrases), the auditory realization of an utterance takes the form of an uninterrupted

stream of acoustic energy, resulting in ambiguous and multi-determined percepts from the same signal. For example, Miller found speech rate manipulations changed the perception of a /w/ to a /b/ (Miller, 1983). This "many-to-many mappings" problem permeates language research (see Magnuson & Nusbaum 2007 for a summary of this problem). As described above, there are many variables that affect speech perception. For example, pitch, voice quality, articulation style, and loudness interact to alter perceptions of gender in male-to-female transsexuals who seek to feminize their voice, suggesting that very different signals can give rise to the same perception (Dacakis, Oates, & Douglas, 2012). Vocal cues simply do not map onto indexical and social percepts in a one-to-one manner.

The current study. Motivated in part by other research on affective perception in phone versus microphone recordings by children with autism, the present study aimed to examine the perception of stimuli recorded via both cell phone and microphone recordings. We used stimuli that had been morphed on a continuum from neutral to a strong emotion (happy or angry); this approach permitted a nuanced measure of perception through both recording modalities. One program, Tandem-STRAIGHT (hereafter, STRAIGHT), was created for this very purpose (Kawahara et al., 2008). STRAIGHT decomposes the speech signal into *source* (i.e., vocal folds) and *filter* (i.e., larynx, etc.) information, resulting in five parameters: 1) F0 trajectory; 2) frequency; 3) duration; 4) spectrotemporal density (which frequencies are contain the most energy); and 5) aperiodicity (e.g. jitter and shimmer components) (Kawahara et al., 2008; Kawahara, Takahashi, Morise, & Banno, 2009). STRAIGHT has been used in several studies of emotional speech perception (Bestelmeyer, Maurage, Rouger, Latinus, & Belin, 2014; Bestelmeyer, Rouger, DeBruine, & Belin, 2010; Doi et al., 2013; Skuk & Schweinberger, 2013). Bestelmeyer and colleagues (2010) used an anger-fear continuum to study auditory adaptation.

They showed that adaptation to angry stimuli caused the ambiguous stimuli (middle of the continuum) to be perceived as more fearful, and the fearful stimuli to seem more angry when reversed. Skuk and Schweinberger (2013) used an angry-happy continuum to study the effect of modality (auditory, visual) on perception of emotion in the voice. They found similar results to Bestelmeyer: voices were perceived as happier after adaptation to angry voices and vice versa. Doi and colleagues (2013) used morphed speech to study the perception of affective prosody in adults with high-functioning autism. They found that, compared to a control group, individuals with autism were less accurate at identifying emotions (happy, sad or angry) at high emotional intensities. The current study makes use of STRAIGHT to test the perception of affective cues via cell phone versus microphone. Specifically, we ask whether graded affective stimuli can be perceived with similar sensitivity from cell phone as compared to microphone recordings.

Methods

Participants

Forty-two participants were recruited from the University of Connecticut participant pool. Participants received class credit for participating. Participants were excluded if they were not native speakers of English or if they reported hearing impairment. This convenience sample was used in service of rapid recruitment. Participant age was 19.9 years on average (*SD* = 1.2), and participants were majority Caucasian (M = 67%) and female (70%). Informed consent was obtained from each participant and they received course credit for participating. All procedures were approved by the University of Connecticut Institutional Review Board.

Stimuli

Stimuli were recorded via a Shure PG42USB Cardioid condenser microphone using Audacity recording software on a MacBook Pro, and via an iPhone using a third party

application that records phone calls (Call Recorder v 1.4.3). The microphone recordings had a sampling rate of 44.1kHz, and the cell phone recordings had a rate of 22kHhz and were band-pass filtered during transmission, such that frequencies above 4-5kHz were removed. All stimuli were monophonic. Continua were created using STRAIGHT morphing software for MATLAB (STRAIGHT; Kawahara et al. 2008), such that 0% represented the neutral utterance and 100% represented the emotional utterance (happy or angry). Stimuli were morphed from 0% to 100% in 10% increments, resulting in eleven 'steps'. To prevent clipping and to approximately match approximately, stimuli were scaled to the same duration while maintaining pitch, and scaled to the lowest intensity of all sound files, roughly ~62 dB, using PRAAT.

Two female research assistants and one female theatre major gave informed consent before recording speech stimuli for this experiment. The theatre major (aged 21) recorded a phrase that was semantically neutral (e.g., "The coffee is vanilla"), which was then morphed on two affective continua corresponding to Neutral-Angry and Neutral-Happy. The practice stimuli were designed to orient participants to the task. The two female research assistants (ages 22 and 26) recorded the experimental stimuli. To strengthen the ecological validity of the stimuli (and because subsequent research on mothers of children with autism will utilize amateur speakers), the speakers were non-trained research assistants. Speakers were asked to imagine speaking the stimuli in an angry or a joyful mood. They uttered the phrase "Mr. Anderson" (/*mis-ter-AN-der-sin*/). This phrase is a semantically neutral common name and could be spoken in angry or happy intonations, depending on context. Previous research with morphed emotional speech also used a name (Doi et al., 2013; Hammerschmidt & Jürgens, 2007). The phrase was simultaneously recorded on the microphone and a cell phone call (iPhone to iPhone).

Voice morphing was performed using STRAIGHT (Kawahara et al., 2008) in Matlab (The MathWorks) in the same manner as described by Bestelmeyer and colleagues (2010). This algorithm allowed us to spectrally smooth the pitch between two exemplars, taking into account both source characteristics (including F0) and supralaryngeal filter characteristics (spectral peak distribution, including F1). When using STRAIGHT, the researcher can individually manipulate five parameters: F0, frequency, duration, spectrotemporal density, and aperiodicity, by manually identifying time-frequency landmarks. Using these landmarks, STRAIGHT interpolated intermediate values based on the number of continuum steps specified by the researcher (in this study, nine points internal to the continuum, and two points outside, for a total of eleven 'steps') and synthesized stimuli to form a smooth continuum.

Acoustical Analyses

Acoustic measurements were extracted from all stimulus files. Measurements were made using the "Quantify Source" script in the GSU PRAAT Tools script package with the exception of mean F1 which was extracted using the "Quantify Emotion" script (Owren, 2008). For each utterance, the measurements were made over the entire file. Measures of fundamental frequency were constrained to 100-450 Hz in order to promote stability in this measurement. The decision to constrain the files to 100-450 Hz stems from three justifications: first, the stimuli were entirely recorded by female talkers and literature suggests the average F0 of female talkers is near 200hz; second, the highest pitch and lowest pitch of the original pre-manipulation recordings were within this window; and third, the morphing procedure and the degraded quality of the cell phone recording combined to create artifacts that were often in the 400-600 Hz range. Although this window still likely retains some of these artifacts, we see it as a compromise between reducing

the measurement error and retraining the original signal, which allows for consistency and reproducibility for other researchers.

Speech Perception Task Procedures

Participants sat at a computer with a Cedrus RB730 response box with seven buttons in a row, where the leftmost button was blue, and the rightmost was yellow; the remaining (unused) buttons were gray. Subjects wore Bose QuietComfort 15 Noise-Cancelling headphones; the volume was consistent across participants. The task was presented via SuperLab 5.0.3 experimental software. Trials consisted of a 500ms crosshair, followed by the appearance of the printed words "Neutral" on the left side of the screen in blue font, and "Happy" or "Angry" (depending on trial) on the right side in yellow font. There was a simultaneous presentation of the acoustic stimulus with the printed words. The font color of the printed cues matched the color of the buttons that subjects pressed to indicate their responses.

Training. Before beginning the experiment proper, participants were shown how to respond using the colored buttons to indicate whether a presented sound was neutral or emotional (happy or angry). Participants were instructed that they would be judging voices for emotional expression and were oriented to the response box. Next, they heard clearly neutral (0%) and clearly emotional (100%) stimuli produced by the training voice actor. If they gave an incorrect response, the trial was repeated. This was done to ensure participants understood what was meant by the terms angry, happy, and neutral, as well as to orient them to the use of the buttons; only 4 participants failed on the first attempt and required repetition. Next, participants read, "Sometimes it can be hard to tell what someone is feeling in their voice," before beginning discrimination trials. They completed 22 trials, corresponding to both the Happy and Angry continua of semantically neutral sentences (e.g., "The coffee is vanilla"). They received no

feedback in this phase, as there was no objectively correct answer. They were instructed to "answer as quickly as possible without making mistakes" and that their "job in this experiment is to judge how well someone communicates emotion in their voice." This second training sequence was included to ensure that participants did not think there was a "right" answer and would not expect feedback.

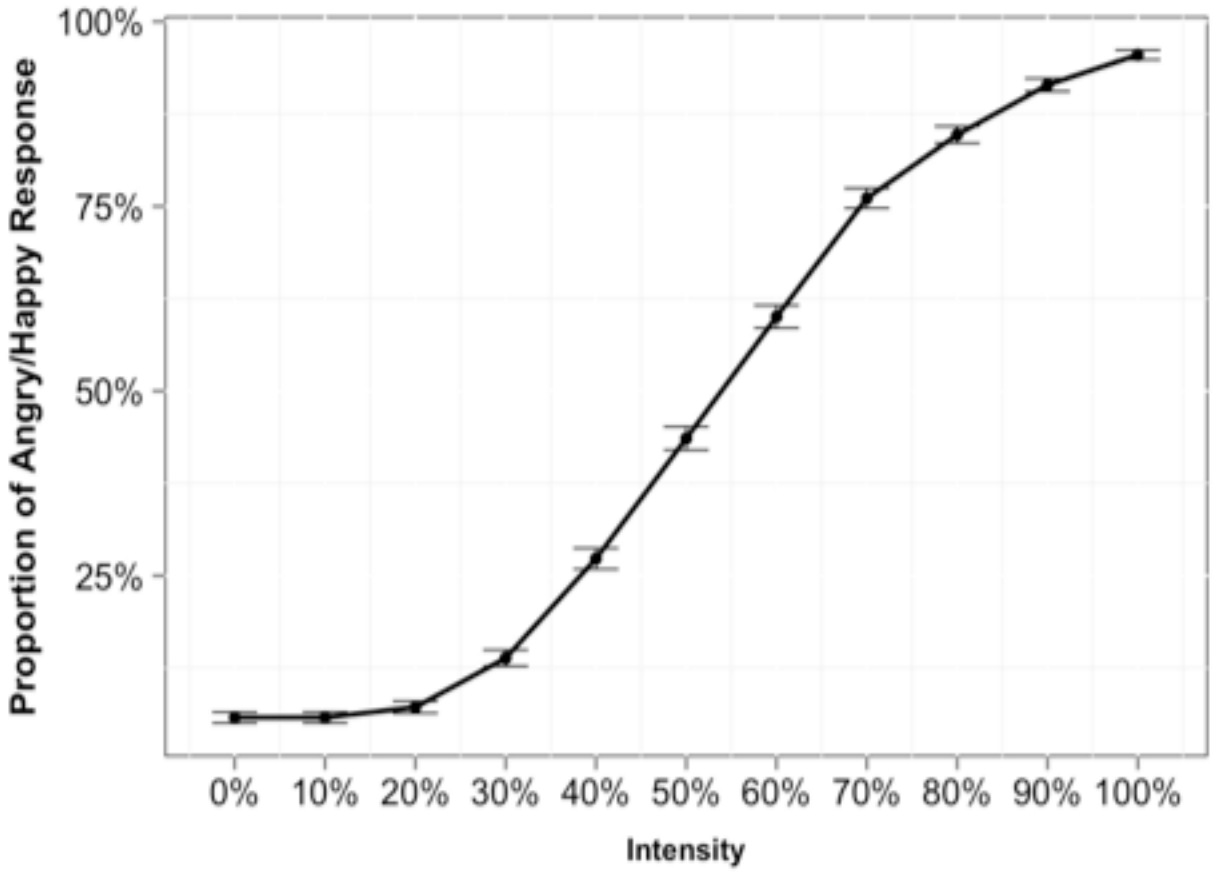
Emotional Judgment Task. After training was completed participants started the experiment. The experiment consisted of four blocks: [Talker A: Happy]; [Talker A: Angry]; [Talker B: Happy]; [Talker B: Angry]. Each block contained eleven trials (one for each step of the continuum). Each of the four blocks was presented using a microphone recording and a cell phone recording. There were eight blocks in total (four blocks x two recording conditions), each of which contained eleven trials corresponding to the eleven steps of the 0-100% continua (e.g., 0-100% Happy). Each block was presented four times in random fashion. Thus, there were a total of 352 trials (eight blocks x eleven recordings x four presentations). The presentation of the trials and of the stimuli within each trial was randomized. Participants were offered a break after every two blocks. The total experiment, including breaks, lasted 20-30 minutes.

