The Pennsylvania State University
The Graduate School
College of Engineering

THE ACOUSTICS OF EMOTION: CREATION AND CHARACTERIZATION OF AN EMOTIONAL SPEECH DATABASE

A Thesis in

Acoustics

by

Peter McPhillips Moriarty

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2017

The thesis of Peter McPhillips Moriarty was reviewed and approved by the following:

Michelle C. Vigeant
Assistant Professor of Acoustics and Architectural Engineering
Thesis Advisor

Pamela M. Cole
Liberal Arts Professor of Psychology and Human Development and Family Studies

Karl M. Reichard
Assistant Professor of Acoustics, Applied Research Laboratory

Victor W. Sparrow
Professor of Acoustics
Director of the Graduate Program in Acoustics

*Signatures are on file in the Graduate School

Abstract

Paralinguistic features of speech communicate emotion in the human voice. In addition to semantic content, speakers imbue their messages with prosodic features comprised of acoustic variations that listeners decode to extract meaning. Psychological science refers to these acoustic variations as *affective prosody*. This process of encoding and decoding emotion remains a phenomenon that has yet to be acoustically operationalized. Studies aimed at sifting and searching for the salience in emotional speech are often limited to conducting new analyses on material generated by other researchers. This project presented an opportunity for analyzing the communication of emotion on a corpus of naturalistic emotional speech generated in collaboration with Penn State's Psychology Department. To this end, fifty-five participants were recorded speaking the same semantic content in angry, happy, and sad expressive voicings in addition to a neutral tone. Classic parameters were extracted including pitch, loudness, timing, as well as other low-level descriptors (LLDs). The LLDs were compared with published evidence and theory. In general, results were congruent with previous studies for portrayals of more highly aroused emotions like anger and happiness, but less so for sadness. It was determined that a significant portion of the deviations from the scientific consensus could be explained by baseline definitions alone, i.e. whether deviations referenced neutral or emotional LLD values.

A listening study was subsequently conducted in an effort to qualify and contrast the objectively determined effects with perceptual input. Only three of the fifty-five speakers were sampled due to practical concerns for testing time. The study tested whether the sampled recordings reflected naturally recognizable emotion, and the perceived intensity of these emotions. Listeners were able to discriminate the intended emotion of the speaker with success rates in excess of 87%. Perceptual intensity ratings revealed that some of the prototypical acoustical cues did not significantly correlate with the perception of emotional intensity. Results from both rounds of analysis indicate that a wealth of emotionally salient acoustical information has yet to be fully characterized.

Table of Contents

List of Figures

List of Tables

Acknowledgments

Funding for this project was made possible through the National Institutes of Health grant #1R21 MH104547-01 and the Joseph and Irene Tobis Graduate Fellowship in Acoustics. Their combined support has been instrumental in keeping this work strong and focused.

I would like to extend my most sincere thanks to my thesis committee for the guidance, discussion, and generosity provided throughout this entire process. The flexibility and timeliness of all of our interactions is has gone beyond all believable expectations. I am deeply appreciative of my advisor, Dr. Michelle Vigeant for the attention and support she has given me throughout my time at Penn State. Thank you for bringing me here to study acoustics and for fostering my growth as a scientist.

Thanks to the members of my research team, specifically Dave Dick, Martin Lawless, and Matthew Neal. From digital signal processing to advanced statistics, your advice and generosity has made much of this work possible.

Lastly, I would like to thank my family and friends for their unwavering love and support. Your interest in my work and optimism for our future has fueled my development as a human and respect for the world we live in.

1 Introduction

The work presented in this thesis is the product of a collaboration between acousticians and psychologists at The Pennsylvania State University. This effort, called the Processing of the Emotional Environment Project (PEEP) has both short- and long-term goals that shaped the scope and methods of the presented works in acoustics. First and foremost, the primary topic of interest is the production and processing of emotional speech or more specifically *affective prosody*. This term has several meanings, but the features of speech to which it refers are generally consistent. The most presently relevant definition of *affective prosody* is “the suprasegmental patterns of stress and intonation in speech that contain emotional information” [1].

PEEP’s short-term goals were to examine how affective prosody in natural speech is processed by children at the neurological level. The primary focus was maternal affective prosody, so the vocal connection between mothers and their children was considered. A large corpus of emotional speech recordings was created specifically intended for playback in a highly structured and noisy environment. Differences in neural activity of the children were measured through the use of functional magnetic resonance imaging (fMRI) as they listened to these recordings of emotional speech.

These design objectives fundamentally shaped the direction for the work in this thesis. Firstly, only female voices were recorded and analyzed. This detail, as will be discussed in later sections, shifts many of the expected ranges for acoustics by an octave, a factor of 10, or many other ratios relevant to analyzing the voice. Second, every recording was limited to exactly ten seconds in length. So the spectrum of emotional content was practically constrained in a temporal manner whilst natural inclinations may extend beyond those bounds.

The primary objectives summarized in this thesis are to create and characterize a database of realistic emotional speech. Although the initial and broader purpose of the database is for neurocognitive evaluation, the wealth of its acoustic information bears a unique opportunity for further study. In Chapter 3, details regarding the design and construction of the speech corpus will be discussed. Here, greater time will be devoted to the procedures and processing as they affect the acoustics rather than the neuroimaging goals.

With a finished corpus ready for examination, the remaining chapters will attempt to characterize the stimuli via acoustical and subjective means. Chapter 4 will focus on acoustic analysis of the measured stimuli. Acoustic features will be included on the basis of comparability to previous works. This feature set will go hand in hand with details specific to calculation methods and challenges discovered along the way. In Chapter 5, justification, design, and results will be discussed for a study aimed at evaluating decoders' perception of the new speech corpus. The collection of listener's perceptual ratings will then be compared to patterns found in the acoustical analysis. With a full set of objective and subjective information, this thesis will attempt to answer the following five questions:

i. Does the presented speech corpus exhibit acoustic cues that agree with the literature?
ii. Do the acoustic cues known to carry emotional information correlate independently of semantic content?
iii. How well does the subjective perception of emotion in the corpus agree with the intended emotion of the speaker?
iv. Does the subjective perception of emotion in the corpus correlate with the acoustic patterns?
v. To what degree does the semantic content interfere with the perception of emotion?

2 Background Information

2.1 Psychology of Human Emotion

The word 'emotion' is more easily defined in terms of a few key compositional elements, and the functional role that emotion plays in life. A fair consensus in the literature identifies the following components: cognitive appraisal, subjective feeling, physiological arousal, expression, action tendency, and regulation. Scherer combines the aforementioned elements into the following definition for emotion: "emotions are episodes of coordinated changes in several components (including at least neurophysiological activation, motor expression, and subjective feeling but possibly also action tendencies and cognitive processes) in response to external or internal events of major significance to the organism" [2]. There are, however, two general interpretations of emotion: 1. The 'narrow sense' in which emotion "…is (temporarily) the dominant feature of mental life…" and "directs people strongly towards a course of action…" and 2. The 'broad sense,' which encompasses "underlying emotion"(s) that, to a lesser extent, modify a person's thinking and interactions [3]. Accepting the broader interpretation of emotion, impure as it may be by the standards of others, enables a greater depth of research into human-computer interaction. After all, training a machine to predict the occurrence of "full-blown" anger based on the underlying precursors merits inclusion in the analysis if possible [3].

2.1.1 Theories on human emotion

One of the primary roadblocks facing the quantification and analysis of emotional speech is reconciling current theoretical models of emotion with its real natural mechanisms. In pursuit of the ground truth, researchers have extrapolated from the theory of evolution while others have built theories from observationally substantiated modern cognitive theory. Attempting to fully characterize emotion without a theory-driven hypothesis is like linearly regressing apples with oranges; searching in the dark risks a fruitless yield.

Scientists have postulated evolution's role in the development of emotion in the human race since the time of Charles Darwin [4]. Evolution as theorized creates a causal link between natural selection and neurophysiological advantages of the fittest. It would seem fitting that emotions were born of the same process; stronger and faster neurological reactions to goal-oriented environmental changes would be favored in the presence of an adversary. e.g. "…an event may be appraised as harmful, evoking feelings of fear and physiological reactions in the body; individuals may express this fear verbally and nonverbally and may act in certain ways rather than others…" [3]. A theoretical implication of the link between emotions and evolution is the prioritization of survival-specific emotions over others [5]. Years of propagating this hierarchy would result in a set of core or "basic" emotions from which humans would efficiently convey the desired information with the least chance of misjudgment in perception.

Categorical or discrete theories concerning emotion construct a hierarchy of emotions in order of their significance to survival. While there is little agreement on a definitive set of basic emotions, researchers in support of the discrete theory agree on the principal that natural selection has advantaged hard boundaries between the basic emotions. Confusing emotional signs of fear with happiness while in the presence of a life-threatening stimulus would have catastrophic results. Furthermore, the biological grounds of the origin of emotion suggest that as humans evolved alike their emotions changed similarly [4]. Over the course of evolution, changes to the nervous system would be reflected in unique sets of cognitive appraisal, physiological arousal, expression, action tendency [2] [4] [5]. Implications of how emotional states produce observable differences in the central, somatic and autonomic nervous systems is discussed in Chapter 4.

Unfortunately, societal norms across most cultures disfavor frequent expression of strong emotions, thus there exists a disagreement between the evolutionary favor of "full-blown" emotions and the discouragement seen today. Cognitive theorists have proposed that emotions are related on the basis of a coordinate system comprised of a small set of underlying continuous dimensions. Two primary dimensions include a horizontal "valence" and vertical "activation" axes illustrated in Figure 2.1. Valence refers to the positive (right) or negative (left) evaluation people assign to a given stimulus. Furthermore, the activation level indicates each emotion's proclivity to cause a person to take an action.

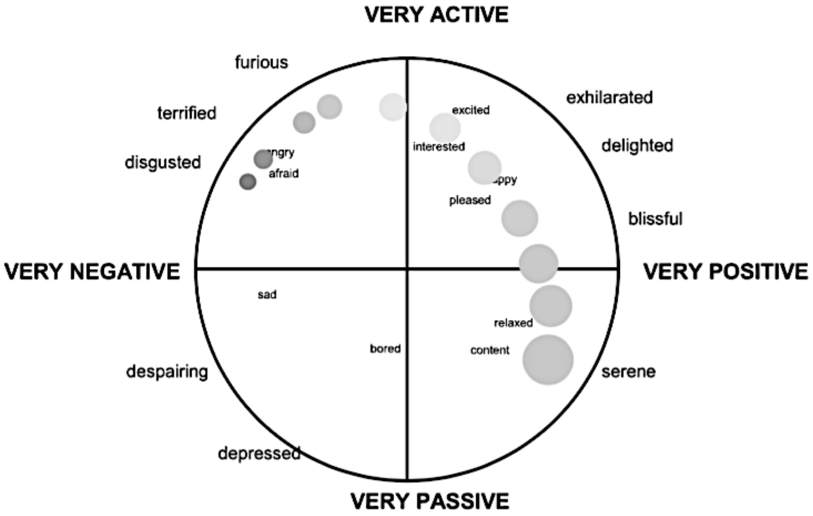


Figure 2.1. Dimensional representation of emotion (Fig. 2 from Cowie and Corneilus 2003).

The distinction between discrete and dimensional emotion theories is drawn here because each theory offers differing views on how emotion is expressed and perceived. If the perception of emotion is categorical and discrete, then the dimensional approach of placing emotions in a fixed space relative to one another loses support. Scherer proposed a theory of emotion that models emotion as a response to a series of successive checks on internal and external stimuli [6]. Although he does not directly support discrete emotion theory, the initial predictions he provided matched the same five basic emotions of anger, fear, disgust, happiness, and sadness. Although many of these details are beyond the immediate scope of this thesis, attention must still be drawn to the fact that there is still debate over the definitions and functions of emotion.

2.1.2 How people convey emotion

The communication of emotion is accomplished by two main roles: encoders and decoders. Encoders (speakers) perform the act of imbuing their emotional state into speech by changing specific modes of articulation and phonation. It should be noted that the social environment of the speaker also helps determine the code by which their emotions may be patterned into speech [7]. As a result, the origin of what influences the way a speaker encodes emotion is differentiated in terms of push (internal) and pull (external) effects [8]. While the encoder may feel the urge to yell loudly at a competitor, social norms (e.g. in the office environment) may curb the magnitude of the expression [9].

There is, of course, an inevitable difference between the natural phenomenon of communication and that which can be measured objectively. Scherer's modified version of Brunswik's functional lens model delineates a model for vocal communication of emotion by objectivity and respective roles as illustrated in Figure 2.2 [10].

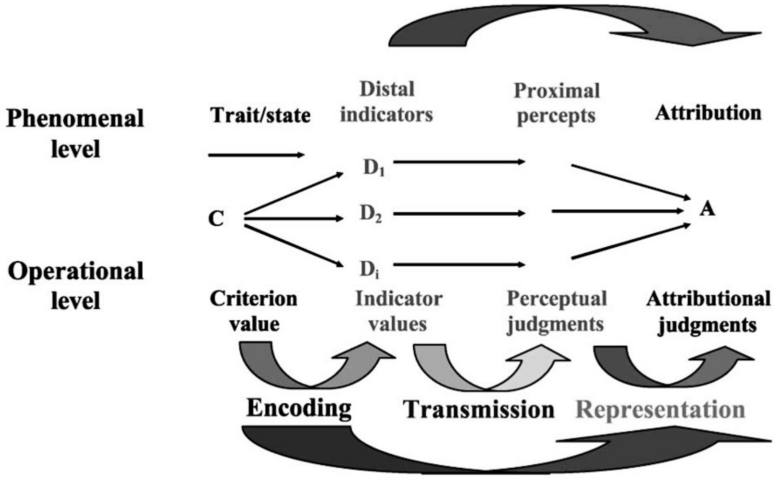


Figure 2.2 A Bruswikian lens model of the vocal communication of emotion (From Fig. 1 Scherer 2003).

The upper half of the diagram, which is labeled the "phenomenal" level, represents the flow of emotional information as it is heard and felt in the natural world, whereas the lower "operational" level represents this flow by way of measureable and experimentally surveyed information. The chain of communication is further divided horizontally into an encoding, transmission, and decoding (representation) processes. Clearly the efforts in vocal expression research are applied to operationalizing the presented phenomenon, but the terminology within the model must first be explained.

The speaker is represented by the letter C in the top left portion of the model. Given an emotional state, this person will encode their speech with acoustic cues in a way that is *ecologically valid* or accurate to the specific emotional state. From the perspective of the listener at letter A, these cues (D) are labeled as *distal* as they are distant from the listener. The acoustic signal then moves from the source to the receiver through a transmission pathway which will modify the sound by way of a transfer function. Both the acoustic transfer function of the transmission pathway and the psychophysiological biases of the receiver contribute to the perception of these cues, which are given the term *proximal percepts*. Finally, the listener attributes these modified cues in a cumulative manner to decode the emotional information they perceive.

If the goal is indeed to understand how an emotion moves between people, both the encoding and decoding process require attention. While the neurological inception of emotion in the brain remains outside the scope of this paper, comparison of the acoustic changes of the voice to subjective ratings offers a feasible path forward. The simplest experimental architecture would compare measured airborne vibrations propagating from a source that has assumed multiple emotionally charged states. The ecological validity of a few of these extracted acoustic cues would then be tested through listener judgements. Ideally, the first step of this process would use an infinite number of acoustic cues, however this is far beyond practical to the computational power of today. Instead, the proposed set can be limited to parameters unique to humans and indicative of physiological change. It is only through a comprehensive understanding of the way humans produce sound that the acoustic predictors of emotion may be determined in a practical manner. The following section will attempt to characterize the general principles of voice production before covering the acoustical descriptors specific to human emotion.

2.2 The Human Voice

2.2.1 Anatomy and Definitions

The human speech production system is comprised of several physical mechanisms cooperating in harmony. Exchange of distinguishable semantic information is made possible through the manipulation of these mechanisms; the glottis excites the air in the vocal tract, the mouth and tongue move to form vowels, and the lips create constrictions necessary for many consonants and so forth. Combined, these effects encode audible information ranging from the content of the language to the emotional state of the speaker.

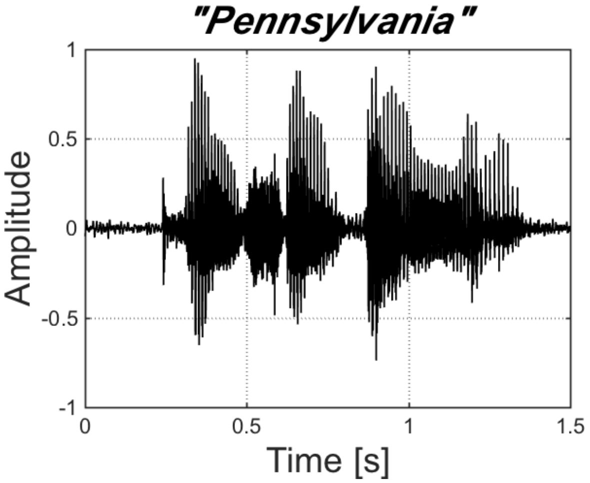


Figure 2.3 This is a time series of a male speaking the word "Pennsylvania."

From the surface, the measured signal produced by this dynamic system suggests content rich in temporal and tonal complexity as depicted in Figure 2.3. Performing acoustical analysis without prior knowledge of the system responsible for this signal would yield very little understanding of the state of the source, and may often mislead the researcher tasked with deciphering the encoded content. Understanding the nature of the human speech production system and each mechanism's contribution informs and refines such analyses towards a more profitable end.

Although the entire body contributes in some form to the production of speech, the vast majority of the acoustical work is confined to the vocal tract. The vocal tract has several dimensional and functional features that contribute to the overall function of the voice. Figure 2.4 is an illustration of the vocal tract, which begins at the glottis and terminates at the mouth. A key gender-specific dimensional difference of the vocal tract is its length, which ranges on average from 14.0 cm for women to 17.5 cm for men [11]. The length of the vocal tract is divided largely into two primary volumes: the pharynx (located above the vocal cords), and the mouth cavity. Located below the pharynx and just between the vocal cords is the glottis, which is the fundamental source of acoustical energy in the system. At the top of the pharynx is the velum, which divides the mouth cavity from the nasal cavity. Located at the end of tract are the tongue and the lips. The importance of the tongue mirrors that of the glottis: the tongue moves with exceptional dexterity and range of motion, and acts as the dominant manipulator of airflow in the vocal tract.

During the process of speech production, humans employ a mixture of independent and dependent control over the responsible muscular network. Muscles that govern diaphragm contraction, vocal cord tension, velum position, oral cavity shape, and formation of the lips can all be operated independently of one another. Much of this system, however, is hardwired to work under predetermined shapes and motions that are physically necessary for sound production; the lungs supply air pressure continuously until the utterance is finished, and the velum and mouth provide the open termination necessary for both pressure release and acoustic radiation. The interdependent nature of this system is emphasized here because it is the cooperative effort of these mechanisms that determines many fundamental characteristics of speech such as the movement of the glottis.

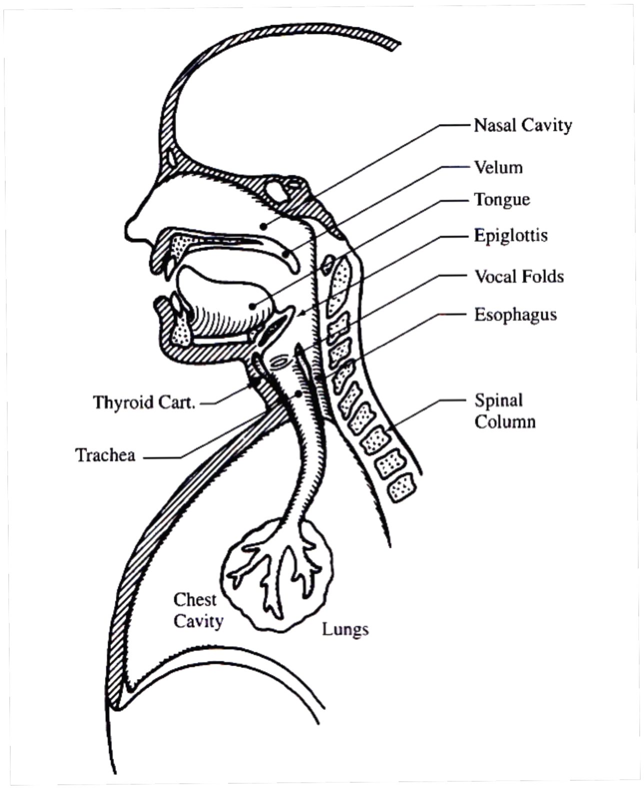


Figure 2.4 Cross-section view of the human vocal tract. The arrow indicates the location of the glottis (From Fig. 3.3 Rabiner & Schafer 2011).

2.2.2 Mechanisms of Excitation and Transformation

The most common form of glottal excitation in humans resembles the functionality of a relaxation oscillator [11]. The diaphragm applies increasing subglottal pressure until the pressure exceeds the tension exercised in the vocal cords; thus the cords break open. By virtue of the Bernoulli Effect, the resulting high-velocity stream of air between the cords creates a low-pressure region, which, with the help of the cord tension, pulls the glottis to a closed state. A generic example of one such glottal pulse is illustrated in Figure 2.5.

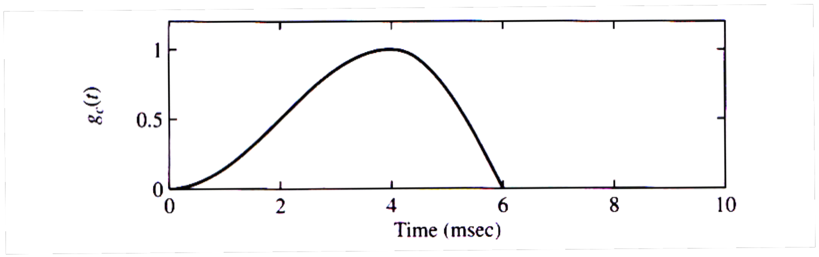


Figure 2.5 Time series of Rosenberg's glottal pulse model (From Fig. 5.21 Rabiner & Schafer 2011).

This process repeats as long as there is sufficient pressure from the diaphragm, tension in the vocal cords, and space for the air to exit the vocal tract. The importance of the pressure release at the end of the vocal tract is easily demonstrated by holding the nose and mouth closed while attempting to "hum." Thus, the movement of air together with the manipulated tension in the vocal cords enables the oscillatory motion of the vocal cords. The rate at which the glottis opens and closes defines the fundamental frequency of a vocalized speech segment and is more commonly known as the glottal pulse rate. For the purposes of consistency to the related works in vocal expression, the notation *F0* will be used to represent pitch. Average *F0* values for men and women are 120 [Hz] and 210 [Hz] respectively.

The origin of acoustic energy in speech is not limited to the impulsive excitation by the glottis; constrictions and sharp edges anywhere along the vocal tract offer additional modes of sound production. The physical basis for this phenomenon is most easily explained in terms of turbulence. Assuming incompressible flow, the equation of continuity given in (2.1) forms the relationship between fluid velocity in two different diameter pipes as depicted in Figure 2.6.

$$\rho_1 A_1 v_1 = \rho_2 A_2 v_2 \qquad (2.1)$$

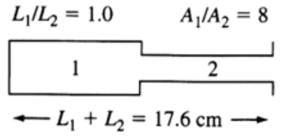


Figure 2.6 Pipe of two different diameters (From Fig. 5.31 Rabiner & Schafer 2011)

Given the assumption of incompressible flow, the density is not affected by the change in velocity or area. Thus, the fluid velocities and areas can be related by (2.2).

$$\frac{v_2}{v_1} = \frac{D_1^2}{D_2^2} \qquad (2.2)$$

Here it is seen that the ratio of the fluid velocities is proportional to the ratio of square of the diameters. Consequently, if D_1 is twice the size of D_2, the velocity v_2 will increase by a factor of four. This relationship is especially important when predicting onset of turbulence with Reynold's number (*Re*) given in (2.3).

$$Re = \frac{\rho V_p D_p}{\mu} \qquad (2.3)$$

Here, ρ is the fluid density in [kg/m^3], V_p is the velocity in [m/s], D_p is the diameter of a circular pipe in [m]. For fluid flow in a pipe, a Reynold's number less than 2300 is considered laminar, while *Re* between 2300 and 4000 is transient, and *Re* greater than 4000 is considered turbulent. The higher the Reynold's number is, the greater the amount of turbulence in the flow [12]. Considering the numerator of (2.3), if the diameter decreases by a factor of two from one section to the next, the velocity of the fluid in the smaller section of pipe goes up by a factor of four. In the context, typical trachea diameters for men and women are 26 mm and 22 mm respectively [13], and typical breathing velocities range from 0.79 to 3.16 [m/s] [14]. For an average women breathing at rest, the Reynold's number in the trachea is approximately 1161, and is therefore considered laminar. If the glottis above the trachea closes to just 11 mm in diameter, the velocity increases by a factor of four, and Re increases to 2323. Halving the diameter again to 5.5 mm, the Reynold's number jumps to 4647 which is well within the turbulent region. As a direct result of its chaotic nature, turbulent flow created by a simple constriction generates acoustic energy in the form of broadband noise. Constrictions at the glottis enable whispering, while constrictions at the teeth and lips give rise to consonants.

The impulsive and noise-like modes of excitation provide a wealth of spectral content, which the vocal tract selectively filters. If modeled as a constant area tube with one end open and the other closed, a 14 cm vocal tract would have resonances with decreasing amplitude starting at 612.5 [Hz], followed by harmonics at odd-integer multiples. For the average women speaking with an *F0* of 210 [Hz], the first tube resonance occurs at nearly three times the glottal pulse rate. The effect of this tube resonance is illustrated in Figure 2.7. More accurate conformity of the pipe model to the human vocal tract is accomplished by allowing the cross-sectional area of the pipe to change with distance from the glottis. Subsequently, modifications of the area along the vocal tract form the physical filters that transform initial impulse and broadband excitation to intelligible parts of speech.

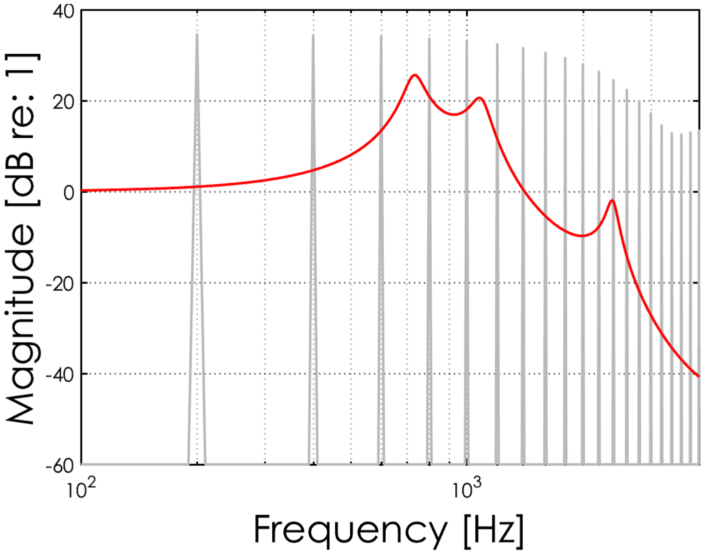


Figure 2.7 Magnitude spectrum of the vocal tract's frequency response (red) overlaid on top of a spectrum of a glottal pulse train (gray).

Figure 2.8 depicts a simple model of the voice that employs the mechanisms delineated in the following procedure:

1. The diaphragm pushes air from the lungs up through the trachea to the larynx.
2. Muscles in the larynx pull the vocal cords tense, which creates a build-up of pressure below the vocal cords.
3. When the pressure below the vocal cords exceeds their tension, air is released upwards into the vocal tract.
4. By virtue of the Bernoulli Effect, the high velocity of the air between the vocal cords decreases the pressure at the site, which pulls the cords back together. This closing motion is also assisted by the restorative tension in the cords supplied by the larynx.
5. Repeated pulses of air moving up through the vocal tract are spectrally contoured by the mouth cavity and the nasal cavity.
6. The combined work of the tongue, lips, and velum determine the effective size and shape of these cavities by directing flow or filling the volume.

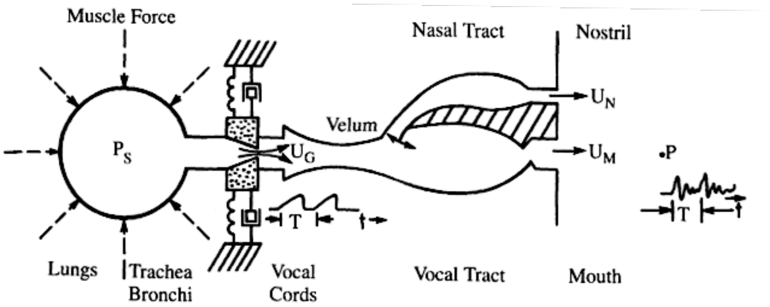
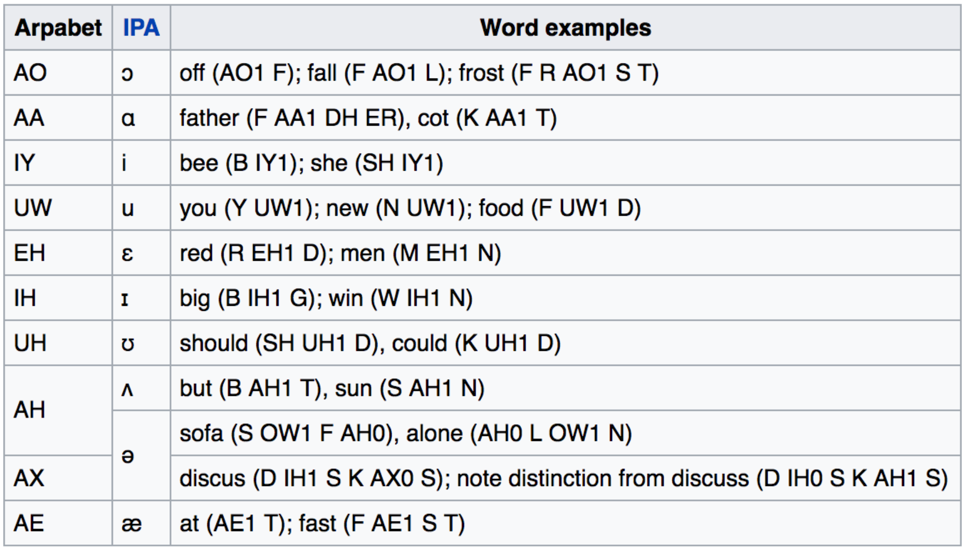


Figure 2.8 Schematic of the voice production mechanism (After Flanagan et al. 1970 IEEE)

2.2.3 Phonetic Descriptions of Speech

As the discussion moves towards translating mouth positions to spoken language, a code is needed or describing human phonetic capabilities, and a set of vocabulary for identifying units of speech. The most basic unit of speech that will be considered in the presented work is an utterance, which is defined as “an uninterrupted chain of spoken or written language” [15]. As the focus of this work is on expressive speech, there will be cases where an utterance may have brief moments of silence. Created by the Advanced Research Projects Agency (ARPA) the ARPAbet describes every phoneme, or unit of spoken sound, in American English in two letter combinations as detailed in Table 2.1. For example, the vowels in the words “bot” and “bee” are represented by /AA/ and /IY/ respectively.

Table 2.1 Monothongs of the American ARPAbet (Wikipedia).



Arpabet	IPA	Word examples
AO	ɔ	off (AO1 F); fall (F AO1 L); frost (F R AO1 S T)
AA	ɑ	father (F AA1 DH ER), cot (K AA1 T)
IY	i	bee (B IY1); she (SH IY1)
UW	u	you (Y UW1); new (N UW1); food (F UW1 D)
EH	ɛ	red (R EH1 D); men (M EH1 N)
IH	ɪ	big (B IH1 G); win (W IH1 N)
UH	ʊ	should (SH UH1 D), could (K UH1 D)
AH	ʌ	but (B AH1 T), sun (S AH1 N)
	ə	sofa (S OW1 F AH0), alone (AH0 L OW1 N)
AX		discus (D IH1 S K AX0 S); note distinction from discuss (D IH0 S K AH1 S)
AE	æ	at (AE1 T); fast (F AE1 S T)

It is the differences in the physical state of the vocal tract that give rise to perceived distinctions between /AA/ and /IY/. In the case of /AA/, the lips form a broad circle while the jaw is lowered and the tongue retracted to the back of the mouth as illustrated in Figure 2.9. By contrast, /IY/ is produced by pulling the lips closer to the teeth in a grin while the tongue is raised up to the roof of the mouth.