Results

First, the data were checked for internal validity (i.e., that participants rated low intensity stimuli as neutral and high intensity stimuli as happy/angry), by conducting a logistic regression predicting response from intensity collapsed across all conditions. Logistic regression analysis allowed for the prediction of a dichotomous outcome variable, such as participant response in the current study. The results of a logistic regression are most easily understood when the output is transformed into odds ratio (OR; the likelihood of an outcome based one or more predictors), and confidence intervals. If the confidence interval contains 1.0, than there is no significant

difference in the likelihood of either outcome. This initial regression was significant, OR =747, 95% CI [594:939], indicating that as the stimuli increased in intensity, participants were more likely to rate them as happy or angry. The results follow an 'S' curve typical of categorical perception research (See Figure 2).

FIGURE 2. *Proportion of Happy/Angry response as a function of Intensity. Error bars are standard error of the mean.*



The y-axis is the mean proportion of "1" vs. "0" button presses where "1" signifies the participant rated that event as happy for a happy trial or angry for an angry trial. A "0" therefore signifies they rated that event as 'neutral'. The x-axis plots the intensity level of the stimuli. Each point plotted at each intensity level represents the mean across all participants for that intensity level. Greater intensity refers to greater expressed affect in the stimulus (i.e. more 'anger' or 'happiness'). Error bars are the standard error of the mean. A logistic regression predicting response from intensity was significant, OR =747, 95% CI [594:939].

Cell Phone vs. Mic: Emotional Judgment

Next, all four predictors (Talker, Affect, Recording Type, and Intensity level) were entered into a logistic regression predicting participant response to test all main effects and

interactions. The results can be found in Table 1. Results are broken down into main effects and interactions.

TABLE 2: Results of a logistic regression predicting participant Response from Talker, Affect, Intensity and Recording Type.

Predictor	OR	Lower CI (95%)	Upper CI (95%)
Intercept	.07	.05	.10
Talker (B) *	.12	.07	.21
Affect (Happy) *	.39	.24	.63
Recording Type (Phone)	1.15	.76	1.75
Intensity *	282.06	160.41	495.99
Talker x Affect *	3.76	1.65	8.58
Talker x Recording Type	1.36	.63	2.98
Affect x Recording Type *	.31	.15	.66
Talker x Intensity *	4.53	1.80	11.40
Affect x Intensity *	6.36	2.53	16.00
Recording Type x Intensity	1.07	.48	2.38
Talker x Affect x Recording Type	2.12	.65	6.91
Talker x Affect x Intensity	.37	.09	1.52
Talker x Recording Type x Intensity	.46	.13	1.63
Affect x Recording Type x Intensity	3.33	.83	13.38
Talker x Affect x Rec. Type x Intensity	.53	.07	4.04

Results from a logistic regression, transformed to Odd Ratios (OR) predicting participant response from Talker, Affect, Recording Type, and Intensity. An OR that does not include 1.0 in its confidence interval indicates a significant effect, p =0.05, indicated by an asterisk (*).

Main effects. There was a significant main effect of Talker. Compared to Talker A, participants were 12% less likely to rate Talker B as Happy/Angry (OR = 0.12, 95%CI [.07, .21]). This could be interpreted to mean that Talker A was more "expressive." There was a main effect of Intensity (OR=282.06, 95%CI [160.41,495.99]). This indicates that as the level of intensity rose, so too did the likelihood that a participant would press the emotion button. There was no main effect of Recording Type (OR = .97, 95%CI [.90, 1.05]). There was a significant main effect of Affect. Compared to Angry stimuli, participants were 39% less likely to press happy in the happy conditions than angry in the angry conditions (OR = .39, 95%CI [.24, .63]). Thus, the Angry stimuli were overall more expressive, however this effect is less meaningful because of a disordinal interaction found between Affect and Talker, reported below.

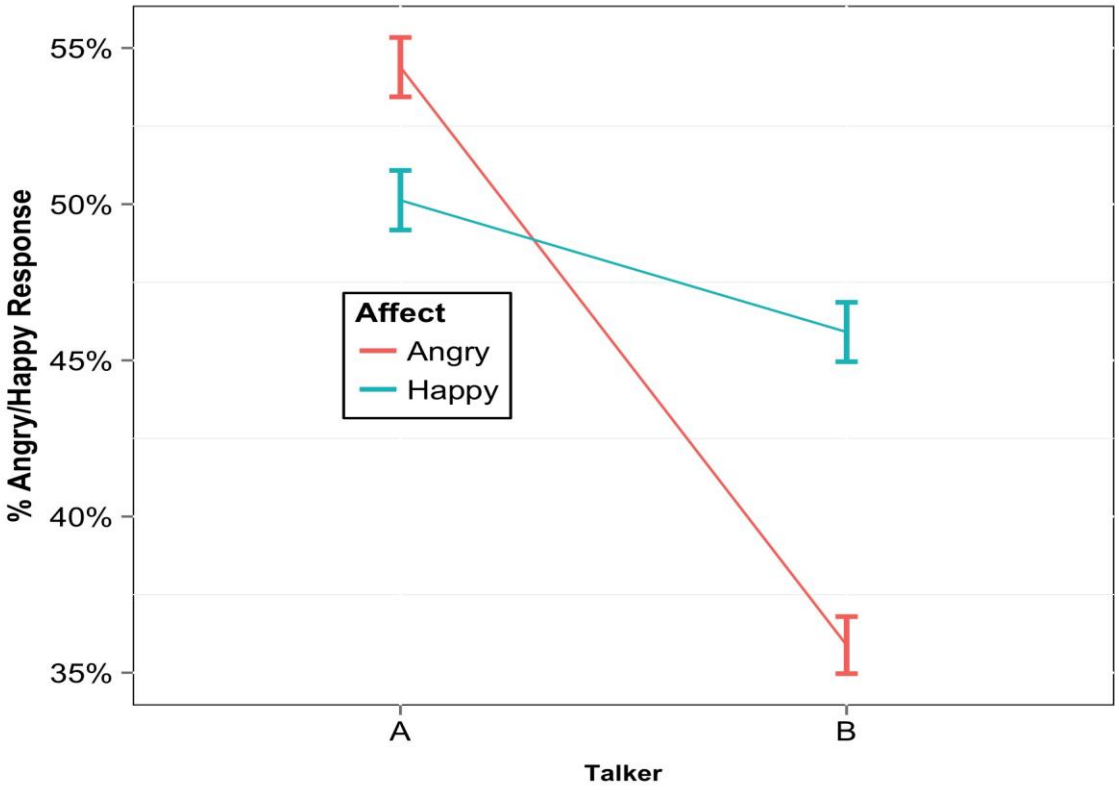
Interactions. There were no significant three-way or four-way interactions (see Table 3). Four out of the six two-way interactions were significant. There was a significant interaction between Talker and Affect (OR = 3.76, 95%CI [1.65, 8.58]), indicating that effect of Talker on participant response was dependent on their Affect. To determine the direction, Welch t-tests were performed. T-tests revealed that all comparisons were significantly different from each other after Bonferroni correction for multiple comparisons (see Table 3). These differences indicated that participants indicated hearing emotion most frequently in the Talker A-Angry block, followed by Talker A-Happy, Talker B-Happy, and finally Talker B-Angry. Thus, there was a disordinal interaction between Talker and Affect (See Figure 3).

TABLE 3: T-test results for Talker x Affect Interaction

Comparison	t	p	df
[Talker A, Happy] X [Talker A, Angry]	-3.16	.002	5496
[Talker A, Happy] X [Talker B, Angry]	10.78	<.001	5497
[Talker A, Happy] X [Talker B, Happy]	3.17	.002	5498
[Talker B, Happy] X [Talker A, Angry]	-6.31	<.001	5494
[Talker B, Happy] X [Talker B, Angry]	7.60	<.001	5496
[Talker A, Angry] X [Talker B, Angry]	14.03	<.001	5494

Results for each individual t-test of the means of participant response for each combination of Talker and Affect condition. Note that the degrees of freedom vary slightly as a result of some data values being removed due to outliers. All values are significant at the 0.05 level after Bonferroni correction ($p < 0.00833$).

FIGURE 3: *Proportion of Happy/Angry response as a function of Talker x Affect. Error bars are standard error of the mean. All comparisons are significantly different from each other (See TABLE 3).*

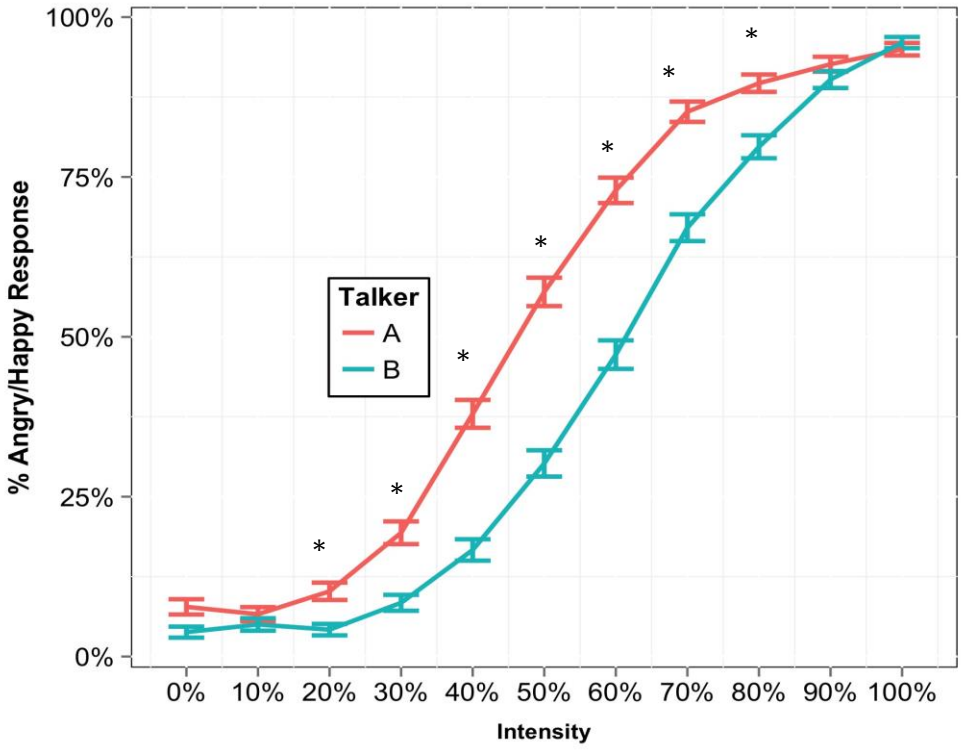


Talker x Intensity. There was a significant Talker x Intensity interaction; OR=4.53, 95%CI [1.80, 11.40]. This indicates that the effect of Talker on participant response depended on the level of Intensity. The results, plotted in Figure 4, show that across Intensity, Talker A was more expressive, with the exception of the ends of the continua, where there were floor and ceiling effects. T-tests at each level of intensity revealed that participant response was significantly greater at levels 20%-80% (See Table 4).