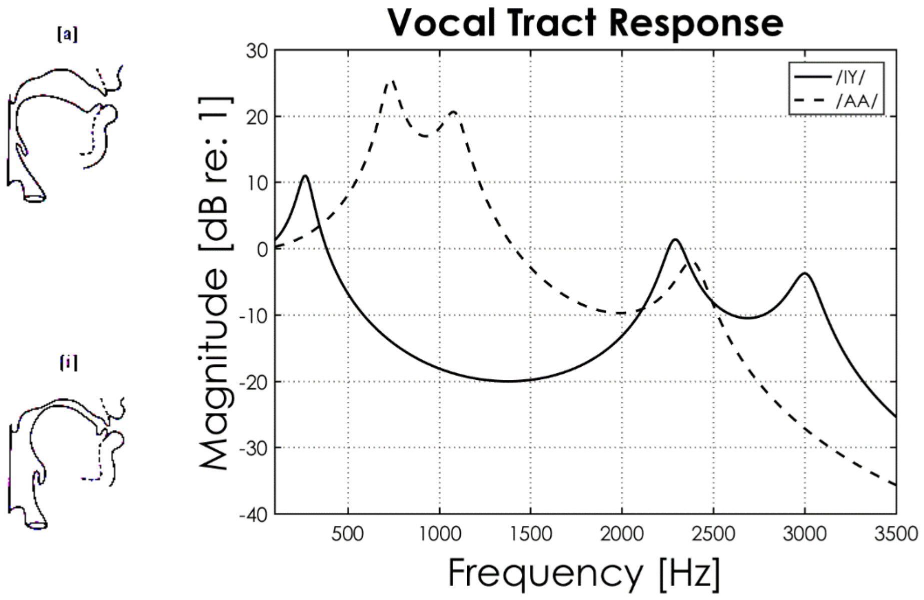


Figure 2.9 Mouth positions for the vowel /AA/ (top left), /IY/ (bottom left), and their corresponding magnitude spectra (figures to the left are from Fig. 3.18 Rabiner & Schafer 2011).

The effect of these physical differences between /AA/ and /IY/ as illustrated in Figure 2.9 is evident in the amplitude and placement of the harmonics or “formants” in their spectra. In /IY/, the lowest formant occurs at a lower frequency than in /AA/, however the second and third formants in /IY/ are much higher than those of /AA/. Lowering the jaw as in the case of /AA/ reduces the volume behind the tongue and subsequently increases the frequency of the first formant. The change in oral cavity in /AA/ has a similar effect; the larger volume lowers the frequency of the second formant in contrast with /IY/.

2.3 Physical mechanisms modulated by emotion

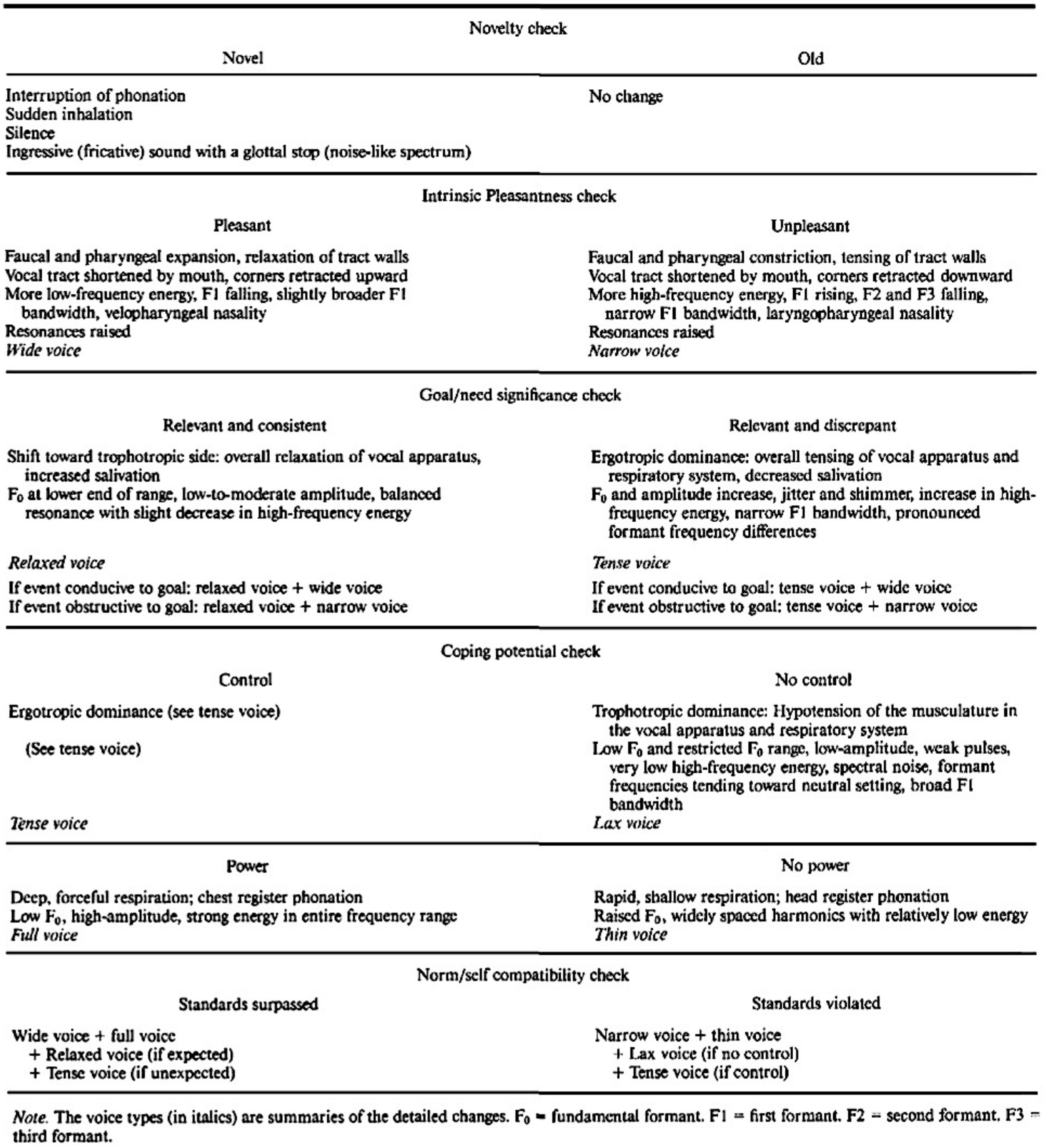
2.3.1 Scherer’s 1986 Predictions

Facial patterns and their relationship to theories on the expression of emotion are numerous and well-studied [16] [17] [18] [19]. Only one modern theory on vocal expression of emotion presents a rigorous argument for how the mind, the body, and the voice coordinate an acoustic signal. Broadly speaking, Scherer hypothesized that emotions which are known to manifest physically in the body must also have some effect on the way we speak [6]. Such physiological symptoms range from stress-induced muscular tension to increased salivation that results from a feeling of pleasantness. This section will briefly detail the connection between emotions and the body along with some predicted acoustic markers specific to each case.

Although not completely aligned with discrete emotion theory, Scherer supports the idea that emotions are “a series of interrelated adaptive changes in several organismic subsystems following antecedent events evaluated to be of major relevance to an organism’s goals” [6]. In Scherer’s proposed “Component Process Theory”, the organism performs a continuous series of cognitive stimulus evaluation checks (SECs) to determine the appropriate emotion to express. Evaluated in hierarchical order, the SECs include a novelty check, intrinsic pleasantness check, goal/need significance check, coping potential check, and finally a norm/self-compatibility check.

Scherer holds that the outcome of each successive SEC moving from top to bottom along Table 2.2 produces a specific physiological change through either the somatic nervous system (SNS) or the autonomic nervous system (ANS). These two nervous systems differ both in the physical mechanisms they govern and the speed with which they respond to stimuli. The SNS is responsible for motor control and responds quickly to both involuntary (tonic) responses to stimuli and voluntary (phasic) attempts to possibly control the organism’s expression. For example, this system would be responsible for tensing the vocal cords and shortening of the vocal tract when a stressful stimulus is presented, therefore raising the ***F0*** and formants of the speaker. The facial muscles involved with smiling are also under the control of the SNS, which has been shown to increase the formants F2 and F3 [20]. The ANS, on the other hand, is responsible for slowly changing, involuntary responses to stimuli. Systems under the control of the ANS include respiration and salivation, which can have a significant effect on both the subglottal pressure and damping within the vocal tract.

Table 2.2 Scherer's Component Patterning Theory (from Table 4 Scherer 1986).



Novelty check	
Novel	Old
Interruption of phonation Sudden inhalation Silence Ingressive (fricative) sound with a glottal stop (noise-like spectrum)	No change
Intrinsic Pleasantness check	
Pleasant	Unpleasant
Faucal and pharyngeal expansion, relaxation of tract walls Vocal tract shortened by mouth, corners retracted upward More low-frequency energy, F1 falling, slightly broader F1 bandwidth, velopharyngeal nasality Resonances raised *Wide voice*	Faucal and pharyngeal constriction, tensing of tract walls Vocal tract shortened by mouth, corners retracted downward More high-frequency energy, F1 rising, F2 and F3 falling, narrow F1 bandwidth, laryngopharyngeal nasality Resonances raised *Narrow voice*
Goal/need significance check	
Relevant and consistent	Relevant and discrepant
Shift toward trophotropic side: overall relaxation of vocal apparatus, increased salivation F_0 at lower end of range, low-to-moderate amplitude, balanced resonance with slight decrease in high-frequency energy	Ergotropic dominance: overall tensing of vocal apparatus and respiratory system, decreased salivation F_0 and amplitude increase, jitter and shimmer, increase in high-frequency energy, narrow F1 bandwidth, pronounced formant frequency differences
Relaxed voice	*Tense voice*
If event conducive to goal: relaxed voice + wide voice If event obstructive to goal: relaxed voice + narrow voice	If event conducive to goal: tense voice + wide voice If event obstructive to goal: tense voice + narrow voice
Coping potential check	
Control	No control
Ergotropic dominance (see tense voice) (See tense voice)	Trophotropic dominance: Hypotension of the musculature in the vocal apparatus and respiratory system Low F_0 and restricted F_0 range, low-amplitude, weak pulses, very low high-frequency energy, spectral noise, formant frequencies tending toward neutral setting, broad F1 bandwidth
Tense voice	*Lax voice*
Power	No power
Deep, forceful respiration; chest register phonation Low F_0, high-amplitude, strong energy in entire frequency range *Full voice*	Rapid, shallow respiration; head register phonation Raised F_0, widely spaced harmonics with relatively low energy *Thin voice*
Norm/self compatibility check	
Standards surpassed	Standards violated
Wide voice + full voice + Relaxed voice (if expected) + Tense voice (if unexpected)	Narrow voice + thin voice + Lax voice (if no control) + Tense voice (if control)

Note. The voice types (in italics) are summaries of the detailed changes. F_0 = fundamental formant. F1 = first formant. F2 = second formant. F3 = third formant.

With these governing principles in mind, Scherer drafted predictions of specific acoustic cues that would result from the emotional state of the organism, summarized in Table 2.3. While there are a multitude of emotions that humans can express, just the emotions studied in the present work (anger, happiness[i], and

[i] Note: The intended meaning and target expression of happiness in this thesis more closely aligns with Scherer's definitions for joy.

sadness) plus an additional two (disgust, fear) have been provided. The remaining information can be found in the appendix. This list of predictions, while based on many assumptions, is the first attempt of its kind to connect emotion theory to acoustic symptom through a system of neurophysiological interactions. As a result, most of these acoustic cues have become the basis by which successive studies have made objective comparisons of emotional expression. The following section will detail definitions and functional relationships of these cues to human expression of emotion.

Table 2.3 Predictions for changes in acoustic cues affected by emotion (Adapted from Scherer, 1986).

	Emotion				
Acoustic Cue	**JOY**	**DISG**	**SAD**	**FEAR**	**RAGE**
F0					
Mean	≥	>	<>	≥≥	<>
Range	≥		≤	≥≥	≥≥
Variability	≥		≤	≥≥	≥≥
Contour	>		≤	>>	=
Intensity					
Mean	≥	>	≤≤	>	≥≥
Range	>		<	>	>
Variability	>		<	>	>
Voice Quality					
Frequency Range	>	>	>	>>	>
HF. Energy	<>	>	<>	≥≥	>>
Spectral Noise			>		
F1 mean	<	>	>	>	>
F2 mean		<	<	<	<
F1 bandwidth	<>	<<	<>	<<	<<
Formant precision	>	>	≤	>	>
Duration					
Speech Rate	≥		≤	≥≥	≥
Transition Time	<		>	<	<

2.3.2 Low-level Descriptors (acoustic cues) of emotional speech

Humans encode an enormity of time-varying information during expressive speech. The framework for classifying emotional speech conveniently summarizes the raw information by order of measurement and analysis. The workflow begins with the recording of an utterance, e.g. a meaningful unit of speech that contains at least one or two spoken words. Within a recorded utterance, the first order of acoustical analysis is conducted either frame-by-frame or over then length of the utterance. Such first order calculations are described as *low-level descriptors* (LLD), which are subdivided into four categories: pitch, voice quality, duration, and intensity [21].

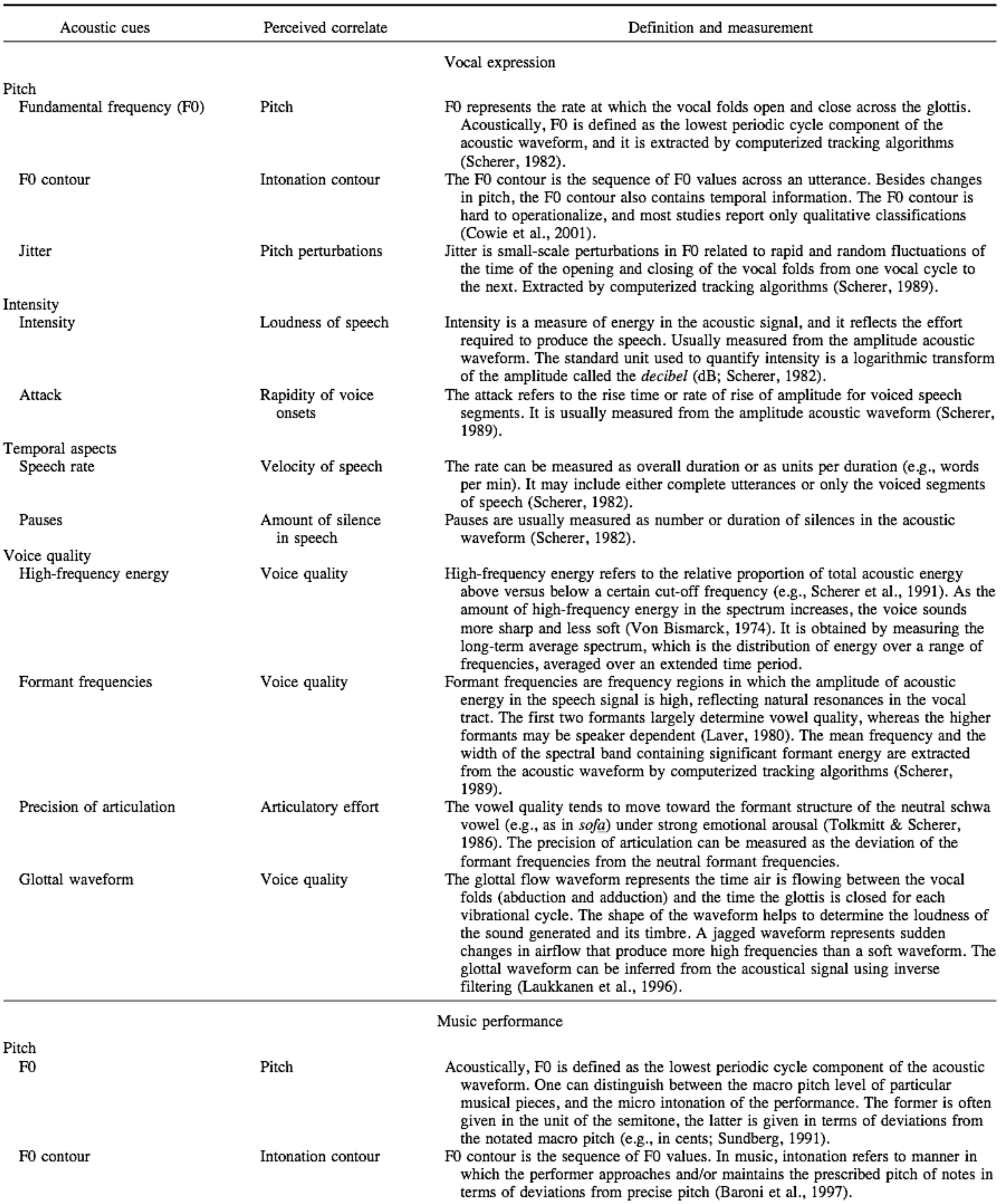
Intonation and stress are largely the result of three main components: pitch, loudness, and duration. Although these primary components of prosody do help characterize much of what is perceived from an utterance, the category of voice quality is included to account for timbre, harshness, and other perceived characteristics of speech that would be otherwise ignored.

Workflow and Definitions

As stated earlier, acoustic LLDs pertaining to emotion are largely grouped into four categories: pitch (*F0*), duration, intensity, and voice quality. Table 2.4 lists the key parameters by category and their respective perceptive definitions. The category of *voice quality* is intended to encompass the characteristics of speech that reflect the physical state of the speaker apart from *duration*, *intensity*, and *pitch*. *Jitter*, *shimmer*, and the *harmonics-to-noise ratio* are often included in the category of *voice quality*, while a fourth category called the *long-term average spectrum* is created to include cues like *high-frequency energy* and *formant frequencies* [21]. Given that the long-term average spectrum contributes to the perception of *voice quality* in many cases, this thesis will restrict the number of overall categories to four, and subdivide *voice quality* into spectral and short-term variability as is shown in Table 2.5.

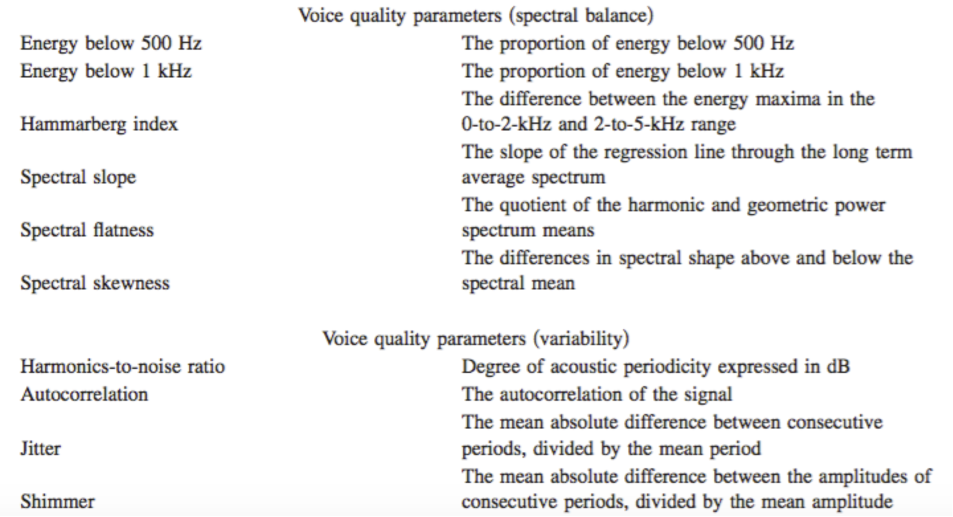
Functionals applied to the LLDs make up the second order of analysis by way of simple statistical calculations, regression, and higher level modeling. Some examples of functionals include taking the mean of the pitch over the utterance, the standard deviation of the intensity, or the average absolute slope of the *HNR*.

Table 2.4 Definitions of acoustic cues and perceived correlate on vocal expression (From Table 6 Juslin & Laukka 2003).



Acoustic cues	Perceived correlate	Definition and measurement
		Vocal expression
Pitch		
Fundamental frequency (F0)	Pitch	F0 represents the rate at which the vocal folds open and close across the glottis. Acoustically, F0 is defined as the lowest periodic cycle component of the acoustic waveform, and it is extracted by computerized tracking algorithms (Scherer, 1982).
F0 contour	Intonation contour	The F0 contour is the sequence of F0 values across an utterance. Besides changes in pitch, the F0 contour also contains temporal information. The F0 contour is hard to operationalize, and most studies report only qualitative classifications (Cowie et al., 2001).
Jitter	Pitch perturbations	Jitter is small-scale perturbations in F0 related to rapid and random fluctuations of the time of the opening and closing of the vocal folds from one vocal cycle to the next. Extracted by computerized tracking algorithms (Scherer, 1989).
Intensity		
Intensity	Loudness of speech	Intensity is a measure of energy in the acoustic signal, and it reflects the effort required to produce the speech. Usually measured from the amplitude acoustic waveform. The standard unit used to quantify intensity is a logarithmic transform of the amplitude called the *decibel* (dB; Scherer, 1982).
Attack	Rapidity of voice onsets	The attack refers to the rise time or rate of rise of amplitude for voiced speech segments. It is usually measured from the amplitude acoustic waveform (Scherer, 1989).
Temporal aspects		
Speech rate	Velocity of speech	The rate can be measured as overall duration or as units per duration (e.g., words per min). It may include either complete utterances or only the voiced segments of speech (Scherer, 1982).
Pauses	Amount of silence in speech	Pauses are usually measured as number or duration of silences in the acoustic waveform (Scherer, 1982).
Voice quality		
High-frequency energy	Voice quality	High-frequency energy refers to the relative proportion of total acoustic energy above versus below a certain cut-off frequency (e.g., Scherer et al., 1991). As the amount of high-frequency energy in the spectrum increases, the voice sounds more sharp and less soft (Von Bismarck, 1974). It is obtained by measuring the long-term average spectrum, which is the distribution of energy over a range of frequencies, averaged over an extended time period.
Formant frequencies	Voice quality	Formant frequencies are frequency regions in which the amplitude of acoustic energy in the speech signal is high, reflecting natural resonances in the vocal tract. The first two formants largely determine vowel quality, whereas the higher formants may be speaker dependent (Laver, 1980). The mean frequency and the width of the spectral band containing significant formant energy are extracted from the acoustic waveform by computerized tracking algorithms (Scherer, 1989).
Precision of articulation	Articulatory effort	The vowel quality tends to move toward the formant structure of the neutral schwa vowel (e.g., as in *sofa*) under strong emotional arousal (Tolkmitt & Scherer, 1986). The precision of articulation can be measured as the deviation of the formant frequencies from the neutral formant frequencies.
Glottal waveform	Voice quality	The glottal flow waveform represents the time air is flowing between the vocal folds (abduction and adduction) and the time the glottis is closed for each vibrational cycle. The shape of the waveform helps to determine the loudness of the sound generated and its timbre. A jagged waveform represents sudden changes in airflow that produce more high frequencies than a soft waveform. The glottal waveform can be inferred from the acoustical signal using inverse filtering (Laukkanen et al., 1996).
		Music performance
Pitch		
F0	Pitch	Acoustically, F0 is defined as the lowest periodic cycle component of the acoustic waveform. One can distinguish between the macro pitch level of particular musical pieces, and the micro intonation of the performance. The former is often given in the unit of the semitone, the latter is given in terms of deviations from the notated macro pitch (e.g., in cents; Sundberg, 1991).
F0 contour	Intonation contour	F0 contour is the sequence of F0 values. In music, intonation refers to manner in which the performer approaches and/or maintains the prescribed pitch of notes in terms of deviations from precise pitch (Baroni et al., 1997).

Table 2.5 Modern categorization of voice quality features (From Table II Goudbeek & Scherer 2010).



Voice quality parameters (spectral balance)	
Energy below 500 Hz	The proportion of energy below 500 Hz
Energy below 1 kHz	The proportion of energy below 1 kHz
Hammarberg index	The difference between the energy maxima in the 0-to-2-kHz and 2-to-5-kHz range
Spectral slope	The slope of the regression line through the long term average spectrum
Spectral flatness	The quotient of the harmonic and geometric power spectrum means
Spectral skewness	The differences in spectral shape above and below the spectral mean
Voice quality parameters (variability)	
Harmonics-to-noise ratio	Degree of acoustic periodicity expressed in dB
Autocorrelation	The autocorrelation of the signal
Jitter	The mean absolute difference between consecutive periods, divided by the mean period
Shimmer	The mean absolute difference between the amplitudes of consecutive periods, divided by the mean amplitude

Duration

LLDs that fall into the *duration* category characterize the length of speech characteristics on a macroscopic scale. Of common interest and utility is determining the duration of the utterances within a speech episode. This type of measurement requires an accurate and repeatable tool for calculating the endpoints of an utterance. One such method employs a combination of exponential averaging a sampled time series and selecting amplitudes above a certain value as is shown in Figure 2.10.

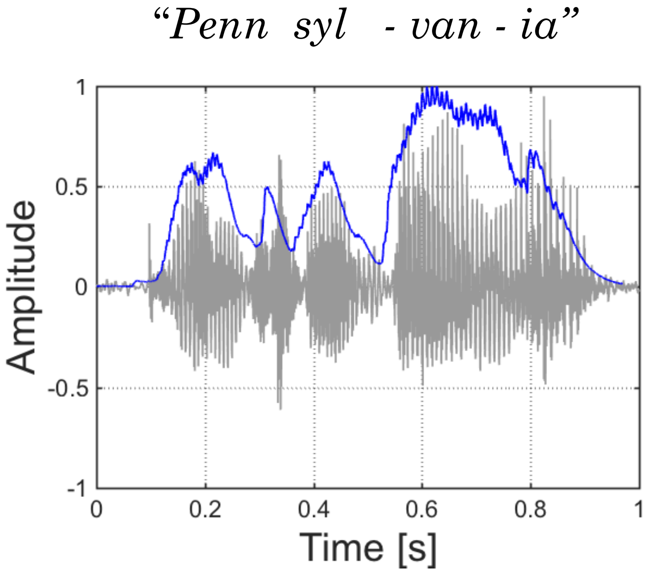


Figure 2.10 The word "Pennsylvania" spoken by a male adult. Original timeseries is in grey, the exponentially averaged time series is in blue.

With the endpoints determined, other features can be known such as the length of the silences between the utterances, and the ratio of the silent and utterance durations.

On a finer scale, duration features also include measures relating to voiced and unvoiced classification of speech. A region of speech is considered “voiced” if the glottis is active and exciting the vocal tract impulsively. All other regions of non-silent speech where the glottis is not active are therefore considered “unvoiced”. Examples of unvoiced parts of speech include /F/, /TH/, /S/, and /SH/. It should also be noted that whispered speech is also classified as “unvoiced”. Classification of voiced and unvoiced speech leads to LLDs such as the length and relative proportions of these regions.

The last characteristic that will be considered in the presented work in the duration category presents a computational challenge in the absence of prior knowledge of the spoken content. The syllabic rate, defined as the number of syllables per second, is often calculated by dividing the total number of syllables within an utterance by the duration of that utterance. Syllables are vowels, diphthongs, or consonants, that often occur before or after consonants [11]. In the context of phonetics, syllables serve as the building blocks for the construction of words. These building blocks largely determine the relative temporal location of stress or emphasis within a word. Unfortunately, the task of coding speech computationally is complicated by slurs and various cases of imprecise articulation that are often manifest by emotional expression. A simple solution to this problem is to count the syllables phonetically from the written text if the language content is known and divide the manual count by the computed duration.

The consensus from most studies in vocal expression relevant to the presented work maintain that emotion modulates duration LLDs primarily on the basis of arousal [22]. When spoken in a highly aroused state, utterances tend to exhibit shorter total durations and increased syllabic rates. In corresponding fashion, lower arousal generally produces speech with decreased durations and syllabic rates.

Intensity

The body of work on affective prosody features the word “intensity” in the context of a subjective and objective measure. In the subjective domain-specifically in the context of emotion, the term “intensity” refers to the strength or magnitude of the emotion as it is perceived by the listener or encoded by the speaker. To avoid confusion, this term will be qualified as emotional intensity. In this section however, the term “intensity” refers to a category of LLDs related to the physical loudness of speech. It is an unfortunate reality that the intended use of the word intensity in many of the related works does not match the actual definition of acoustic intensity. Measurement of acoustic intensity, given in watts per meter squared, requires either a two microphone array or a sensor that can measure pressure and particle velocity simultaneously. The majority of the aforementioned studies actually refer to sound pressure level (SPL) in decibels (dB), which thankfully can be measured with a single element transducer.

In much the same way that duration changes with emotional arousal, differences in intensity have been found to correlate mostly with arousal. The trend between emotions for average values and variability of intensity are the same. Anger and happiness are marked by increases in this category while sadness tends to exhibit lower values.

Pitch (F0)

A primary component in the intonation of speech, the *F0* (pitch) represents the third and final categorical variable in the traditional definition of prosody. As discussed earlier, *F0* is the frequency of opening and closing of the glottis, reported in Hertz [Hz]. Few studies have measured *F0* using electroglottography (EGG), but technical difficulties concerning partial or lack of glottal contact makes acoustic measurement by way of a microphone more feasible [23].

Across languages and cultures, humans modulate their *F0* during both expressive and non-expressive speech [24]. According to the review by Juslin & Laukka, there appears to be strong consensus on the way *F0* changes globally and locally between affects [22]. From the utterance level, average *F0* values increase for both anger and happiness, while sadness is usually characterized by lower *F0* values compared to a neutral expression. In addition to average levels, consensus was found that anger and happiness both demonstrated increased variability and upward directed contours, while the opposite found concluded for sadness.

It was only recently that an attempt was made to quantify what has been described qualitatively as an upwards or downwards-directed contour using polynomial parameterization [25]. Results from Busso et al. suggest that while many of the standard statistical quantities like mean, maximum, and range provided the greatest discrimination power, curvature features consistently boosted their model's performance.

The agility and precision with which singers identify the pitch of a voice starkly contrasts the complexity and often erroneous estimations of many modern audio pitch tracking computer programs. Indeed, such a statement begs a minimum of two interpretations: 1. that technology in its current state lacks the sophistication to accomplish the seemingly elementary task at hand, or 2. pitch estimation itself still presents a computationally formidable problem, one that the human hearing system evolved to solve long ago. Comparative substantiation for either perspective, unfortunately, remains a topic of discussion for future biological acousticians. The following section will instead detail the difficulty of pitch estimation with regards to real and measureable acoustical characteristics that color the pitch-based patterns of speech we have grown to recognize today.

The importance of accurately measuring the fundamental frequency (*F0*) of vocalized speech is perhaps best exemplified with a qualitative example. Generally speaking, news reporters and all persons concerned with conveying unbiasedness or neutrality attempt to vocalize in a manner free of positive or negative inflection [25]. Neutral utterances predominantly feature flat pitch (*F0*) contours, and only deviate from a neutral reference when linguistic clarity necessitates, e.g. an upward inflection when posing a question. Cold, calculated, and machine-like apathy can be conveyed by draining an utterance of its intonation, while perception of neutrality is easily broken by digitally expanding the range of *F0* [26]. From a broader perspective, pitch estimation enables services ranging from speaker identification systems and hearing aids to "auto-tune" plugins found in many Digital Audio Workstations (DAW). Given its essential role in verbal and artistic expression, pitch estimation tops the list of important contributions to speech processing.

Functionals applied to pitch

Assuming the estimated pitch (*F0*) accurately represents the physical ground truth, the following section presents a comparison of pitch trends found in emotionally expressive speech. For the purposes of comparison to the present work, the affective states will be limited to anger, happiness, and sadness.

Extensive review of studies on vocal expression from 1930 to 1985 sheds light on unfortunately broad trends in pitch modulation as a function of emotion [6]. Subjective testing with a panel of judges forms a practical groundwork for validating these findings [27]. With a team of twelve professional actors, Banse et al. created an extensive corpus containing recordings of nonsense phrases spoken in fourteen emotional states. A recognition study including twelve different judges was run on this corpus and the acoustic features extracted. Given the "remarkable consistency" of the findings upon which the predictions were made [6], results found by Banse et al. differed significantly for many of the emotion categories. Many of the disparities appear split by arousal level. Regarding the results for mean *F0* in Figure 2.11, Elation, Hot Anger, and Despair all differ significantly from prediction while their lower arousal level counterparts (Happiness, Cold Anger, and Sadness) roughly agree with prediction. The effect of arousal level positively modulating the mean *F0* agrees with the results of [28].