TABLE 4: T-test results for the Talker x Intensity Interaction

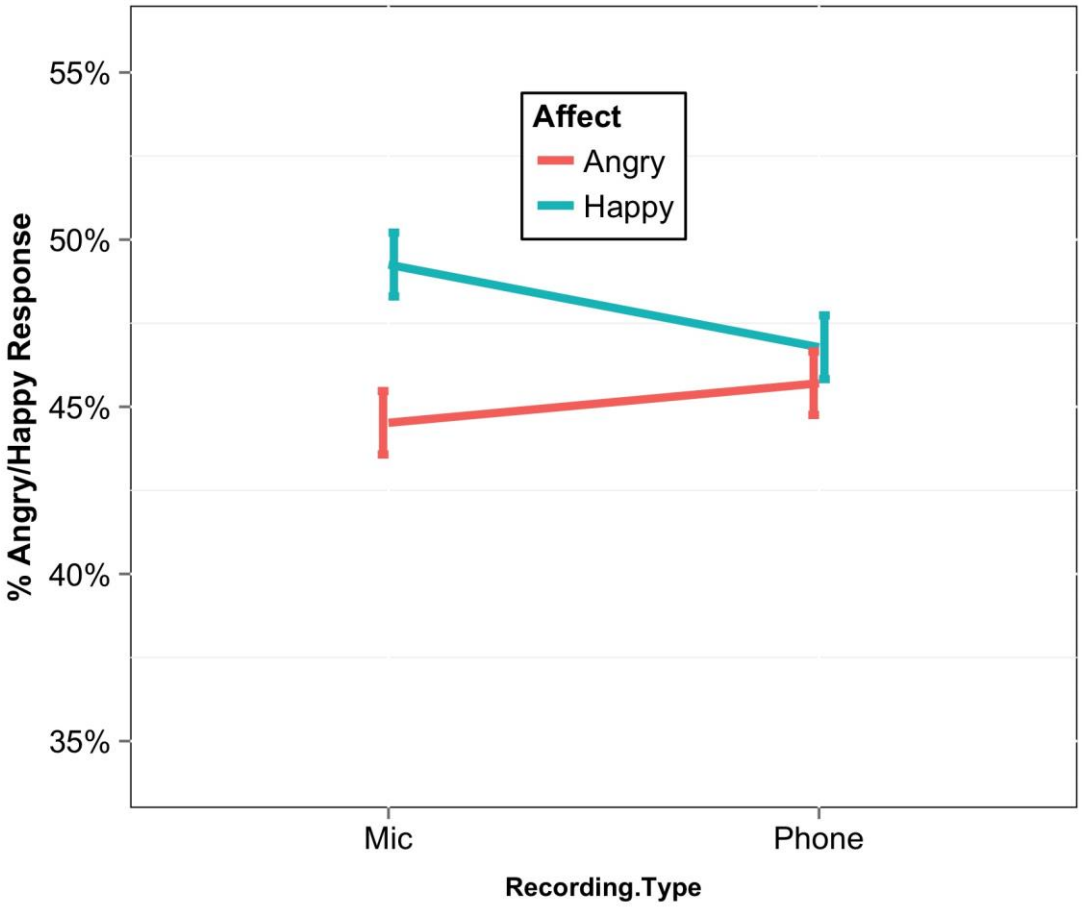
Intensity Level	t	p	df
0%	2.68	.007	908
10%	1.08	.280	980
20%	3.69 *	<.001	866
30%	5.06 *	<.001	887
40%	7.76 *	<.001	932
50%	8.87 *	<.001	990
60%	8.61 *	<.001	989
70%	6.90 *	<.001	936
80%	4.42 *	<.001	935
90%	1.35	.176	986
100%	-.79	.430	984

FIGURE 4. *Proportion of Happy/Angry response as a function of Talker x Intensity. Error bars are standard error of the mean. *Indicates significantly different at p < 0.0045*



*Bonferroni correction: .0045, *indicates significant at p< 0.0045*

FIGURE 5. *Proportion of Happy/Angry response as a function of Affect x Recording Type. Error bars are standard error of the mean.*



Affect x Recording Type. There was a significant interaction between Affect and Recording Type (OR = .31 95%CI [.15, .66]), indicating that the effect of recording type on participant response depended on affect (See Figure 5). To determine the direction, Welch t-tests were performed. T-tests revealed that all comparisons were significantly different from each other after Bonferroni correction for multiple comparisons. These differences indicated that participants indicated hearing emotion more frequently in the Happy, microphone-recorded stimuli than in the Angry stimuli on both the microphone and cell phone. All other comparisons were non-significant (See Table 5).

Table 5: T-test results for Recording Type x Affect Interaction

Comparison	t	p	df
[Mic, Happy] X [Mic, Angry]	3.52 *	<.001	5496
[Mic, Happy] X [Phone, Angry]	2.64 *	.008325	5508
[Mic, Happy] X [Phone, Happy]	1.84	.067	5498
[Phone, Happy] X [Mic, Angry]	1.68	.093	5494
[Phone, Happy] X [Phone, Angry]	0.80	.43	5506
[Mic, Angry] X [Phone, Angry]	-0.88	.38	5504

T-test results for each individual t-test of the means of participant response for each combination of Recording Type and Affect condition. Note that the degrees of freedom vary slightly as a result of some data values being removed due to outliers. All values are significant at the 0.05 level after Bonferroni correction (p must be less than 0.00833).

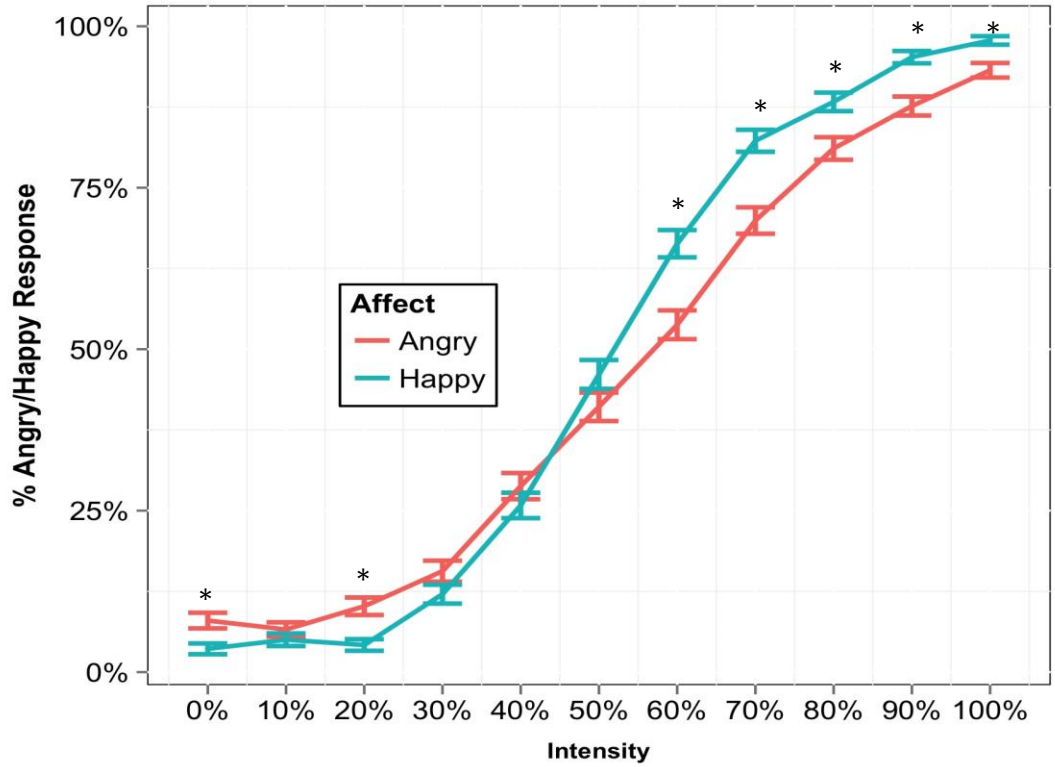
Affect x Intensity. There was a significant interaction between Affect and Intensity; OR = 6.36 95% CI [2.53, 16.00]. This indicated that the effect of Affect on participant response depended on the level of Intensity. The results, shown in Figure 6, suggest that there is a disordinal, logistic interaction, such that participants heard anger sooner than happiness at lower intensities in the microphone stimuli (e.g., intensity level 20%), whereas happiness appears to have been conveyed better at intensity levels above 60%. T-tests revealed that participants made more emotional judgments for Anger at intensity levels 0%, 20%; but made more emotional judgments for Happy at all levels above 60%. The results of these t-tests can be found in Table 6.

TABLE 6. *Affect by Intensity. Negative t-values indicate that Angry has a higher value on participant response.*

Intensity level	t	p	df
0%	-2.97*	.003	887
10%	-1.07	.287	981
20%	-3.69*	<.001	866
30%	-1.63	.104	983
40%	-1.06	.290	993
50%	-1.60	.111	996
60%	4.09*	<.001	999
70%	4.63*	<.001	970
80%	3.19*	.002	966
90%	4.32*	<.001	859
100%	3.54*	<.001	800

Bonferroni correction: 0.0045, *indicates significant at $p < 0.0045$

FIGURE 6. *Proportion of Happy/Angry response as a function of Affect x Intensity. Error bars are standard error of the mean. *Indicates significantly different at p< 0.0045*



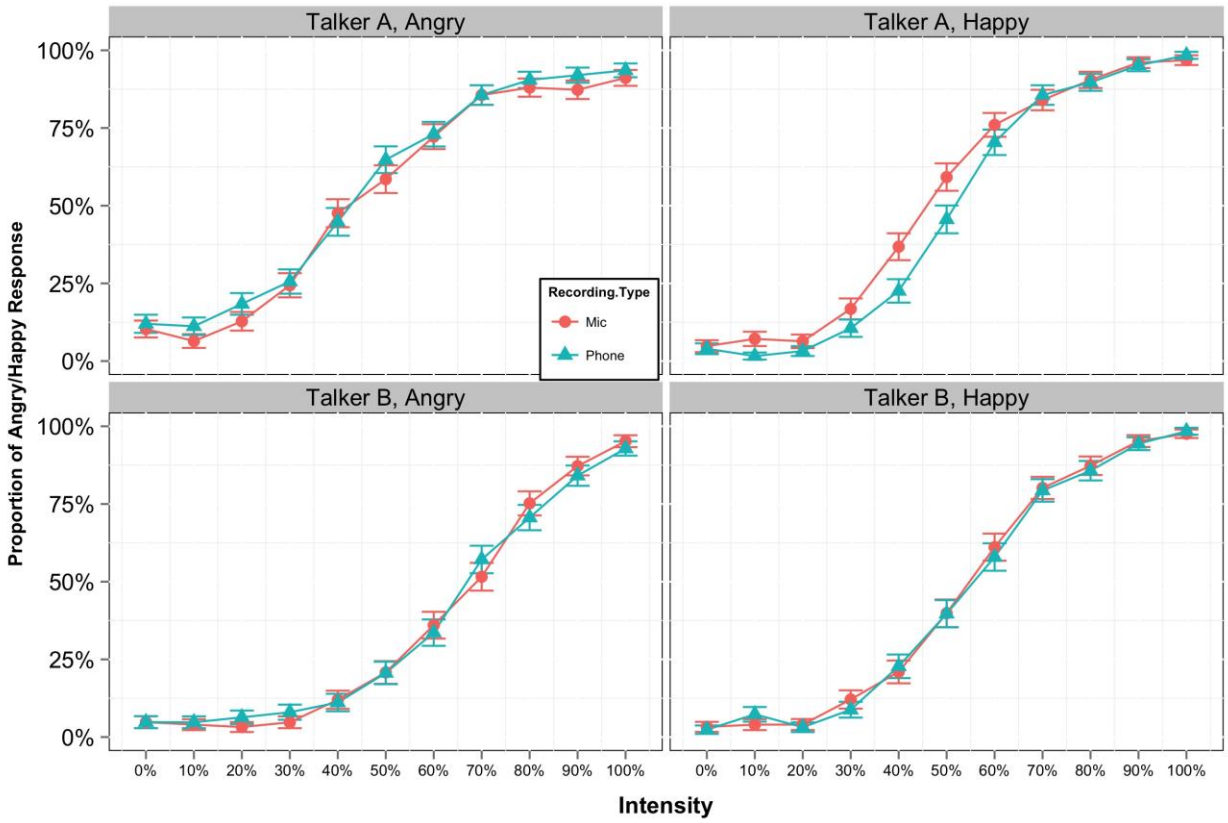
Next, we asked whether Affect and Talker differences were primarily driving the effect of the significant two-way interactions, since one of them was common to all of them. Different talkers, especially non-actors, are likely to have different levels of expressiveness; likewise, the different emotional expressions (i.e., happy and angry) are likely to have different levels of expressiveness because of variable levels of acoustic cues. Indeed, happiness and anger are expressed with different pitch (higher in happiness or joyful emotions) (Belyk & Brown, 2014). Our data also showed a similar pattern in F0 and F0 contour: the happy stimuli were greater and more variable in F0 than anger (see Figures 10-17), and overall F0 for happy stimuli was greater than angry across each intensity level (see Figure 7). Thus, the analysis controlled for the effect of both Affect and Talker to see whether recording type still had a significant effect on participant response.

The reader might ask, however, whether it is also important to control for Intensity. We reasoned that emotion in the voice may have been somewhat masked over the cell phone at different levels of intensity – especially the lower and middle levels. Moreover, the preliminary plots that include intensity (e.g., Figures 4 and 6) suggest that this is a possibility – the perception of emotion in the voice may be more challenging on the phone at low or ambiguous levels, but not at high intensity levels. On the other hand, it is possible that this effect does no differ by recording type, and that these effects are largely driven by affect and talker differences. Therefore, to test whether recording quality had a significant effect on participant response, a logistic regression was conducted to predict participant response from Intensity and Recording Type, while controlling for the effect of Talker and Affect. As expected, there were significant main effects for Talker (OR=0.40, 95%CI [.36, .45]); Affect (OR=1.25, 95%CI [1.12, 1.39]); and Intensity (OR=952, 95%CI [683,1328]). Consistent with the hypothesis that Recording Type would not impact participant response, the results revealed no main effect of Recording Type (OR=.94, 95%CI [.72, 1.23]) and no significant interaction between Recording Type and Intensity (OR=1.03, 95%CI [.65, 1.63]), after controlling for the effects of Affect and Talker (See Figure 8).

Finally, a visual plot of the data suggested a possible result that challenged our primary hypothesis. Specifically, in the set of data that corresponds to the Happy stimuli for Talker A, there was a separation between recording types (See Figure 8). To test whether this difference was significant, we conducted a logistic regression predicting participant response from Recording Type and Intensity for that condition only. The results indicated a significant Recording Type by Intensity interaction: participants were more likely to press the “happy” button for the microphone stimuli over the cell-phone stimuli; OR = 3.55, 95% CI [1.14:11.04].

To determine which intensity levels were significantly different, we ran t-tests on levels 40% and 50%. The results revealed that both were significant; participant response was significantly higher for the microphone stimuli at the 40% level (t (242.66)=2.48, p=.01), and the 50% level (t (247.96)=2.16, p =.03).

FIGURE 8. Relationship between participants' response at each intensity level as a function of recording type for each condition of talker and affect.



Starting in the upper left and moving clockwise, the first graph represents judgments of affect for Angry stimuli from Talker A, followed by Talker A's Happy stimuli, followed by the Happy stimuli for Talker B, and finally the Angry stimuli for Talker B. The axes plot judgment of affect by Intensity of stimulus for each recording type. Greater intensity refers to greater expressed affect in the stimulus (i.e., more 'anger' or 'happiness'). Error bars are the standard error of the mean. A logistic regression predicting response from intensity and recording type was significant for the [Talker A-Happy] block, OR = 3.55, 95% CI [1.14:11.04].

Reaction Time: Effect of Ambiguity

To test whether emotion judgments were most susceptible to or influenced by ambiguity, we collected RT data. Since judgment should be "hardest" for the most ambiguous stimuli at the middle of a continuum, we expected that the RT would be best fit by a quadratic model, and

would not differ as a function of recording quality. Because of dependency in the data resulting from multiple responses to the same stimuli by each individual, multi-level logistic modeling was used. As suggested by Baayen, Davidson, & Bates (2008), a t-value > |2| is considered significant. Results indicated that recording quality was a significant predictor of RT, with slower RT overall for cell phone recordings (β = 38.77, SE = 12.16, t = 3.19). See Figure 9 for a graph of this relationship. Table 7 lists the estimates, standard errors, and t-values for the predictors from the multi-level model.

TABLE 7: *Multilevel model results*

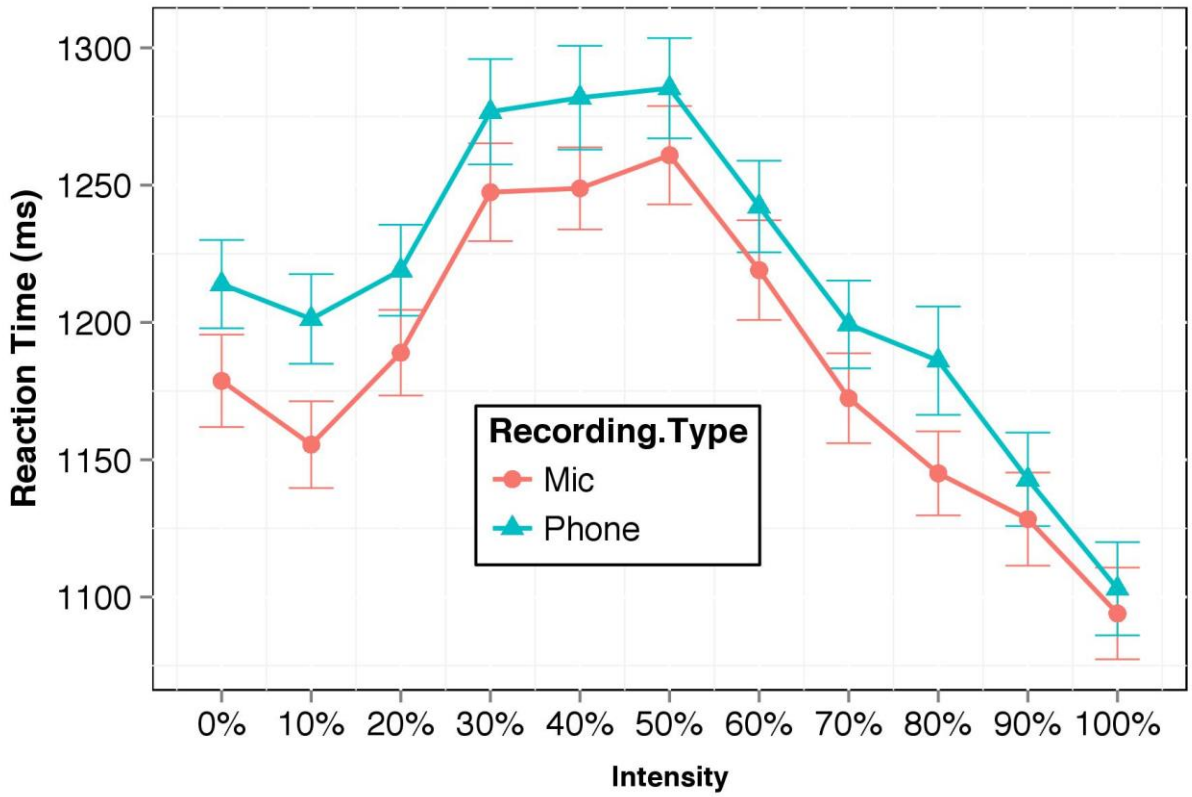
Fixed effects	**Estimate (ms)**	**Standard Error**	**t**
(Intercept)	1154.69	28.49	40.53
Recording Quality	38.77	12.16	3.19 *
Intensity	384.51	39.54	9.73 *
Intensity^2	-462.00	36.77	-12.57*
Interaction	-20.90	20.53	-1.02

**As suggested by Baayen, Davidson, & Bates, (2008), a t-value > |2| can be considered significant. The results here indicate that both the linear model and the quadratic model account for a significant amount of the variance in reaction time. A chi-square goodness-of-fit test comparing the linear and quadratic models revealed that the quadratic model explained significantly more variance than the linear model.*

Both the linear and quadratic models explained a significant amount of the variance in reaction time (linear: β = 384.51 SE = 39.54, t = 9.73 quadratic: β = -462.00, SE = 36.77, t = -12.57). A chi-square test of goodness-of-fit revealed that the quadratic model was a significantly better fit than the linear model, χ^2 = (1, N = 11,006) = 156.8, p < 0.001. Finally, the interaction

between intensity and recording quality was not significant, β = -20.65, SE = 20.67, t = -1.00.

FIGURE 9. Relationship between participants' RT at each intensity level for each recording type. Error bars are the standard error of the mean.



Reaction Time plotted against Intensity of stimulus for each recording type. Greater intensity refers to greater expressed affect in the stimulus (i.e., more 'anger' or 'happiness'). Error bars are the standard error of the mean. RT was significantly slower for the cell phone stimuli, β = 38.77, SE = 12.16, t = 3.19. The data were best fit by a quadratic model: β = -462.00, SE = 36.77, t = -12.57.

To summarize, participant response can be thought of as an indication of the expressivity of a particular stimulus. For example, if one talker is more expressive than another, than participants making judgments on that talkers' stimuli would be more likely to press the Happy/Angry button than the Neutral button for a given stimulus. Using this interpretation, our results indicated that Talker A was more expressive overall (i.e., there were more Happy/Angry button presses for Talker A's stimuli than for Talker B), this was true even when the effects of Affect and Intensity were included. The results also indicated that this talker effect was strong enough to cause a disordinal interaction between Talker and Affect, such that Anger was rated as

more expressive overall than Happiness, but only for Talker A, while the reverse was true for Talker B. An examination of recording type reveals no significant differences with the exception of the interaction between Affect and Recording Type, such that the Happy Microphone stimuli were rated as more expressive than both sets of Angry stimuli (Microphone & Phone), with no other significant differences. This result was driven largely by differences in Affect. We also found that the effect of Affect on participant response differed across Intensity. At low levels (0% and 20%) anger was more expressive, while happiness was more expressive at levels above 60%. It could be that participants relied more on mean F0 at higher levels of intensity, and other cues at low levels of intensity; this idea is expounded upon further in the discussion section.

Finally, because we were primarily interested in whether the emotion in the voice was affected meaningfully by the difference between a cell phone recording and a microphone recording, it made sense to control for the differences found between talkers and emotions. The results of this study indicated that the cell phone stimuli were no more or less expressive than the microphone stimuli after controlling for talker and affect differences. However, a plot of the data suggested that for levels 40% and 50% in the happy stimuli of Talker A, there were differences between recording types. Thus, we also ran an analysis on this condition, and performed t-tests on intensity levels 40 and 50%. Here we found a significant interaction between recording type and intensity, which appears to be largely driven by differences in the middle of the continua (i.e. 40-50% intensity level). T-tests confirmed that participant response at 40% and 50% levels were significantly different (microphone > phone).