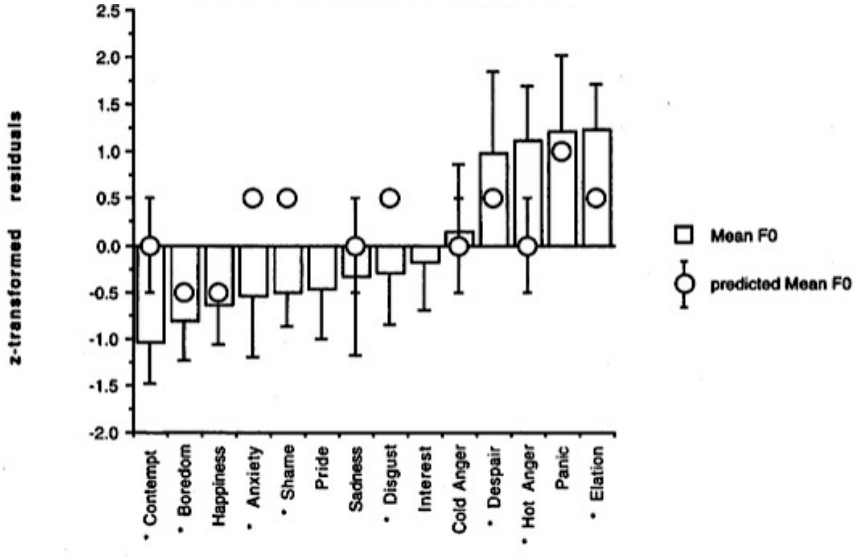


Figure 2.11 Predictions and results for changes in mean F0 as a function of emotion (From Fig. 1 Banse & Scherer 1996).

Voice quality

In order to give machines the tools necessary to quantify the emotional content of speech, qualitative characteristics must be translated into calculable terms. While adjectives such as breathy, sharp, shaky, and resonant all represent audible qualities that denote an underlying physical state of the speech production system, their digital analogs are often only tangentially correlated.

Formants

Formants offer a strong link between the measured spectra and the state of the vocal tract [29]. Humans constantly use formants to differentiate vowels; without the harmonic weighting of just three formants spoken language as it functions today would be impossible. Figure 2.12 illustrates the differences in the spectra for two similar vowels /IY/ and /IH/. The vowel /IH/ involves a constriction in the back of the throat which resembles a stressed version of the vowel /IY/. The stressed voice (dotted line) displays an upward deviation in the first peak's location (*F1*) in addition to a narrowing of that very peak. Here, the higher resonance suggests a shortening of the vocal tract, and the sharper width indicates an increase in the impedance of the walls of the vocal tract.

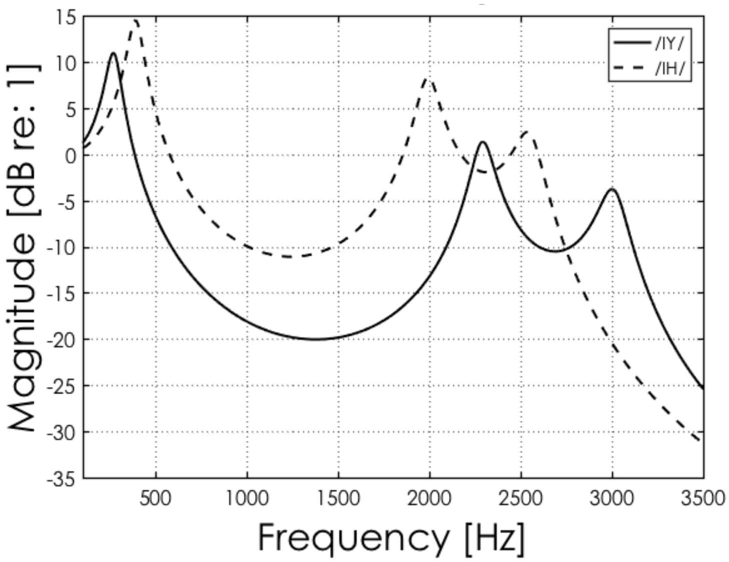


Figure 2.12 This is a plot of the ideal magnitude spectra for the vowel /IY/ (solid line) and a stressed version which resembles the vowel /IH/ (dashed line).

Unfortunately such observations in the frequency domain can be caused by more than one change in the physical state of the vocal tract. To complicate matters further, simultaneously occurring reformations at different places along the vocal tract have the potential to nullify any acoustically observable outcome. The shape the face makes while smiling can raise formant frequencies, but a relaxed and longer vocal tract due to appraisal of pleasantness could work in the opposing fashion [20].

Harmonicity

The harmonics-to-noise ratio (e.i. harmonicity, HNR) is an acoustic measure defined by Boersma (1993). *HNR* is the ratio of the energy of the periodic components to the noise of a signal [30]. Given the relatively recent advent of this specific algorithm, a smaller number of studies have incorporated *HNR* as compared to *F0*. Trends found from previous studies agree with what one might expect in comparing *HNR* values of angry, happy, and sad speech. Other studies have found that hot anger had very low *HNR* values compared to joy [31].

Jitter

The perturbations of *F0* create both a signal processing challenge as well as yet another piece of identifiable information. While mathematical functions such as the standard deviation and mean slope offer a macroscopic measure of the variability of *F0*, shorter time frame analysis is often helpful. Jitter is a measure of the variance of the *F0* taken from point to calculated point throughout the *F0* contour.

Due to limitations in the accuracy of earlier algorithms, only a handful of recent studies have been able to include jitter as a LLD of interest. Juslin (2003) found overall that anger and happiness increased the amount of jitter, while sadness decreased the jitter.

Shimmer

The measure of the short-time variability of the intensity of a speech signal is known as *shimmer*. This is calculated as the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude. Values are reported in percentages. As in the case of jitter, only recent studies in vocal expression have included shimmer as a basis of comparison. There appears to be some limited consensus with the observation that shimmer increases in the following order: sadness, happiness, anger.

Long-Term Average Spectrum

Although the voice exhibits the properties of a dynamic and very quickly changing system, features measured over a longer time scale are often of interest. In this case, the time scale specifically refers to an analysis window with a length an order of magnitude greater than the traditionally accepted time frame over which the vocal tract is approximately time-invariant. In contrast to the estimation of *F0*, the long-term average spectrum (LTAS) is calculated over a window length of 1-10 s in contrast to 20 ms. As previously discussed in the calculation of intensity, longer window lengths afford greater contributions of lower frequencies to these spectral measures. The LTAS provides an additional set of spectral information on which functional operations can be applied. These operations include calculation of the energy above and below center frequencies such as 500 Hz and 1000 Hz [32]. Other such measures include the *Hammarberg Index,* which is defined as the difference in the peak energy within the 0-2000 Hz and the 2000-5000 Hz band [33].[ii]

2.3.3 Summary of Patterns from Previous Studies

An extensive review of 104 studies of vocal expression identified several salient trends in acoustic cues pertaining to specific emotion categories [22]. Comparison of each study's results is complicated by the lack of consistency in the types of emotions that were included, the acoustic cues considered, and the baselines used as reference values for each acoustic cue. While some used cue values specific to neutral portrayals, others simply normalized the acoustic descriptor values across

[ii] Modifications to the cutoff frequency of 2000 Hz has been recommended in order to accommodate speakers with vastly different vocal ranges [40].

each emotion category as one group. As a result, absolute differentiation between findings was substituted for each author's broader interpretation of acoustic cue modulation between emotion category. Table 2.6 provides a brief summary of the consensus as of 2003.

It can be seen here that the last five cues specific to vocal expression of emotion have been studied the least since 2003. From a physiological perspective, these cues offer a great deal of information regarding the shape of the vocal tract as it may have changed from a resting position. Additionally, Scherer's predictions of stress, activation, and power more often included aforementioned physical changes which may be detected acoustically. Although it is known that several of these cues predict to some extent general changes in emotional state, there is still a need for speaker-dependent normalization [34].

2.4 Previous studies in Vocal Expression of Emotion

2.4.1 Types of Studies

There are a variety of ways to examine the communication of emotion between humans. Encoding studies look at the way people produce changes in their vocal patterns during expressive speech. These studies require a large set of speakers in order to account for personal differences in how each speaker conveys emotion. Measurement of the acoustic output of the speaker is a simple task. Obvious ethical concerns of inducing real emotion rightfully limit the assumption of total accuracy of the speaker's performance. And so participant inclusion favors those who are naturally or trained in acting apart from their true emotional state. Ideally, some objective validation of these portrayals would involve some form of non-invasive neuroimaging like functional Magnetic Resonance Imaging (fMRI). Unfortunately, requirements of minimal physical motion within the scanner render this task practically unfeasible. Therefore, researchers depend on self-reported feedback from the speaker to qualify their performance accuracy.

Table 2.6 Patterns of Acoustic Cues Used to Express Discrete Emotions in Studies of Vocal Expression (Adapted from Juslin & Laukka 2003).

	Speech Rate		Proportion of Pauses		High-Frequency Energy	
Emotion	**Category**	**# Studies**	**Category**	**# Studies**	**Category**	**# Studies**
Anger	**Fast**	28	Large	0	**High**	22
	Medium	3	Medium	0	Medium	0
	Slow	4	**Small**	8	Low	0
Happiness	**Fast**	22	Large	1	**High**	13
	Medium	5	Medium	2	Medium	3
	Slow	6	**Small**	3	Low	1
Sadness	Fast	1	**Large**	11	High	0
	Medium	5	Medium	0	Medium	0
	Slow	30	Small	1	**Low**	19

	Voice Intensity		Voice Intensity (var)	
Emotion	**Category**	**# Studies**	**Category**	**# Studies**
Anger	**High**	28	**High**	9
	Medium	3	Medium	1
	Low	4	Low	2
Happiness	**High**	22	**High**	9
	Medium	5	Medium	3
	Low	6	Low	2
Sadness	High	1	High	2
	Medium	5	Medium	1
	Low	30	**Low**	8

	F0 **(M)**		***F0*** **(var)**		***F0*** **Contours**	
Emotion	**Category**	**# Studies**	**Category**	**# Studies**	**Category**	**# Studies**
Anger	**High**	33	**High**	27	**Up**	6
	Medium	4	Medium	4	Down	2
	Low	5	Low	4	-	-
Happiness	**High**	34	**High**	33	**Up**	7
	Medium	2	Medium	2	Down	0
	Low	2	Low	1	-	-
Sadness	High	4	High	2	Up	0
	Medium	1	Medium	1	**Down**	11
	Low	40	**Low**	31	-	-

	Voice Onsets		Micostructural Regularity	
Emotion	**Category**	**# Studies**	**Category**	**# Studies**
Anger	Fast	1	Regular	0
	Slow	1	**Irregular**	3
Happiness	**Fast**	2	**Regular**	2
	Slow	0	Irregular	0
Sadness	Fast	1	Regular	0
	Slow	1	**Irregular**	4

	Precision of Articulation		Formant 1 (M)		Formant 1 (bandwidth)	
Emotion	**Category**	**# Studies**	**Category**	**# Studies**	**Category**	**# Studies**
Anger	**High**	7	**High**	6	**Narrow**	4
	Medium	0	Medium	0	Wide	**0**
	Low	0	Low	0	**-**	-
Happiness	**High**	3	**High**	5	**Narrow**	2
	Medium	2	Medium	1	Wide	1
	Low	0	Low	**0**	**-**	-
Sadness	High	0	High	1	Narrow	0
	Medium	0	Medium	0	**Wide**	3
	Low	6	**Low**	5	-	-

	Jitter		Glottal Waveform	
Emotion	**Category**	**# Studies**	**Category**	**# Studies**
Anger	**High**	**6**	**Steep**	**6**
	Low	1	Rounded	0
Happiness	**High**	**5**	**Steep**	**2**
	Low	3	Rounded	0
Sadness	High	**1**	Steep	**0**
	Low	5	**Rounded**	4

Decoding studies begin with a set of stimuli with some emotional content and specifically examine listeners' responses to the stimuli. Such work necessitates the use of numerous listeners so that within-decoder judgement biases can be mitigated. Each listener must then listen to and rate each and every stimuli in order for the researcher to determine any statistically significant effects between emotional portrayals. Unlike speaking, listening necessitates little-to-no physical motion, so the use of fMRI as a validation tool is mainly resource-limited rather than by methodology. In summary, both decoding and encoding studies are needed to fully characterize the chain of communication. Furthermore, each direction of inquiry requires the generation or access to a corpus of stimuli. The selection and generation methods vary considerably throughout the literature, and will here be discussed in further detail.

2.4.2 Types of Stimuli

All studies in vocal expression of emotion require some kind of set of stimuli to work on. Schuller et al. categorizes speech databases at the highest level by the "type of speech", followed by the "type of scenario" [21]. Speech type refers to whether the speakers were instructed to express a specific emotion ("prompted"), or if the speech was recorded without specific direction for the desired emotion. The creation of prompted speech databases usually involves recording several actors speaking a small set of generic scripted words or sentences with several types of affects. As Schuller notes, the increase in control over emotion category and textual content costs prompted speech the authenticity of origin and contextual realism. Examples of such corpora include the Munich corpus [27], the Danish Emotional Speech Database (DES) [35], Emotional Prosody Speech and Transcripts acted database [36], the Berlin Emotional Speech Database (BES) [37], the Geneva Multimodal Emotion Portrayals (GEMEP) [38], Montreal Affective Voices (MAV) [39].

Conversely, the natural gains of non-prompted speech are often hampered by excessive emotional and contextual specificity. Even more challenging are the legal concerns regarding personal privacy in the context of obtaining purely candid recordings. Methods in obtaining non-prompted speech vary from human-to-human interaction to creative implementations of robot-human interaction. For more information and examples see Schuller et al. 2011. The main message here is that although non-prompted speech provides a window into a diverse and authentic world of realistic speech, prompted speech is more practically obtained and controlled.

2.4.3 Scripts and the effects of language

In addition to declaring the target emotion, studies in prompted speech must consider the scripted content to be spoken. Several forces govern the debate over this requirement.

a. The script shouldn't produce bias in the speaker's portrayal or listener's perception of the emotion
b. The script should also be representative of realistic, naturally occurring speech.

To this end, researchers have devised creative approaches to the construction of scripts expressed in a variety of emotional styles. Table 2.7 gives a small sampling of the text given to past participants in vocal expression studies.

Table 2.7 A selection of vocal expression studies and a corresponding example script.

Author(s)	Year	Native Languages	Script Description	Script Example
Scherer et al.	2015	French	2 sentences of scrambled nonsense syllables	"Ne kal ibam soud molen!"
Liu & Pell	2012	Mandarin	35 pseudo sentences, semantically meaningless but with real words.	"她在一个门文上走亮" (like "The fector jabbored the tozz")
Patel et al.	2011	French	Sustained vowel /a/, they used 10 versions per emotion	"aaa….a!"
Lima & Castro	2011	Portuguese	16 sentences & 16 reorganized pseudo sentences.	"O quadro está na parede" (The painting is on the wall) vs. "O juadre está na pafêne"
Castro & Lima	2010	Portuguese	16 short sentences and pseudo sentences	"The painting is on the wall."
Pell et al.	2009	English, German, Hindi, Arabic	Pseudo sentences, all content words replaced with sound strings.	"The dirms are in the cindabal."
Belin et al.	2008	French	Short interjections using the vowel /a/.	"ah!"
Bänziger & Scherer	2005	German	2 meaningless sequences of syllables, derived from various European languages.	"Ha¨t san dig prong nju ven tsi."
Juslin & Laukka	2001	English, Swedish	2 sentences, 1 statement + 1 question in either English or Swedish.	"Klockan a¨r elva", "Is it eleven o'clock?"
Engberg et al.	1997	Danish	"yes", "no", nine sentences (4 questions), two passages of fluent speech.	"Jeg er ikke sulten." (I am not hungry)
Banse & Scherer	1996	German	2 sentences of scrambled nonsense phonemes from Indo-European languages	"Hat sundig pron you venzy."

While it is hardly a comprehensive list, the trends found in the table do represent much of the language used in the literature today. Often employed are generic sentences that talk about the time of day, or the existence of a painting somewhere in a room. Generally speaking, strong and evocative language is avoided. This observation follows reason: introducing a keyword or subject of conversation typical of emotional feeling presents a pathway for bias to confound the perception or expression of emotion.