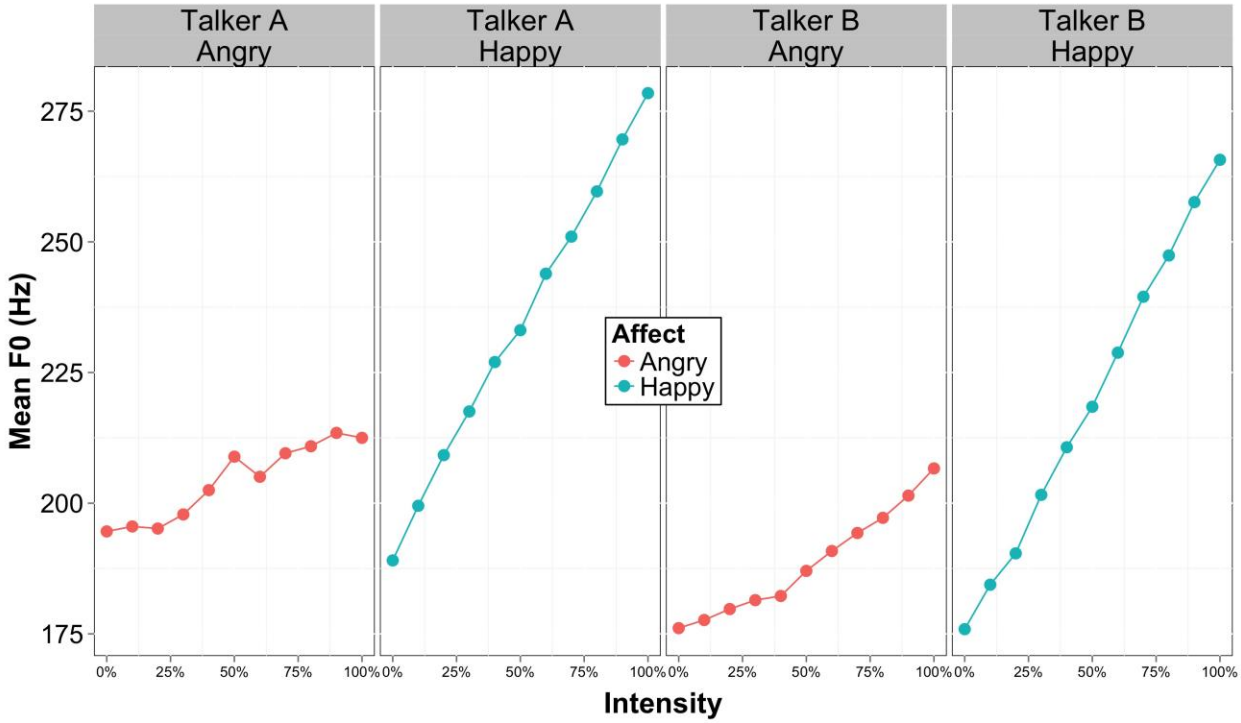
Discussion

Research on the perception of emotion in speech has grown exponentially over the last few decades (of 4,268 PubMed citations for the period of 1949-2015, 28% appeared within the

last five years); this increase may reflect both a growing interest in speech and technological advances. Many researchers may, for practical reasons, prefer to record speech via cell phone; however, little is known about the perceptual implications of this methodology. This study compared listener judgments of graded emotional speech created from cell phone transmission to recordings via high-quality microphone in four conditions, representing two talkers expressing two emotions. These data add to our understanding of the perception of emotion in speech by comparing two recording qualities and reporting measures of acoustic cues. Findings inform the research community about the viability of collecting speech data through cell phone transmissions.

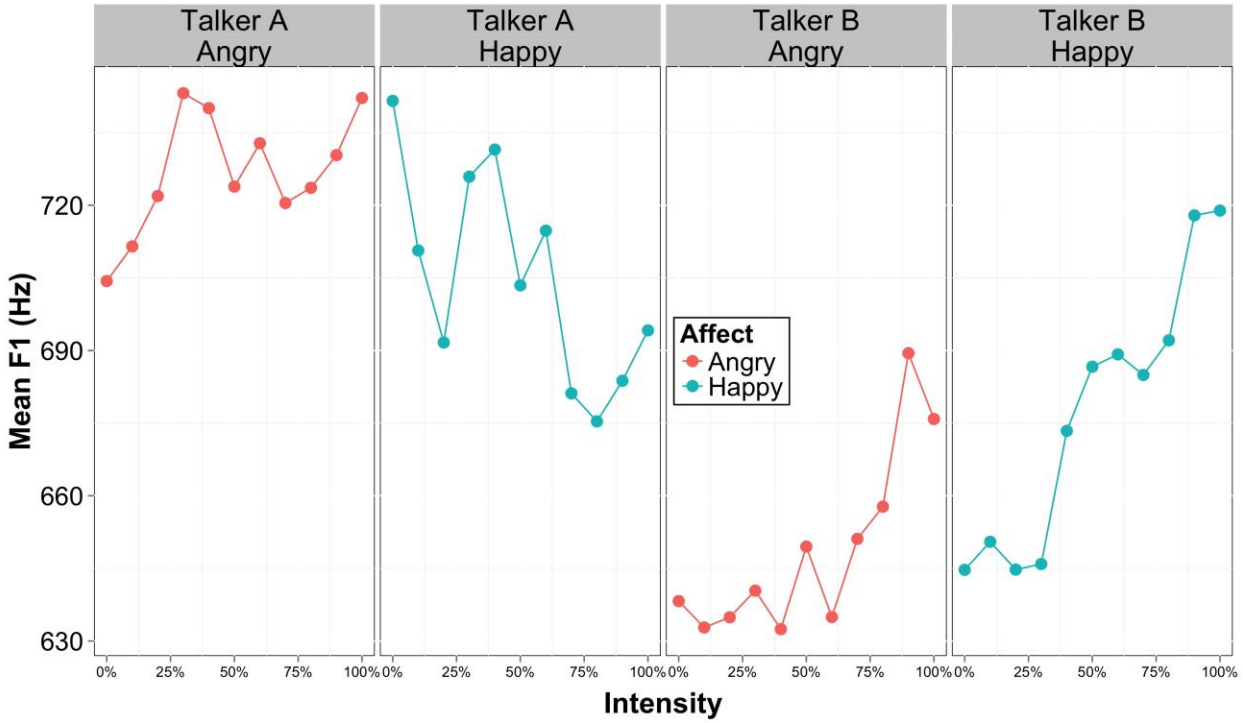
When we look at how participants responded to stimuli across talkers and emotions, we find that the two talkers differed in their level of expressivity (i.e., more participant happy/angry judgments) such that Talker A was more expressive than Talker B. This was true regardless of emotion (Figure 3). Within talkers, we found that participants rated the Angry stimuli of Talker A as more expressive than the Happy stimuli of Talker A; the reverse was true for Talker B. In previous research joyful stimuli – which are similar to Happy stimuli in the current study – had higher and more variable F0 than anger (Belyk & Brown, 2014). Our results were similar; the mean and standard deviation of F0 were greater for Happy stimuli as compared to the Angry stimuli across both talkers (See Figure 10 for mean F0).

FIGURE 10. Mean F0 of stimuli across Intensity levels for each condition.



This finding is largely driven by the increased expressiveness of Talker A, who may have simply "acted" anger better than she did happiness, and was also overall more expressive across both emotions than Talker B. To better describe this finding, we examined acoustic measures to see which may explain the increased "expressiveness" of Talker A. In Figure 10 we see that mean F0 appears to be somewhat higher for Talker A as compared to B, but it is unclear if these small differences are meaningful. Next, we looked at mean F1. Here we see more dramatic differences (Figure 11). First, Talker A has a higher mean F1 than Talker B overall; second, Talker A has a higher mean F1 for the Angry condition compared to the Happy, while the reverse is true for Talker B. This pattern mirrors the results of the disordinal Talker x Affect interaction reported earlier and thus may partially explain that finding (See Figure 3). Thus, it may be that listeners relied more on F1, or the combination of F0 and F1, than on F0 alone for their judgments.

FIGURE 11. Mean F1 of stimuli across Intensity levels for each condition.



With regards to our significant two-way interactions, we believe these interactions were significantly affected by the differences between the two talkers, which then caused unexpected differences in our affect variable. That is, while we would expect Happy stimuli to be rated as more expressive than the Angry stimuli, Talker A was so much more expressive that her angry stimuli overcame the effect of affect. This explains why Figures 3, 4 and 6 show their respective patterns. The last significant interaction (Figure 5) resulted in an interaction between Affect and Recording Type: Happy stimuli recorded on the microphone were rated as more expressive than Anger recorded on either recording type. This effect seems to stem from Talker A. We found that only for Talker A's Happy stimuli were there any significant differences between microphone and cell phone; the microphone recordings were more likely to be rated happy than the cell phone stimuli for intensity levels 40% and 50%. Thus, there is some evidence that the cell phone

was unable to retain important speech cues for the perception of happiness at the more ambiguous intensity levels.

Overall, an evaluation of how participants rated emotion in the human voice for speech recorded on a cell phone as compared to a high-quality microphone reveals that the primary hypothesis was largely supported: listeners' judgment of affect did not differ as a function of recording quality when compared within talker or within affect at every intensity level except for 40% and 50% in the Talker A-Happy condition (See Figure 8; quadrant I). It is unclear what caused this difference; the analyses suggest that F0 contour, F1, and HNR may be the most important influences. This does mute somewhat the blanket conclusion that cell phone recordings will yield similar results to microphone recordings in all cases. A conservative interpretation of our data suggests that at least high levels of intensity (i.e. >60%), the cell phone appears to be a reasonably good substitute for the microphone in speech research on emotion in the voice.

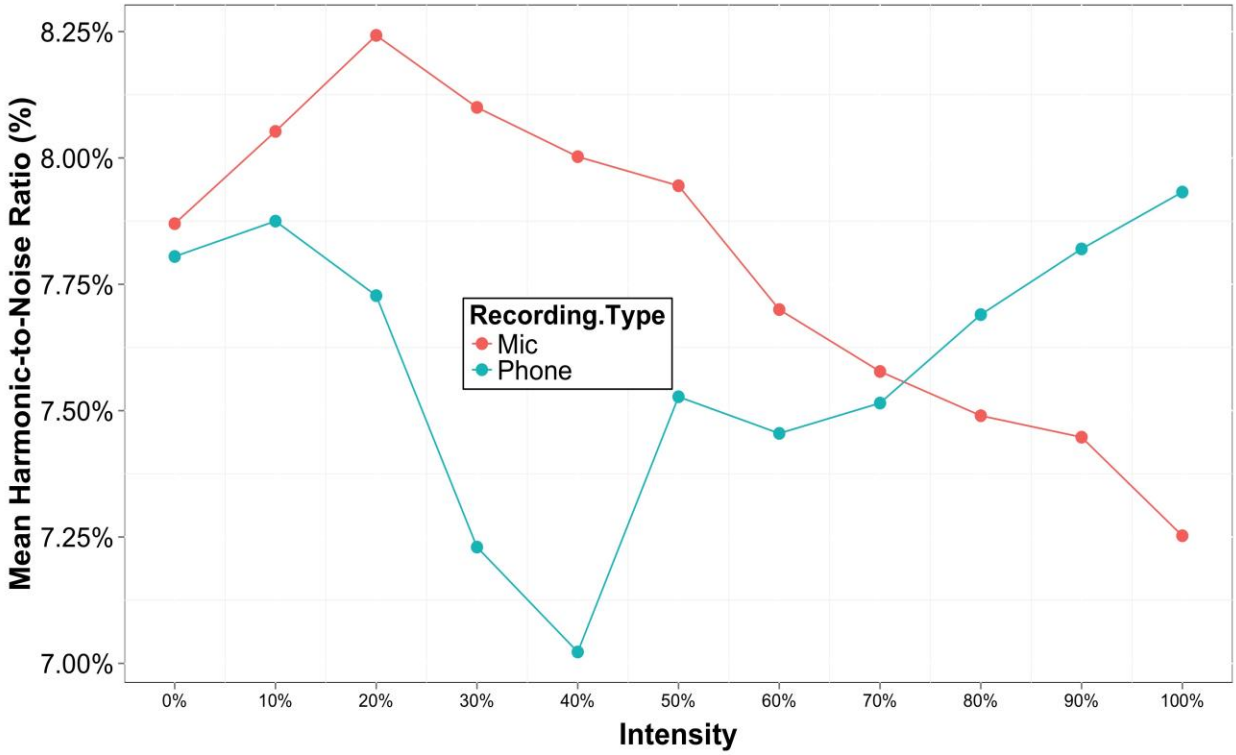
To better understand this finding, we extracted pitch statistics for our stimuli, including various measures of F0, HNR, jitter, shimmer, and F1. A visual examination of these acoustic parameters revealed no clear patterns of differences between the two recording types for the 40-50% intensity levels of Talker A-Happy. Instead, there are large differences across all stimuli between the recording types with the exception of mean F0. These widespread differences in multiple acoustic parameters likely reflect the lower bandwidth of the cell phone transmission, which disrupts some fine-grained acoustic cues. Because participants did not differ in their judgments overall across recording types, these differences may not be large enough to elicit a measurable effect, or these cues may not be critical to affect judgments.

Next, we looked at F0 contour. Previous research suggests that F0 contour is a large determinant of emotion category, above and beyond F0 values (Rodero, 2011). Thus, differences in the F0 contour for the two stimulus sets (microphone and cell phone) for the Talker A-Happy condition may explain the results. If this is the case, the F0 contours ought to be similar for the three blocks where no effect was found, and disrupted or dissimilar for the block with a significant difference. Each set of stimuli was plotted in PRAAT using the "draw pitch object" tool, which portrays the pitch contour values at each step of the 0%-100% intensity continua (see Figures 13-20). Although we do not report an objective measure of "amount of F0 contour disruption" across the different sets of stimuli, there appears to be at least one unique difference for the two sets of stimuli in question (Figures 15 & 16). A comparison of these two figures with the remaining figures indicates that there are disruptions of the red line in the cell phone stimuli, which corresponds to the most intense stimulus in that continuum. This may have affected participant judgment of the "happiest" stimuli, leading to fewer judgments of happy for the more ambiguous stimuli. That is, if the expressive "anchor" of the continuum was disrupted in the cell phone recordings, the set may have been less expressive overall. Alternatively, there may have been an interaction of multiple cues to cause this effect that were not detected in these acoustic analyses. Overall, no especially clear patterns emerged that explain the difference in that bloc. Future researchers should be vigilant in characterizing their stimuli both pre- and post-manipulation to check for unexpected or large acoustic differences.

An examination of reaction times revealed that participants responded significantly slower to the cell phone stimuli. Given that the cell phone stimuli were distorted in several acoustic dimensions, these distortions may have led to a processing cost. That is, participants responded more slowly because the signal was more ambiguous/noisy. Although participants

responded more slowly to the cell phone stimuli, they responded more slowly to the most ambiguous stimuli (middle of the intensity spectrum), regardless of recording type. Follow-up analyses probed the mean harmonic-to-noise (HNR) ratio, where a lower value indicates a noisier signal. As shown in Figure 12, the phone-recorded stimuli have a lower mean HNR, especially for the middle of the continuum.

FIGURE 12. Comparison of Mean HNR at each intensity level for the microphone and cell-phone stimuli.



Looking at the graph (Figure 9), one question is why RT early in the continuum may have been relatively slower than in the latter part of the continuum. This difference was significant (0% compared to 100%; $t(1991.1) = 5.88$, $p < 0.001$). Responses to the 0-10% stimuli may have been slower because this response reflected a judgment about the absence of a quality (i.e., 'not happy/angry'), whereas the intense stimuli involved an affirmative judgment (i.e., 'that sounds happy').

Limitations.

The current study is limited in several dimensions. First, the results may not generalize to emotions other than happy/angry. In addition, results may not generalize beyond the two individual talkers studied, however, it highlights how even two females of similar demographics (e.g., age, race) can nevertheless have very different levels of expression, and thus exert effects on research of this nature. Both the talkers and the participants were a convenience sample that has been described as homogeneous and over-studied (sometimes called "WEIRD": Western, Educated, Industrialized, Rich & Democratic) (Henrich, Heine, & Norenzayan, 2010). The results may also be limited in applicability to populations with a high degree of familiarity with cell phones. The design used a two-alternative forced-choice response; future studies might utilize an open-ended identification response, or choices between multiple emotions, to permit stronger conclusions regarding the transmission of emotion in speech through a cell phone signal and to provide more detail on which acoustic measurements correlate with listeners' decision making. Moreover, this study controlled for the duration of the stimuli, but research suggests that duration is an important variable in distinguishing between emotions, such as between 'rage/hot anger' and 'joyful surprise' (Hammerschmidt & Jürgens, 2006). Finally, the use of a single utterance, "Mr. Anderson," limits conclusions regarding the way emotion is conveyed in cell phone *conversations*. This approach does suggest that cell phone recordings of short utterances can be used in speech research on emotion. Future research should expand this design to include more emotional categories, more talkers, and more diversity of the talkers and listeners. Although not the focus of the current study, future research could manipulate different variables - or manipulate the same variables more precisely - to better discern which vocal cues are most

related to the perception of different emotions. This would allow a broader understanding of both the cues responsible for conveying affect and the interaction of the cues in that process.

The present study sought to test whether listeners would judge emotion in the voice equally for identical, graded emotional stimuli that differed only in recording quality; cell phone versus high-quality microphone. The results largely supported the use of a cell phone transmission for use in this manner. Although conclusions are limited to short, second-long utterances, they tentatively suggest that a cell phone transmission retains important vocal information that listeners use to make affective judgments. Researchers using this method can expect slower reaction times than would be found for higher-quality recordings, but fairly robust mean F0 and F0 contour preservation. Other variables, such as mean F1 and mean HNR, may be more affected by the recording quality and should thus be considered in methodological design.

References

Addington, D. W. (1968). The relationship of selected vocal characteristics to personality perception. *Speech Monographs*. http://doi.org/10.1080/03637756809375599

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. http://doi.org/10.1016/j.jml.2007.12.005

Babel, M., McGuire, G., & King, J. (2014). Towards a more nuanced view of vocal attractiveness. *PloS One*, *9*(2), e88616. http://doi.org/10.1371/journal.pone.0088616

Belyk, M., & Brown, S. (2014). The acoustic correlates of valence depend on emotion family. *Journal of Voice : Official Journal of the Voice Foundation*, *28*(4), 523.e9–523.e18. http://doi.org/10.1016/j.jvoice.2013.12.007

Bestelmeyer, P. E. G., Maurage, P., Rouger, J., Latinus, M., & Belin, P. (2014). Adaptation to vocal expressions reveals multistep perception of auditory emotion. *The Journal of Neuroscience*, *34*(24), 8098–105. http://doi.org/10.1523/JNEUROSCI.4820-13.2014

Bestelmeyer, P. E. G., Rouger, J., DeBruine, L. M., & Belin, P. (2010). Auditory adaptation in vocal affect perception. *Cognition*, *117*(2), 217–23. http://doi.org/10.1016/j.cognition.2010.08.008

Blood, G., Mahan, B., & Hyman, M. (1979). Judging personality and apperance from voice disorders. *Journal of Communication Disorders*, *12*, 63–68. http://doi.org/10.4319/lo.2013.58.2.0489

Bruckert, L., Liénard, J.-S., Lacroix, A., Kreutzer, M., & Leboucher, G. (2006). Women use voice parameters to assess men's characteristics. *Proceedings. Biological Sciences / The Royal Society*, *273*(1582), 83–9. http://doi.org/10.1098/rspb.2005.3265

Byrne, C., & Foulkes, P. (2004). The "mobile phone effect" on vowel formants. *International Journal of Speech, Language and the Law*, *11*. http://doi.org/10.1558/sll.2004.11.1.83

Doi, H., Fujisawa, T. X., Kanai, C., Ohta, H., Yokoi, H., Iwanami, A., … Shinohara, K. (2013). Recognition of facial expressions and prosodic cues with graded emotional intensities in adults with Asperger syndrome. *J Autism Dev Disord*, *43*(9), 2099–2113. http://doi.org/10.1007/s10803-013-1760-8

Eskenazi, L., Childers, D. G., & Hicks, D. M. (1990). Acoustic correlates of vocal quality. *Journal of Speech and Hearing Research*, *33*(2), 298–306. http://doi.org/10.1044/jshr.3302.298

Evans, S., Neave, N., & Wakelin, D. (2006). Relationships between vocal characteristics and body size and shape in human males: an evolutionary explanation for a deep male voice. *Biological Psychology*, *72*(2), 160–3. http://doi.org/10.1016/j.biopsycho.2005.09.003

French, P., & Howard, D. (1995). Studies in General and English Phonetics: Essays in Honour of Professor J.D. O'Connor, 1995 | Online Research Library: Questia. In J. Windsor Lewis (Ed.), *Studies in General and English Phonetics: Essays in Honour of Professor J.D. O'Connor* (pp. 230–240). New York: Routledge. Retrieved from https://www.questia.com/read/103842881/studies-in-general-and-english-phonetics-essays-

in

Hammerschmidt, K., & Jürgens, U. (2007). Acoustical correlates of affective prosody. *Journal of Voice : Official Journal of the Voice Foundation*, *21*(5), 531–40. http://doi.org/10.1016/j.jvoice.2006.03.002

Harnsberger, J. D., Shrivastav, R., Brown, W. S., Rothman, H., & Hollien, H. (2008). Speaking Rate and Fundamental Frequency as Speech Cues to Perceived Age. *Journal of Voice*, *22*(1), 58–69. http://doi.org/10.1016/j.jvoice.2006.07.004

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2-3), 61–83; discussion 83–135. http://doi.org/10.1017/S0140525X0999152X

Hughes, S., & Rhodes, B. (2010). Making age assessments based on voice: The impact of the reproductive viability of the speaker. *Journal of Social, Evolutionary, and …*, *4*(4), 290–304. Retrieved from http://shell.newpaltz.edu/jsec/articles/volume4/issue4/HughesVol4Iss4.pdf

Jong, G. De, Hudson, T., Nolan, F., & Mcdougall, K. (1995). The telephone effect on F0, *0*, 1–2.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (pp. 3933–3936). http://doi.org/10.1109/ICASSP.2008.4518514

Kawahara, H., Takahashi, T., Morise, M., & Banno, H. (2009). Development of exploratory research tools based on TANDEM-STRAIGHT. In *Proceedings : APSIPA ASC 2009* (pp. 111–120). Retrieved from http://eprints.lib.hokudai.ac.jp/dspace/handle/2115/39651

Künzel, H. J. (2001). Beware of the "telephone effect": the influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics*, *8*(1), 80–99. http://doi.org/10.1558/sll.2001.8.1.80

Lahey, M., & Bloom, L. (1988). *Language disorders and language development*. New YorkLondon: Macmillan ;Collier Macmillan.

Lass, N., Barry, P., Reed, R., Walsh, J., & Amuso, T. (1979). The effect of Temporal speech alterations on speaker height and weight identification. *Language and Speech*, *22*(2), 163–171.

Leemann, A., Kolly, M. J., & Dellwo, V. (2014). Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. *Forensic Science International*, *238*, 59–67. http://doi.org/10.1016/j.forsciint.2014.02.019

Mayo, R. (1994). Vocal fundamental frequency and vowel formant frequency characteristics of normal African- American and European-American adults. *Texas Journal of Audiology and Speech-Language Pathology*, *20*, 33–36.

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say "hello"? Personality impressions from brief novel voices. *PloS One*, *9*(3), e90779. http://doi.org/10.1371/journal.pone.0090779

McAllister, A., & Sjölander, P. (2013). Children's voice and voice disorders. *Seminars in Speech and Language*, *34*(2), 71–9. http://doi.org/10.1055/s-0033-1342978

Owren, M. J. (2008). GSU Praat Tools: Scripts for modifying and analyzing sounds using Praat acoustics software. *Behavior Research Methods*, *40*(3), 822–829. http://doi.org/10.3758/BRM.40.3.822

Protopapas, a, & Lieberman, P. (1997). Fundamental frequency of phonation and perceived emotional stress. *The Journal of the Acoustical Society of America*, *101*(4), 2267–2277. http://doi.org/10.1121/1.418247

Ptacek, P. H., & Sander, E. K. (1966). Age recognition from voice. *Journal of Speech and Hearing Research*, *9*, 273–277.

Rodero, E. (2011). Intonation and emotion: influence of pitch levels and contour type on creating emotions. *Journal of Voice : Official Journal of the Voice Foundation*, *25*(1), e25–34. http://doi.org/10.1016/j.jvoice.2010.02.002

Rossing, T. D. (1982). *The science of sound*. Addison-Wesley Pub. Co.