A key question remains to be answered: does language have a significant effect on the perception of emotion or do humans rely entirely on paralanguage? One of the first studies to attempt to answer this question compared the subjective perception of speech consisting of matched and mismatched language-affect pairs [40]. Mehrabian & Weiner found that tone more greatly influenced the perception of attitude than the language content of their stimuli. Their study, however, considered a listener's general perception of positivity in the speaker's expression rather than recognition or quality judgements towards a specific emotion. More recent work by Morton & Trehub reinforced these findings, noting a significant age-dependence to human's focus on paralanguage [41]. Although paralanguage dominated adults' focus on emotion classification, conflicting language-affect content unequivocally increased the response latency, and "unusualness" of these utterances was unanimously reported in participating adults. If priority goes to the manner and not the message of the speaker, it would seem wise to approach the chain of communication from an acoustical perspective.

To examine the effect of a single variable in any experiment, one should ideally control for all other possible confounding variables while varying the characteristic of interest. Unfortunately, speech is rich with an incredible amount of information, so care must be taken in selecting the parts of the signal which are believed to convey emotion. Validation of each variables' contribution to the perception of emotion in speech can be done with a decoding study, where changes in perception are compared to independent changes in select acoustic properties of a filtered speech signal. When considering just the extracted pitch and intensity of a set of recordings, Lieberman & Michaels reported a reduction in correct identification from 85% (before filtering) to 47% [42]. Although higher than chance (12.5%) it is clear that a large fraction of the information necessary for the perception of emotion had been removed. Ladd et al. synthetically modulated pitch, intensity, and lastly included two versions of voice quality as produced by the speaker [26]. While differences in speaker type and text did not interact with the independently manipulated acoustic variables, both still significantly affected listener judgements throughout all three experiments.

More recently, Castro & Lima created a database of emotional speech in Portuguese in an effort to examine the effect of language on the perception of emotion [43]. Although their results support the general comparative utility of pseudosentences to sentences in terms of recognition rates, happy expressions were still significantly recognized better as full sentences. Additionally, response times between sentences and pseudosentences were not significantly different for expressions of happiness, fear, surprise, or no affect. The most promising result in favor of scrambled language content is the similarity of intensity ratings between sentences and pseudosentences across all expressions. Although a significant step forward in determining a more optimal code for prompted speech, the aforementioned scripts still lack contextually realistic content.

2.5 Contributions of this thesis

2.5.1 Limitations of previous work

Previous work seems to originate from three primary fields: linguists/speech pathologists interested in psychology, psychologists interested in acoustics, and acousticians interested in psychology. Unfortunately for the psychologists and acousticians, the speech pathologists hold the keys to the castle on this one. One of the concluding remarks from a comprehensive machine learning study that included the data of 64 emotional speech data collections since 2006 stated that "pitch, the formants, the short-term energy, the MFCCs, the cross-section areas, and the Teager energy operator-based features" are the areas of interest moving forward with emotion recognition [34]. Surprisingly enough, many of these overlap with traditional prosodic cues such as pitch and intensity (energy). Most of these features have a concrete basis for prediction and a theoretical function, and features like formants and cross-section areas directly correlate with one-another.

Schuller begins his concluding remarks on the state of the art in emotion recognition: "Obtaining more realistic data will still be the most important issue in the foreseeable future." [21]. It is easy to see why the state of emotional speech databases have had such limited positive impact: the conflicting interests of control versus realism have produced stimuli that is little of either. Although preference clearly goes to the collection of vast quantities of non-prompted speech, the lack of control over lexical content raises the variability of both subjective ratings and acoustic information [44]. Diminished counts of 'exemplar' emotional vocalizations in everyday communications further drain the pool of ideal candidates to feed into machine learning models.

3 Creation of the Speech Corpus

3.1 Objectives

Before any further elaboration, the general and specific objectives must first be stated. Three affects have been included for examination: anger, happiness, and sadness. Each encoder (speaker) will be instructed to speak a provided piece of text with each of those three affects. A fourth version will also be produced with as little expression of emotion as possible, resulting in a total of four styles of expression for each script. With four topics and two versions of wording for each topic, the speaker will be asked to apply the four styles of expression to each of the eight (8) scripts. So in summary, the participant will produce four (4) affects over eight (8) scripts for a total of thirty-two (32) stimuli samples. The labels for each of these stimuli versions is given in Table 3.1.

Table 3.1 Summary of the scripts obtained separated by script topic, version, and affect.

		Affect			
Script Topic	**Version**	**Anger**	**Happiness**	**Neutral**	**Sadness**
Dinner	a	ang-din-a	hap-din-a	neu-din-a	sad-din-a
	b	ang-din-b	hap-din-b	neu-din-b	sad-din-b
Checkbook	a	ang-chk-a	hap-chk-a	neu-chk-a	sad-chk-a
	b	ang-chk-b	hap-chk-b	neu-chk-b	sad-chk-b
Help	a	ang-hlp-a	hap-hlp-a	neu-hlp-a	sad-hlp-a
	b	ang-hlp-b	hap-hlp-b	neu-hlp-b	sad-hlp-b
Talk	a	ang-tlk-a	hap-tlk-a	neu-tlk-a	sad-tlk-a
	b	ang-tlk-b	hap-tlk-b	neu-tlk-b	sad-tlk-b

3.1.1 Participant selection

PEEP required the inclusion of mother and child pairs. Mothers would speak in an expressive style typical of anger, happiness and sadness, thus “encoding” emotion into their speech. The children were to listen to these recordings of emotional expressions, and fMRI scans made of their brains during this time. The children therefore assumed the role of the “decoder” in the communication of emotion. Each child would also listen to recordings of an unfamiliar mother, thus enabling analytical comparisons of speaker familiarity on brain activity. The primary characteristics of the participants affecting inclusion and use in PEEP were the age of the child and the fluency of the mother.

Ideally, researchers would prefer to examine how neural processing of vocal emotion changes from birth to adulthood. Scanning the brain for functional differences (fMRI) requires the head to be nearly motionless for several minutes at a time- a task that can be quite challenging for individuals in an early stage of motor control development. As a result, the success rate for scans on younger children is typically reduced. Some studies have had moderate success with children aged 8-10. But the transition in focus between language content and paralanguage occurs at a slightly earlier age. Thus the target age of the children in the present study is 7-9 years old. Child participants that had previously participated in an earlier study (PEEP I) would be excluded. Children that were included in PEEP II took the diagnostic assessment of nonverbal accuracy (DANVA) test [45]. The DANVA test measures an individual's ability to recognize emotion from a set of facial expressions and recordings of emotional speech.

Because natural speech is the focus of this study, only mothers that are fluent in English will be included. Although many cues specific to affective prosody are often universal across language, stimuli from mothers who have a strong foreign accent will not be used as the "unfamiliar" mother. The reason for this is that the question of familiarity could be more directly influenced by a child's perception of their country of origin or native language than subtler differences not caused by an accent.

Practical concerns regarding the testing methods and schedule limited the radius within which participants were recruited. As this study is examining mothers and children, the researchers were more than likely to deal with entire families when booking appointments. Additionally, differences in privacy laws between states could complicate the storage and use of the recorded stimuli. To these ends, participating families were primarily recruited from towns and cities close to State College, Pennsylvania. A fortunate benefit of this effort is the decreased likelihood of recruiting participants that speak with vastly different interstate or international accents.

3.2 Stimuli Development

3.2.1 Emotions Considered

The main objective for much of the functional analysis was to look at the effects of affect (emotion vs. no emotion), valence (positive vs. negative) and arousal (active vs. passive) dimensional differences. The first point of analysis thus requires a set of stimuli with affective prosody and without (neutral, expressionless). To compare the effects of valence which is illustrated in Figure 3.1 as the horizontal axis, both positive and negative emotions needed to be included. Emotions like anger (negative) and happiness (positive) enable this type of contrast. The arousal dimension is depicted as the vertical axis of Figure 3.1, and differentiates emotions on the basis of activity level. As both anger and happiness are on the higher end of the arousal dimension, a more passive emotion like sadness could provide a suitable contrast. In total participants were to produce vocal expressions of anger, happiness, and sadness. A non-expressive version of each stimulus would be obtained as a control.

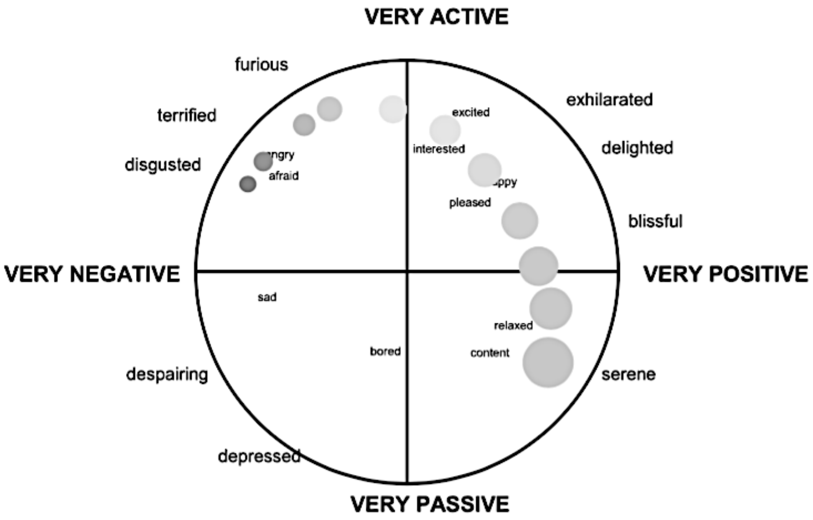


Figure 3.1 Dimensional representation of emotion (From Fig. 2 Cowie and Corneilus 2003).

3.2.2 Verbal content

The scripts were constructed to serve several objectives of the psychological study "Processing of the Emotional Environment Project" (PEEP). Its primary aim is to study children's neural processing of affective prosody using natural speech samples spoken in four prosodies: angry, happy, sad and neutral (linguistic prosody only). Most neuroimaging research on affective prosody has been conducted with adults, and the speech content of the stimuli is masked in one of several ways (e.g., using nonsense or foreign words [Banse & Scherer 1996, Castro & Lima 2010, Abrams, etc]. By removing semantic content, most studies mitigate interference from word meaning and isolate affective prosody. That benefit, however, renders the expression contextually unrealistic, and may be less personally meaningful than the human voice's natural speech. Thus, to both the encoder (participant being recorded) and the decoder (participant hearing the recordings), this cost lessens ecological validity and the import of affective prosody for real human functioning and relationships. Thus to study children's neural processing of emotion in the voice, PEEP used natural speech samples in different prosodies. To this end, the scripts were designed to be credible when spoken with happy, angry, sad, or neutral prosody. The scripts were designed to be similar to one side of phone conversations between adults. Those conversations usually consist of real and meaningful content as is illustrated by the scripts used by Shifflett-Simpson and Cummings in 1996 [46]. Through consultation with psychologists Mark Cummings and Patrick Davies, the PEEP scripts were designed to be meaningful and understandable to young children while allowing affective prosody to vary. The final scripts comprise statements, questions, and colloquial interjections spoken in a logical sequence of utterances that preserve the original topics of conversation provided by Shifflett-Simpson and Cummings. In addition to striving for greater ecological validity than nonsensical or semantically-meaningless words, the scripts were also designed to sound like one side of a phone conversations between two people. The following is one of the resulting scripts:

> *"Oh, hi, it's you."*
> *"When will you be home?"*
> *"Dinner won't be ready then."*
> *"Okay, I'll fix dinner."*

In this example the subject of the conversation is dinner, and the speaker learns that the existing dinner plan does not accommodate the schedule of the person with whom the speaker is speaking. The speaker realizes the situation and agrees to make dinner.

In total, the scripts feature four different conversations. For the sake of clarity, each of the four topics is identified by 3-letter codes: dinner (din), talk (tlk), checkbook (chk), help (hlp). Each topic had 2 script versions, labeled as "a" or "b". The complete set of eight scripts (4 topics X 2 phrasings) is shown in Table 3.2.

Table 3.2 Complete set of eight scripts by topic and phrasing version.

Topic	Version	Text
Checkbook (chk)	**a**	*"Where is the checkbook? It's gone, I can't find it. I don't have it. Do you have it?"*
	b	*"Do you have the checkbook? You had it last. It's just not here. I'll look for it."*
Dinner (din)	**a**	*"Oh, hi, it's you. When will you be home? Dinner won't be ready then. Okay, I'll fix dinner."*
	b	*"I'm fixing dinner. It will take an hour. I have a lot to do. I'll see you later."*
Help (hlp)	**a**	*"Hi, I hoped you'd call. You're running late? I will need some help. Can you change your plans?"*
	b	*"I could use your help. There's so much to do. Can you change your plans? See you when you get here."*
Talk (tlk)	**a**	*"Oh, you're tired? Sorry to hear that. We should talk. About lots of things."*
	b	*"Can you talk now? About lots of things. Money, the weekend. Okay, we won't talk now."*

3.3 Development of the Stimuli Recording Process

A number of constraints were considered when developing the protocol to obtain stimuli for the study. As is often the case in the design of systems, resources in work force, space, and time drove the iterative system design towards an optimal level. A total of three people were present throughout the recording process: a study participant (the mother being recorded), a vocal coach, and a recording engineer. The primary role of the recording engineer (author of this thesis) was to digitally capture the vocal expressions of the subject and ensuring the highest degree of audio quality possible.

3.3.1 Recording Location

All stimuli measurements took place at the Pennsylvania State University, University Park Campus in the Moore Building, Room 214. The floor plan was divided up into three smaller rooms: 1. meeting room, 2. recording room, and 3. control room pictured in Figure 3.2. The control room is located on the left side of Figure 3.3, and the recording room is pictured in the middle.

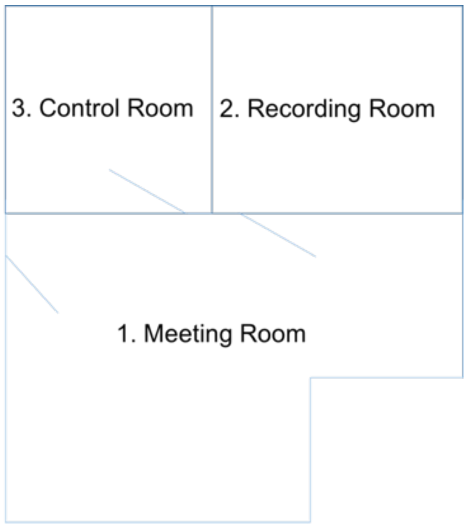


Figure 3.2 Floorplan of the testing location, Moore Building, Room 214.

Acoustical Considerations

The acoustical quality of the recordings was assured from the evaluation and optimization of reverberation, isolation, and noise levels. Each of these evaluations was aimed at ensuring that only the sound of the participant would be recorded as they encoded emotion into their speech patterns. An environment with low reverberation is ideal because fewer, lower amplitude reflections from surrounding surfaces will contaminate the direct sound from the participant speaking into the microphone. The recording room was relatively dry. This was likely due to the wall and ceiling treatments, which feature perforated metal instead of tile or plaster. The mid-frequency average reverberation time of the room was measured using a Brüel & Kæjr 22560 sound level meter at T30 = 70 ms (400Hz-1.25kHz).

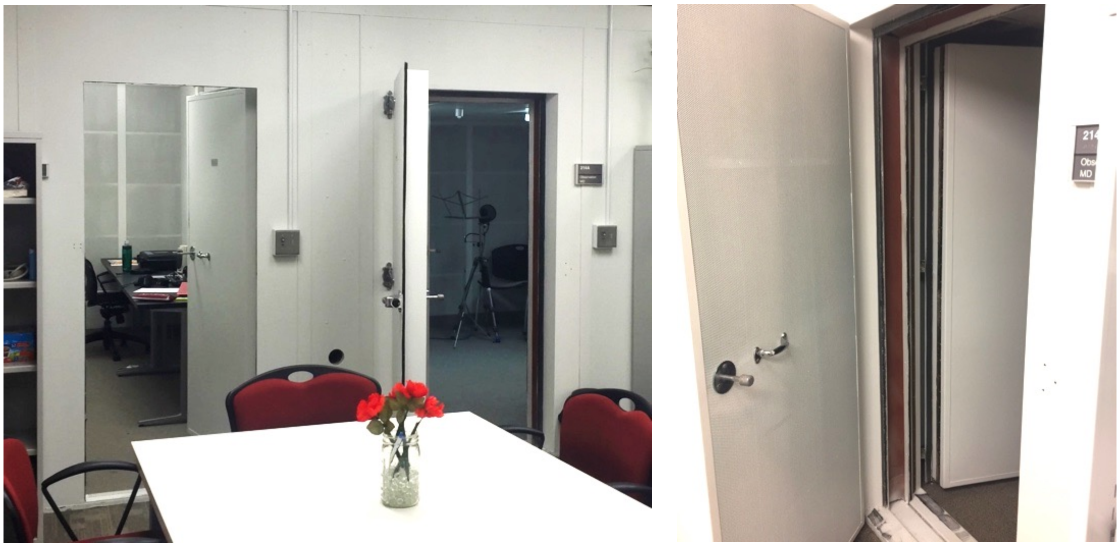


Figure 3.3 On the left is a picture of the three rooms: 1. Meeting Room (foreground), 2. Recording Room (background right), 3. Control Room (background left). The picture on the right shows the gap between the walls of the connected rooms.