Skuk, V. G., & Schweinberger, S. R. (2013). Adaptation aftereffects in vocal emotion perception elicited by expressive faces and voices. *PloS One*, *8*(11), e81691. http://doi.org/10.1371/journal.pone.0081691

Teixeira, J. P., & Fernandes, P. O. (2014). Jitter, Shimmer and HNR Classification within Gender, Tones and Vowels in Healthy Voices. *Procedia Technology*, *16*, 1228–1237. http://doi.org/10.1016/j.protcy.2014.10.138

Traunmüller, H., & Eriksson, A. (1994). The frequency range of the voice fundamental in the speech of male and female adults. *Department of Linguistics, University of Stockholm*, *97*, 1905191–5.

van Dommelen, W. A., & Moxness, B. H. (1995). Acoustic parameters in speaker height and weight identification: sex-specific behaviour. *Language and Speech*, *38 (Pt 3)*, 267–87. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8816083

Walton, J. H., & Orlikoff, R. F. (1994). Speaker race identification from acoustic cues in the vocal signal. *Journal of Speech and Hearing Research*, *37*(4), 738–45. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7967558

Wilcox, K. A., & Horii, Y. (1980). Age and changes in vocal jitter. *Journal of Gerontology*, *35*(2), 194–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7410776

Wolf, R., & Murray, H. a. (1937). An Experiment in Judging Personalities. *The Journal of Psychology*, *3*(2), 345–365. http://doi.org/10.1080/00223980.1937.9917506

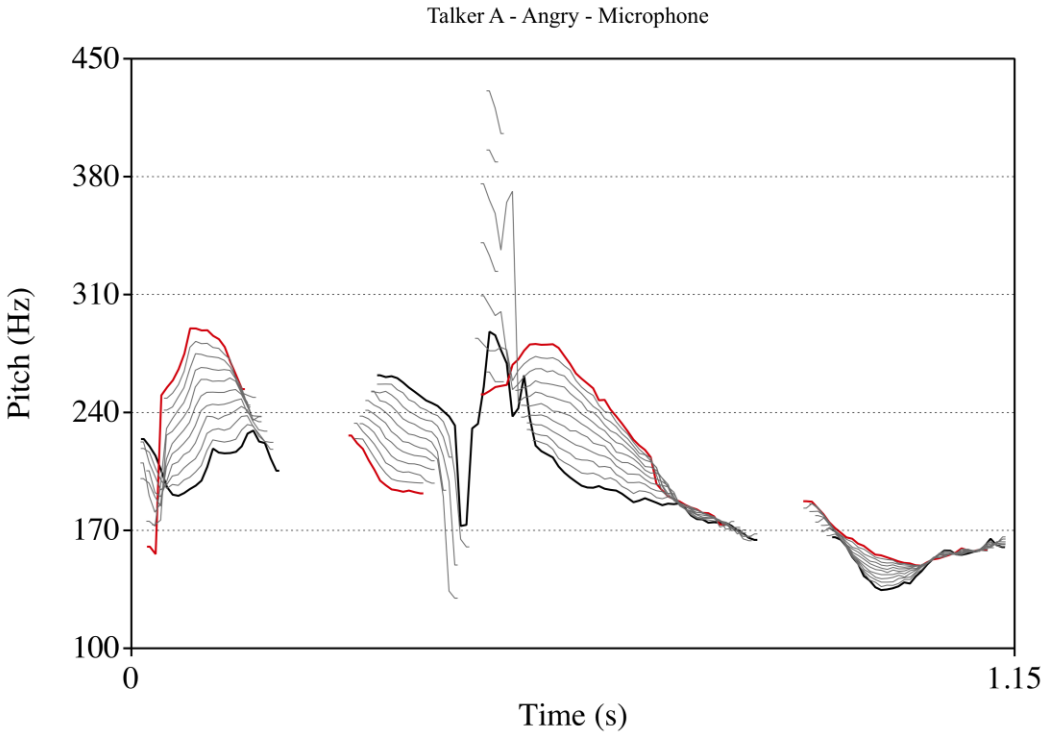
Zäske, R., Skuk, V. G., Kaufmann, J. M., & Schweinberger, S. R. (2013). Perceiving vocal age and gender: An adaptation approach. *Acta Psychologica*, *144*(3), 583–593. http://doi.org/10.1016/j.actpsy.2013.09.009

Zuckerman, M., & Driver, R. E. (1988). What sounds beautiful is good: The vocal attractiveness stereotype. *Journal of Nonverbal Behavior*, *13*(2), 67–82. http://doi.org/10.1007/BF00990791

Appendix

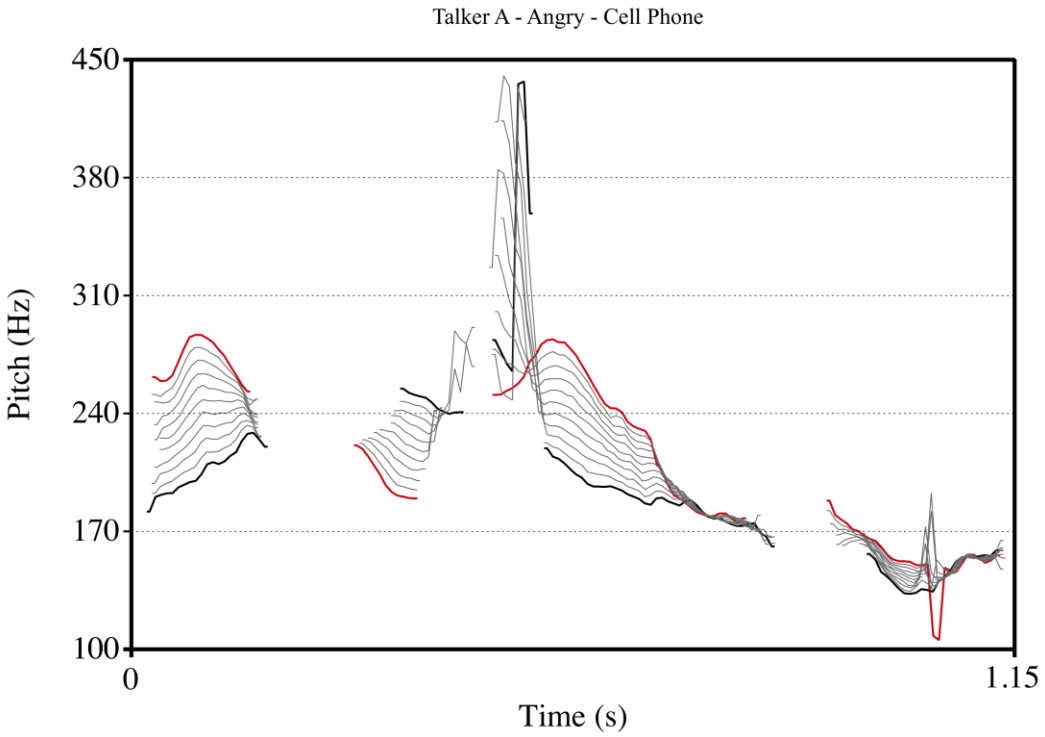
Pitch Contour Plots (f0)

FIGURE 13: Pitch contour graph for stimuli from Talker A, Angry, Microphone



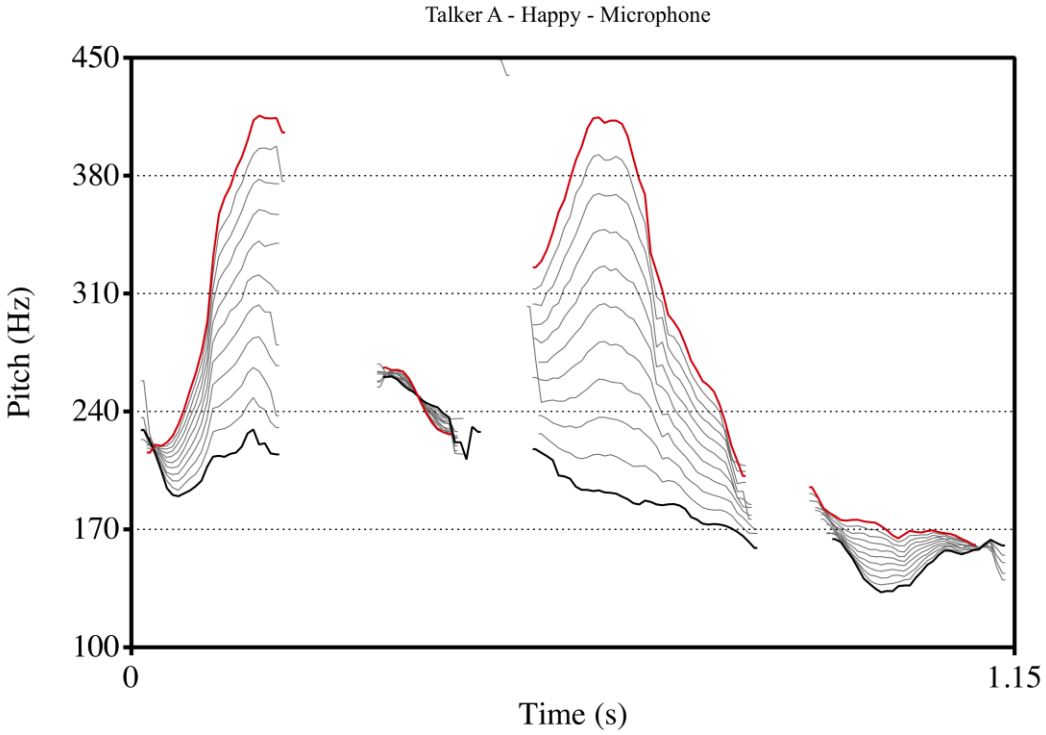
Black represents 0% angry; Red represents 100% angry; the grey lines represent the individual morph steps for the morphed stimuli.

FIGURE 14 Pitch contour graph for stimuli from Talker A, Angry, Cell Phone



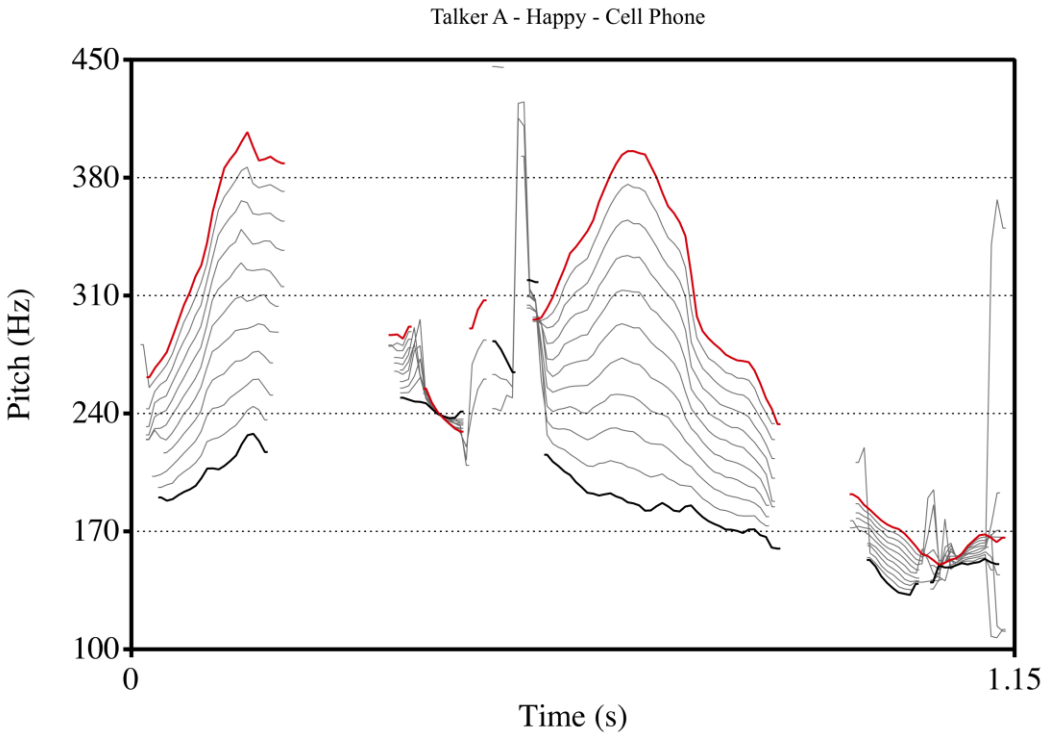
Black represents 0% angry; Red represents 100% angry; the grey lines represent the individual morph steps for the morphed stimuli.