Isolation from exterior environmental noises was of great importance. Without proper isolation, conducted vibrations or conversations from surrounding offices could travel into the diaphragm of the microphone and corrupt the signal of interest. Separating and acoustically isolating the location of the recording engineer and the source of the stimulus helps ensure that the sound engineer monitors only the sound that is entering the recording device. Proper monitoring is critical to protecting the quality of the measurement; the sound engineer and the microphone "hear" sound in vastly different ways. Between the microphone's frequency response, directivity, and physical placement with respect to the encoder, the signal leaving the microphone has already been altered several times. For example, bumps and vibrations conducted into the microphone element via its rigid stand may not be audible to the sound engineer without listening to the recorded signal. The most accurate place to monitor the signal as it is being recorded was either at the sound card or from the live playback via the Digital Audio Workstation (DAW) in a separate room from the signal source. It was therefore necessary to create a pathway for the signal to travel from the microphone to the audio recorder.

The physical treatment of both the recording room and control room met the goals for strong isolation. There are a set of double-doors separating the meeting and recording rooms, and just a single hinged door separating the meeting and control rooms. Each of these doors is approximately 6 inches thick, and is manufactured by the Suttle Corporation (Lawrenceville, IL) as illustrated in Figure 3.3. Between the control room and the recording room are air gaps created by rubber spacers on the exterior sides of the floors and walls. Although sound isolation was not measured directly, loud conversations from either the control or recording rooms were virtually silent with the doors to both rooms closed. Solutions to make it easier to communicate between the rooms are investigated in later text.

To maximize the signal-to-noise ratio at the microphone end of the signal chain, modifications were made to both the microphone placement and microphone directivity selected. The loudest source of noise in the recording room was determined to be the ceiling vent diffuser, which was located at the corner closest to the door. Using a Brüel & Kæjr 2250 sound level meter, the A-weighted equivalent noise level was measured to be 34.2 dBA (re: 20 μPa) at 0.5 meters from the air diffuser. The maximum difference in noise level was found at the opposing corner of the room at 24.8 dBA (re: 20 μPa). This was determined to be the optimal location for the microphone as shown in Figure 3.4. Further noise suppression was achieved by setting the microphone to a cardioid directivity, and aligning the null region towards the noise source at the opposite corner of the room.

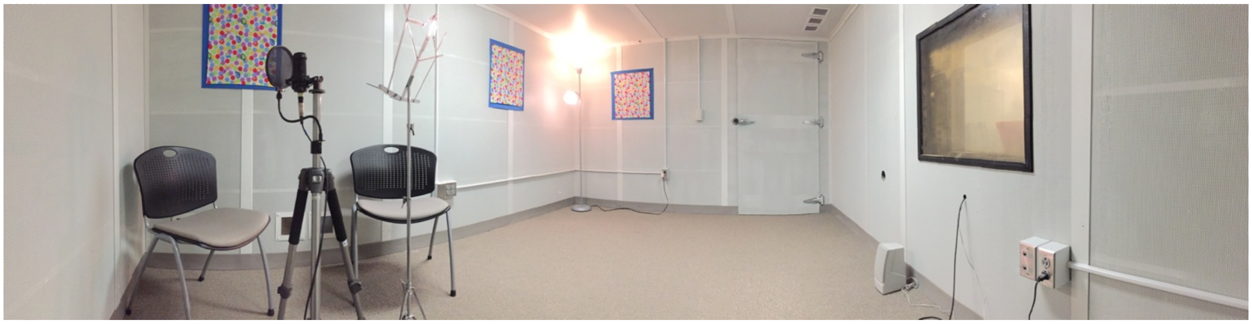


Figure 3.4 Panorama view of the recording room. From left to right: participant's chair, microphone, vocal coach's chair, door, air vent (top), monitor (bottom), and viewing window.

3.3.2 Design of the Signal Chain

Chefs often claim that a meal is only as good as its ingredients. This philosophy is paralleled in the design of the signal chain. In this case, the “meal” is the digital recording, and the ingredients are the sounds propagating from the subject. As it is the emotional qualities of the subject’s vocalizations that determine the better product, personal comfort and care of the participant take an equal, if not higher priority over the acoustical treatment. The source of interest is not a reciprocating machine with an on/off button, but a human subject who is being placed under a “microscope” and asked to emote in a personally accurate manner. From the microphone placement to the patch cables nested in the control room, every modification supported the principle goal of measuring the best possible affective prosody. The next section of text will briefly cover the evolution of the measurement system with respect to modifications targeted at augmenting higher-level functionality.

Initially, the most basic requirements for the measurement system were that a pathway exist for sound to be recorded and monitored, and a pathway exist for the sound engineer to communicate with the vocal coach. With all doors firmly closed, the only practical means of communication without the aid of synthetic amplification or transduction is simply hand gestures communicated visually through a window measuring roughly 2'X2', pictured in Figure 3.5. These initial requirements were met by the first version of the measurement system depicted in a schematic in Figure 3.6.

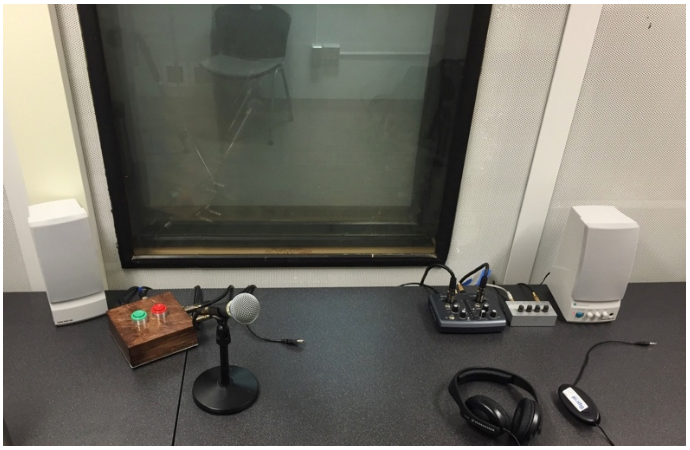


Figure 3.5 This is a picture of the window connecting the control room and the recording room.

The signal path of the stimuli moves in one direction from the recording room to the control room. Traveling initially as propagating acoustical energy from subject to the microphone, the energy is then transduced into an electrical signal represented by the solid red line. Once delivered to the sound card (an M-Audio M-Track USB interface), the sound engineer can then monitor the penultimate recording over a set of headphones just as the signal is saved to the computer's memory. Lastly, walkie-talkies provided a simple mechanism for two-way communication between the sound engineer and the vocal coach.

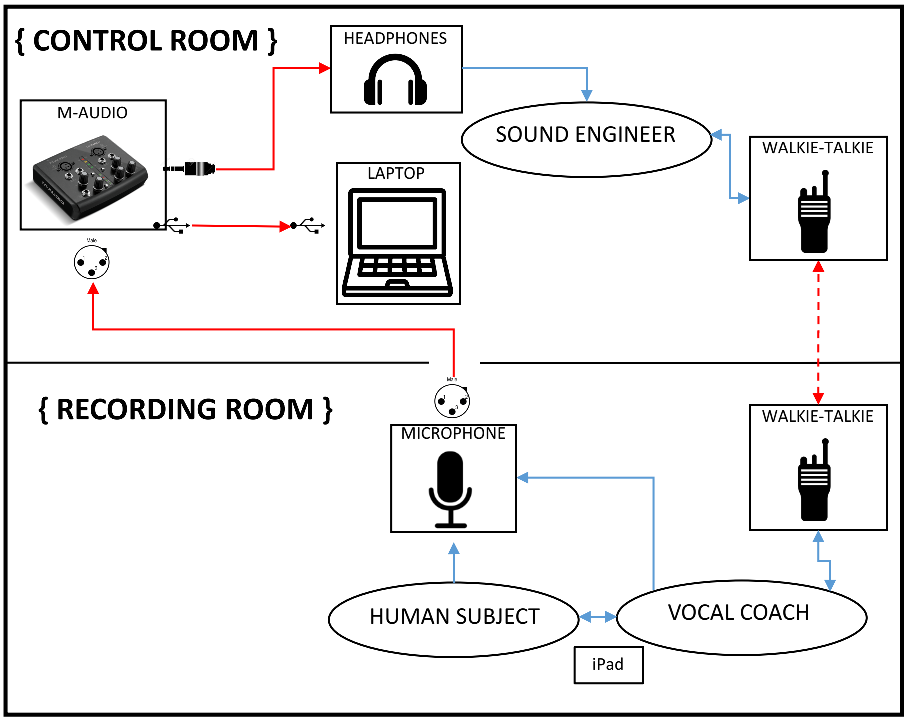


Figure 3.6 Original signal chain. Blue lines indicate acoustic transmission, solid red lines indicate wired electrical signal transmission, and dotted red lines indicate wireless signal transmission.

Although conveniently simple as a first step, experience gained from pilot tests revealed several areas where technical modifications would significantly benefit the process. During piloting, it was determined that a standard set of exemplar recordings be created for the subjects to listen to on request. As the only pair of listening devices was located in the control room, the subject would have to move from their seat in the recording room to the control room. This process could alter the distance between the microphone and the subject in the middle of a recording, which may boost or lower the amplitude beyond the previous amplitude range. This configuration applied too much logistical pressure on the vocal coach, who was tasked with both communicating directly with the subject, holding the walkie-talkie, as well as taking notes quickly in real time. A hands-free alternative to the walkie-talkie would mitigate the dropping of equipment and return focus to coaching the subject. A higher quality set of speakers than the walkie-talkies had would also boost the clarity of communication between the sound engineer and the other room. The headphones used by the sound engineer, although comfortable at first, would have to be worn for recording sessions lasting almost two hours. If the recording room were to be equipped with a proper set of speakers then the control room could easily receive the same treatment to the cumulative comfort level of all. Figure 3.7 illustrates how each of these issues were solved through technical means.

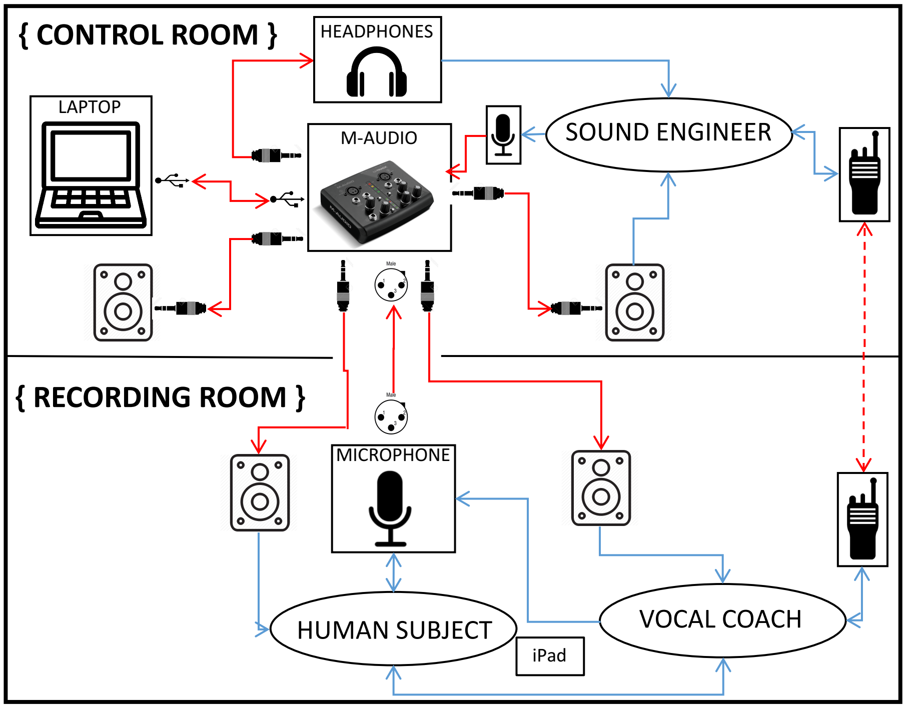


Figure 3.7 System setup #2, featuring a set of monitors in both rooms for example stimuli playback.

A set of monitors in the recording room provided a convenient pathway for the subject to hear their own recorded takes in addition to the standard set of exemplars stored on the laptop. A secondary microphone in the control room largely nullified the use of the walkie-talkies, which could then be used as backups to the new system. The problem of the speed of communication remained. Although there is a path between the sound engineer and the monitors in the recording room, the steps required to initiate the communication without corrupting the recording were multitudinous and often difficult to access quickly. It was concluded that a simple physical button could replace many of these software hurdles when the sound engineer wished to communicate with the other room.

The third and final iteration of the measurement system provided support in all of areas where improvements were needed and was chosen as the best of the options presented. The one difference between setup #2 and setup #3 (see Figure 3.8) is the introduction of the “Push-to-Talk” box.

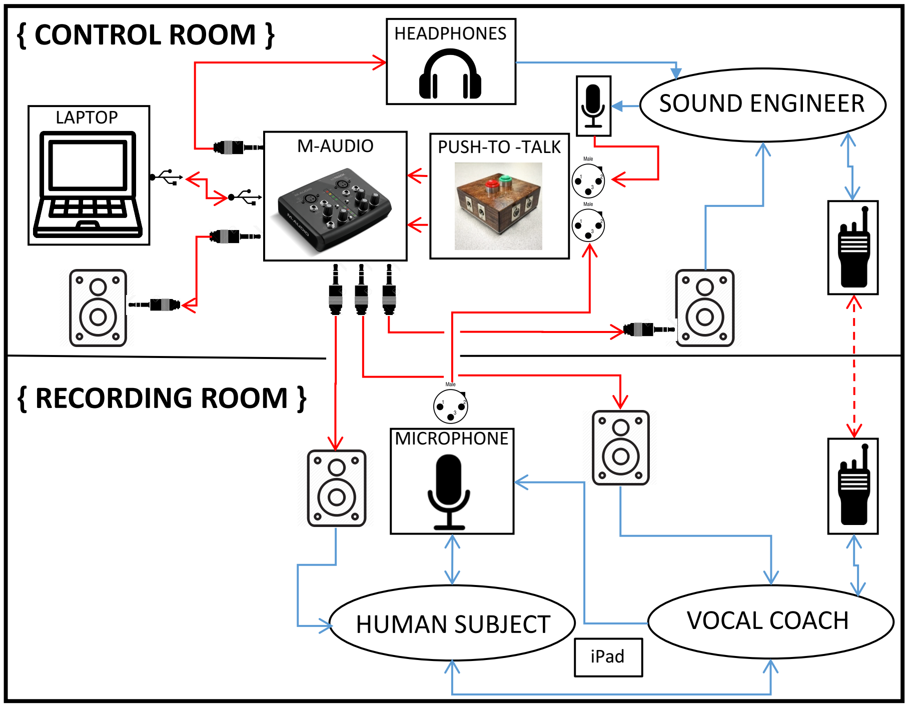


Figure 3.8 Measurement setup #3 features monitors in both rooms as well as an easy-access "Push-to-Talk" box in the control room.

This box features two XLR inputs, two XLR outputs, and a large button between each input-output pair (see Figure 3.9). The green button wired such that a connected microphone is normally muted until the button is pressed. Conversely, the red button is wired to mute a connected microphone when it is pressed. With this simple device, the sound engineer could leave their microphone channel open in the DAW software, and simply press the green button to talk when necessary. With an open loop between the recording room microphone and monitors via the control room's microphone and monitors, the red button could then be pressed to eliminate any potential feedback. See Appendix A for details on the circuitry. The initial complications of wiring the system correctly on the control room side were far outweighed by the hands-free and speedy communication that resulted between both rooms.