FIGURE 15 Pitch contour graph for stimuli from Talker A, Happy, Microphone



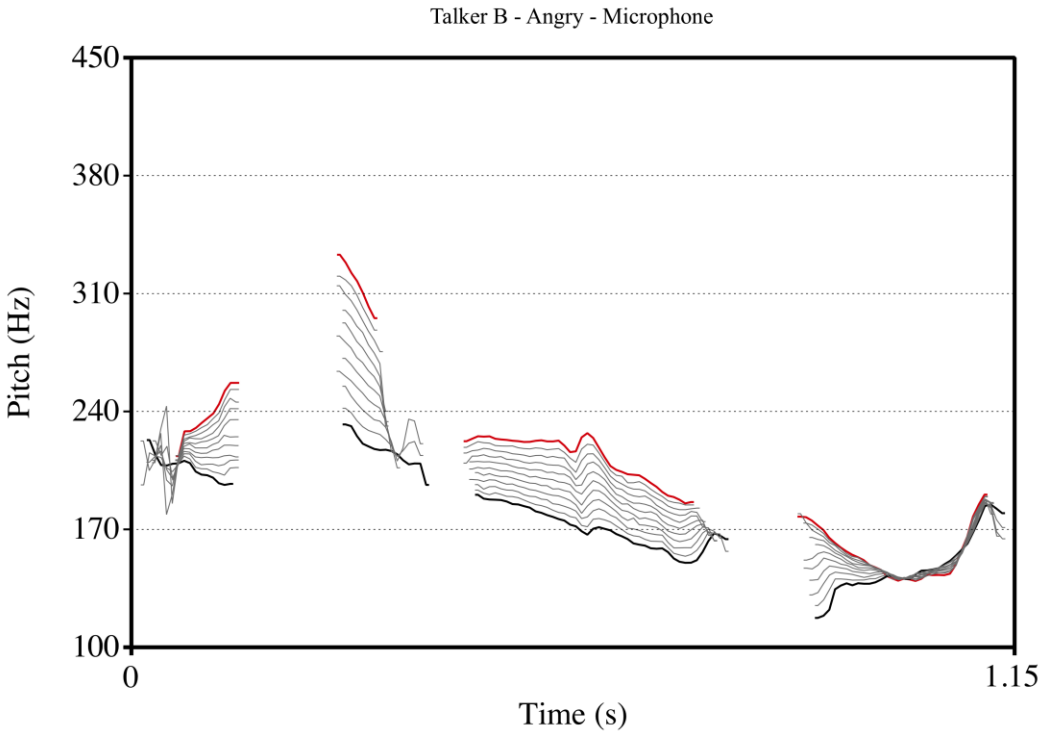
Black represents 0% angry; Red represents 100% angry; the grey lines represent the individual morph steps for the morphed stimuli.

FIGURE 16 Pitch contour graph for stimuli from Talker A, Happy, Cell Phone



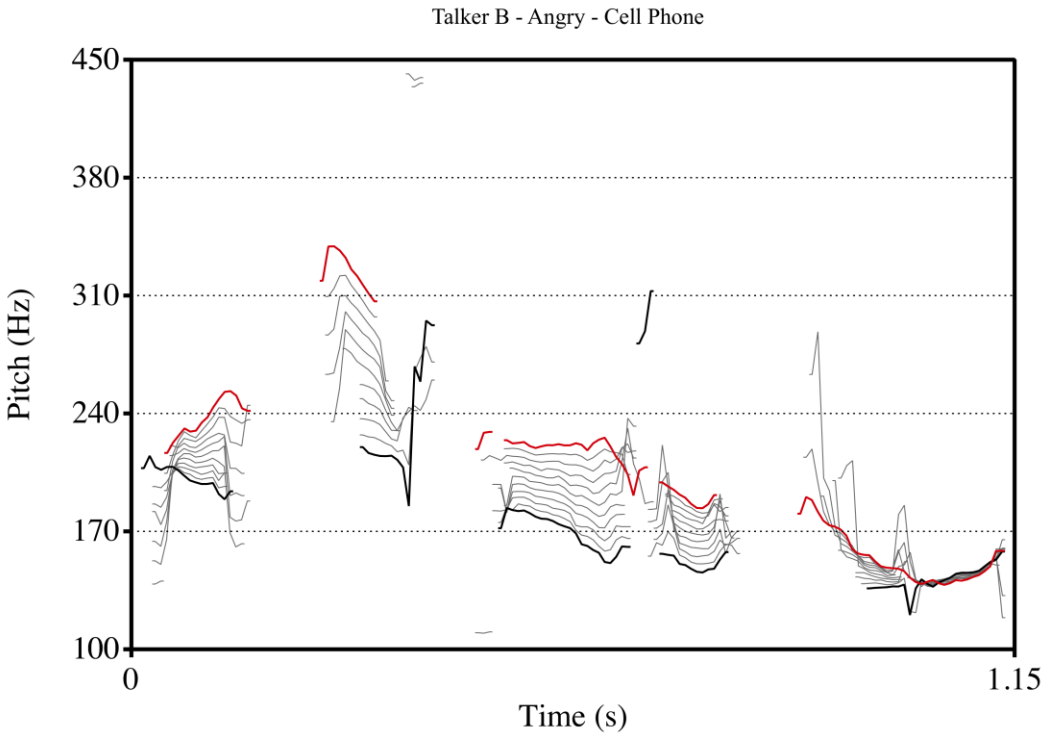
Black represents 0% angry; Red represents 100% angry; the grey lines represent the individual morph steps for the morphed stimuli.

FIGURE 17 Pitch contour graph for stimuli from Talker B, Angry, Microphone



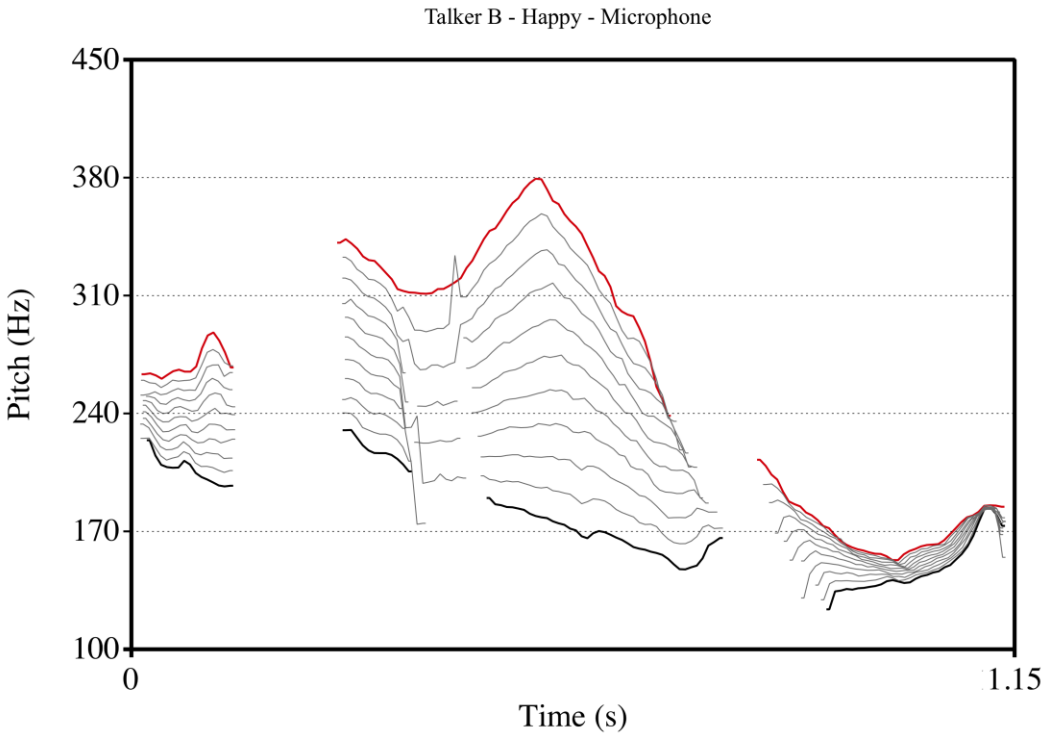
Black represents 0% angry; Red represents 100% angry; the grey lines represent the individual morph steps for the morphed stimuli.

FIGURE 18 Pitch contour graph for stimuli from Talker B, Angry, Cell Phone



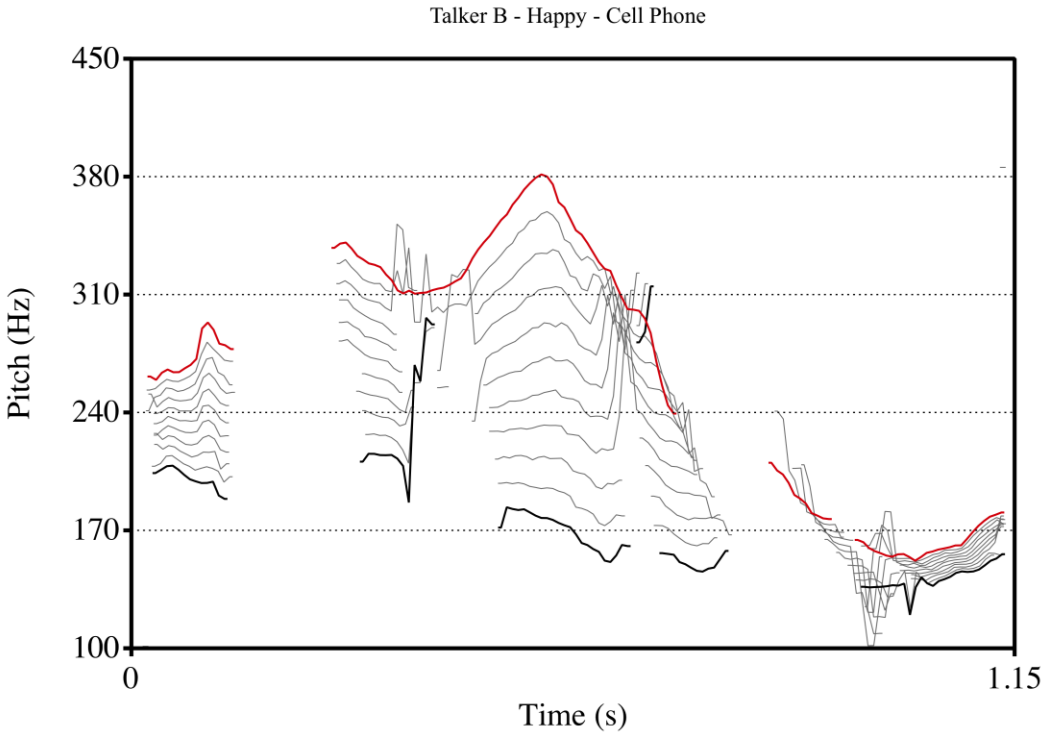
Black represents 0% angry; Red represents 100% angry; the grey lines represent the individual morph steps for the morphed stimuli.

FIGURE 19 Pitch contour graph for stimuli from Talker B, Happy, Microphone



Black represents 0% angry; Red represents 100% angry; the grey lines represent the individual morph steps for the morphed stimuli.

FIGURE 20 Pitch contour graph for stimuli from Talker B, Happy, Cell Phone



Black represents 0% angry; Red represents 100% angry; the grey lines represent the individual morph steps for the morphed stimuli.