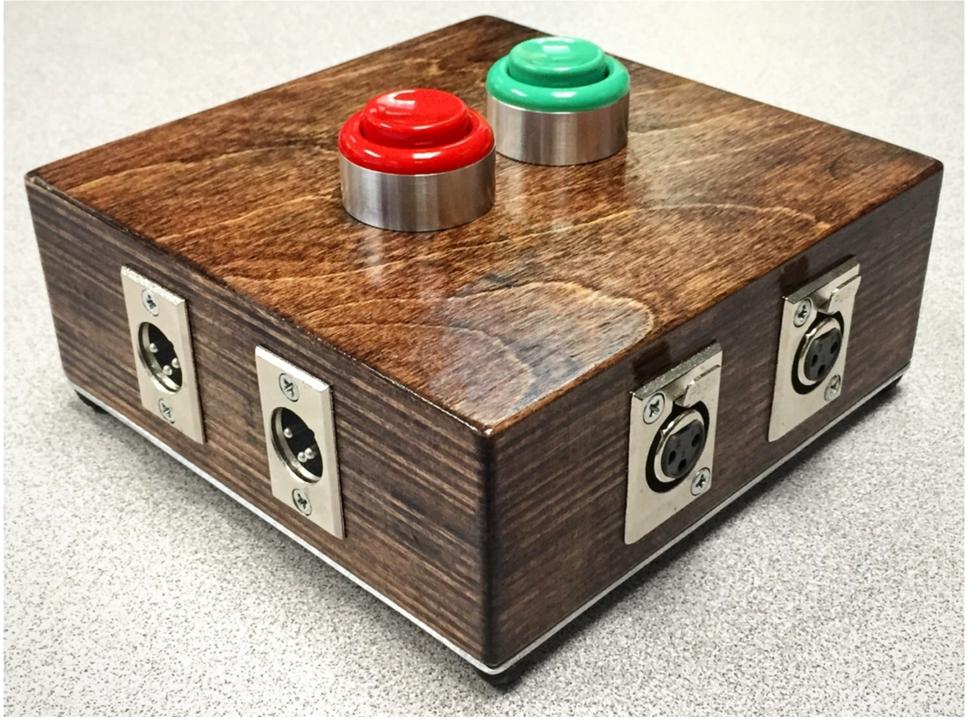


Figure 3.9 This is a picture of the "Push-to-Talk" box. This device features two input-output pairs each accompanied by a button that mutes (red) or unmutes (green) the connection to the microphone.

Source Considerations and Equipment

Several characteristics of the sound sources (the mothers speaking) that were recorded required attention when finalizing an equipment list. The dynamic range was predicted to vary between soft speech (30-40 dB) to shouting (80-95 dB). Accordingly, the selected microphone should be capable of recording sounds at amplitudes within that range. With an expected range of meaningful frequency between 100-8000 Hz, the microphone's cardioid response in this range shouldn't deviate significantly. An Audio-Technica AT-2050 large diaphragm condenser microphone was selected as a candidate due to its practical accessibility and relatively good adherence to the prescribed technical requirements. The magnitude frequency response of the microphone as provided by the manufacturer is shown in Figure 3.10. From this response plot it can be seen that the largest deviation occurs at roughly 4kHz with a peak of less than 5 dB. Included in the purchase of the microphone was a shock mount that would suspend the microphone's body with a system of stretchable cords as depicted in Figure 3.11. This mount would introduce an impedance mismatch between the stand and the body, therefore attenuating possible structural vibrations from the floor. More technical specifications can be found in Appendix B.

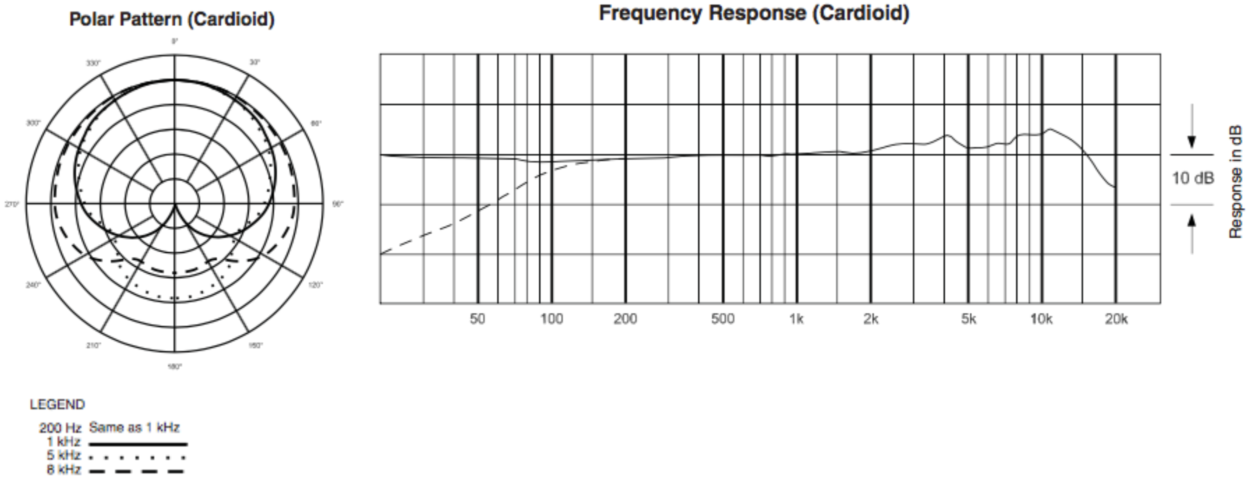


Figure 3.10 Directivity and magnitude frequency response of the Audio-Technica AT-2050. From www.audio-technica.com.

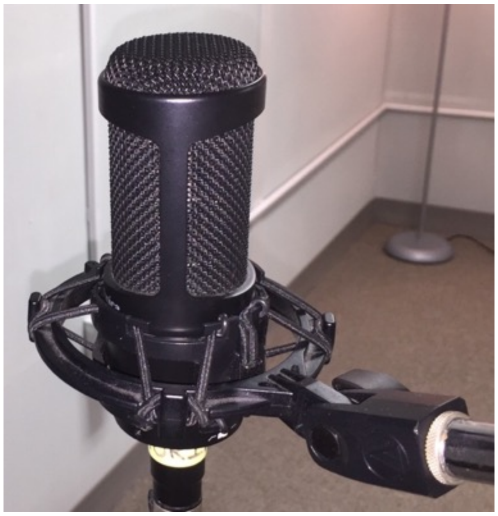


Figure 3.11 This is a picture of the condenser microphone seated within a shock mount. The shock mount protects vibrations from the attached stand (bottom right) from transmitting into the transducer element.

Also of necessity was an external sound card by which the sound engineer could increase or decrease the gain of the input microphone's signal with respect to the recorded output signal. An M-Audio M-Track was selected on the basis of its functionality and financial accessibility. This interface consists of two input channels and two output channels. Each input channel supports +48v bias voltage necessary for the selected condenser microphone, and each output channel could be routed to the control room and recording room independently. Lastly, this device was capable of recording at 48kHz, which is many times higher than the range of interest. Each sample could also be measured with 24-bit depth resolution. For more information on the M-Track interface, see Appendix B.

Because it was not known exactly how much frequency or amplitude resolution some of the future signal processing might benefit from, a sample rate of 48kHz and bit-depth of 24 bits per sample were chosen. At the very least, this sample rate more than accommodates the sampling of measured frequencies beyond the range of human hearing. Given a commonly accepted upper limit of 20kHz [47], the Nyquist criterion holds that aliasing from frequencies below his upper limit can only be avoided by sampling at the Nyquist Rate, which is twice this frequency or 40kHz and above [11]. The chosen maximum sample rate of 48 kHz accommodates this criterion with room to spare. A low-pass filter provides protection from aliasing from frequencies above 24 kHz. Although just the major components of the measurement system have here been described, a comprehensive list of equipment is provided in Appendix B.

3.4 Recording Procedure Development

With people, equipment, and a set of objectives declared, the next steps for the project were to integrate, apply, and refine the cooperative effort where necessary. This section will delineate the basic structure of the stimuli collection process, followed by a short summary of the feedback received from piloting and modifications that were made to optimize the process.

3.4.1 General experimental format

The task for each participant (speaker) in its most fundamental form was to produce 32 exemplar recordings of speech with a style of vocalization representative of the way they produce each affect.

The general procedure was conducted as follows:

a. The participant arrives at the testing location, the Moore Building Room 214, in University Park, PA. The participant and the project coordinator then review and sign the informed consent forms.
b. The vocal coach then presents a behind-the-scenes interview excerpt from the movie “Inside-Out (2015),” in which each voice actor demonstrates multiple versions how they speak when attempting to embody their respective character roles named “Joy”, “Sadness”, and “Anger” [48].
c. The participant is then given the booklet containing each of the eight scripts to practice speaking before entering the recording room.
d. After 2-3 minutes of reading and practicing on their own, the participant is seated in the recording room in front of the microphone.

e. The vocal coach proceeds to rehearse the scripts with the participant. During this time, the sound engineer adjusts the gain of the sound card to accommodate the loudness of the participant. This gain is recorded on the script order sheet for future reference.
f. The measurement order is organized by affect, and randomized over script (see Appendix C). Beginning with the neutral vocal style, the participant is presented with each of the eight scripts in a randomized order.
g. The participant speaks each script for a minimum of three takes: one practice take to familiarize the speaker with the text, and two takes with full effort.
h. Feedback is given by the vocal coach and audio engineer after each take regarding naturalness or pronunciation.
i. After speaking each of these eight scripts without emotion, a new randomized order of scripts is presented to the participant to be spoken with an angry affect. This process repeats through the sad and happy affect.[iii]

A critical question remains unanswered: how do the test administrators, the vocal coach and audio engineer, determine that a particular sample stimulus is good enough to move on to the recording of the next stimulus? Several factors influenced the provisional criteria for assessing the quality of each participant's vocal expression. At a fundamental level, the characteristics of the target expression had to be defined with respect to the participant's ability to act, and the intended use of the stimulus for the rest of the study. In the first place, PEEP sought stimuli that demonstrated affective prosody that was ecologically valid to each participant i.e. the expression of anger should sound and feel natural and accurate to how they would normally vocalize anger. Secondly, the stimuli from one participant would be presented to both their own child as well as the child of another participant. Consequently, each decoder (child participant) would eventually listen to the stimulus of a familiar voice (their mother) and an unfamiliar voice (different mother). This stimulus presentation design reflects a key objective of the fMRI analysis, which is to examine how familiarity alters children's processing of affective prosody from mothers. Thus, the affect should be held constant and recognizable between each mother's expression without significantly altering qualities unique to each mother's voice.

Given these many requirements to achieve high quality stimuli, additional procedural tools were used to assess the quality of the recordings. During the measurement process, it was determined that objective assessment of each stimulus would both prolong the procedure beyond practical means and potentially (and synthetically) fit every encoder's expressions to potentially foreign prosodic

[iii] By holding the affect constant and changing the script, participating speakers are given more time to search and refine ways to act out an emotional state according to memory and feeling that they may not be experiencing at the very moment of making the recordings.

patterns. To this end, the test administrators would offer general direction to express a given emotion by incorporating cues over which there is broad consensus within the literature [3] [22] [24] [27] [43]. For anger, subjects were encouraged to be louder, more abrupt, and direct their pitch downwards during phrases where such motion was typical. For happiness, it was recommended for subjects to raise their average pitch level, direct their pitch in an upwards motion, and speak quickly. During expressions of sadness, each participant was encouraged to be quiet, to not raise their pitch, and to lengthen the duration of their phrases.

3.4.2 Results of Piloting the Recording Procedure

This procedure was piloted from start to finish for eleven (11) female participants (mean age ≅ 32 years) in an effort to detect where improvements could be made. Several valuable lessons were learned throughout this process and each influenced the quality and efficiency of the procedure in some way. While some participants were quite capable of acting out emotions apart from the emotional state they may have been experiencing, several others required substantial assistance in vocalizing anything apart from a neutral tone of voice. Table 3.3 provides a summary of the number of recording takes that were needed to obtain acceptable stimuli during the piloting phase, where number of takes are the total number of attempted portrayals excluding practice readings. Besides the first entry, counts exceeding 64 (2 trials * 32 stimuli versions) takes indicate additional measurement attempts. Generally speaking, the number of takes and the duration of the session correlated with participant's difficulty producing quality affective expressions. The rows highlighted in gray are for this pilot study. As points of reference, information from two other recording sessions were included for the principal investigator and the vocal coach.

Table 3.3 Summary of the results from piloting the stimuli recording procedure. Rows highlighted in gray indicate sessions where excessive guidance was given. The thick box border surrounds entries from participants that are mothers.

Visit Date	Original ID	Mother?	Duration (min)	# Takes	Notes
7/23/15	9999	Yes	32	32	Is the Principal Investigator. Only recorded 8 takes per affect.
9/24/15	9998	No	45	69	Is the Project Coordinator.
10/22/15	9997	No	88	119	Undergraduate Research Assistant
*10/23/15	9996	No	68	105	Undergraduate Research Assistant
10/29/15	9995	No	78	74	Undergraduate Research Assistant
10/30/15	9994	No	49	100	Undergraduate Research Assistant
11/9/15	9993	No	50	69	Graduate Research Assistant (not part of project)
2/24/16	900	Yes	46	76	1st pilot participant with child.
2/27/16	901	Yes	65	70	Co-investigator. Voice was very tired. Constructive conversation lengthened recording duration.
2/27/16	902	Yes	59	69	-
3/4/16	903	Yes	48	67	-

* began monitoring number of takes.

Unfortunately, the exceptional quality and ease of the initial participants (9999-9998) drove the expectation for acting quality higher than was perhaps necessary for the remaining participants (9997-9994). This meant that for underperforming pilot participants, a far greater number of takes were recorded for each stimulus version therefore extending the measurement session well beyond expected lengths. This realization necessitated the introduction of a cap on the number of takes each each script and affect could receive, thus limiting the cumulative fatigue the participant and test administrators may experience. Additionally, each take received undue scrutiny, often drawing the vocal coach and sound engineer to micro-manage until the optimal vocalization was achieved. The effect of this was evident in the excessive patterning of expression and reduction of individuality that resulted. Thus, greater emphasis was given to seeking the participant's natural expression of emotion rather than strict abidance to stereotypical examples.

3.4.3 Voice coaching

Although there are many widely accepted global patterns of affective prosody [3] [22] [24], spontaneous and more unguided expressions are of greater need in a field inundated in prompted speech [21]. The current process of subjectively evaluating the quality of each stimulus overemphasized conformity to predetermined scientific trends. While the product of these methods displayed close adherence to previous cues found in the literature [3] [22] [27], the body of work as a whole still lacked diversity in emphasis placement, pitch stylization and other basic characteristics of expressive speech.

To address this issue, a less direct approach to vocal coaching was formulated. Before each affective category, the vocal coach would define the specific type of affect in terms of groups of narrower emotions. For example, the anger category was more specifically defined as a highly aroused hot anger or rage as opposed to cold anger or annoyance. Happiness was equated with elation or joy rather than contentment or placidity. Sadness was defined more specifically as a depressed, sullen state rather than desperation. The vocal coach would then ask the participant to recall a scenario in their life during which they most strongly felt the particular emotion to be expressed. Additionally, the participant was encouraged to assume a facial expression in accordance with the emotion e.g. smiling for happiness, brow furrowed for anger, upward slanted eyebrows and a frown for sadness. Next, the participant read through the script while attempting to convey the target affect as accurately and intensely as possible.

If the resulting speech contained any lexical or semantic errors with respect to the script, or was pronounced differently from typical American English, then the vocal coach or the sound engineer would offer the appropriate correction. Subsequently, the vocal coach would ask the participant how natural and accurate they felt about their most recent portrayal of emotion, and what stylistic changes might benefit future attempts. If a participant was experiencing considerable difficulty acting out in any way whatsoever, the vocal coach would then offer examples of adjectives and adverbs to describe typical vocal cues without directly prescribing the cue itself. A list of the affect categories and their corresponding perceptual descriptors is provided in Table 3.4.

Table 3.4 Target affects or expressive style with their respective descriptors.

	Target Affect			
	Anger	**Happiness**	**Sadness**	**No Expression**
Descriptors	sharp	melodic	subdued	removed
	biting	sing-song	low-energy	complacent
	harsh	chirpy	exasperated	apathetic
	abrupt	sprightly	lethargic	robotic
	frontal	chipper	depressed	flat
	raised	liltingly	exhausted	factual
	demanding	quick	hopeless	news-report
	coarse	hopeful	breathy	

A close examination of these lists reveal descriptors pertaining to environment, linguistic content, physical feeling, and audible perceptions of a given affect or expressive style. Words such as demanding, delighted, hopeless, and apathetic describe a particular framework for delivery or reception of information within a conversation. For example, a question posed in a hopeful (happiness category) might be delivered as more of a demand when angry. Descriptors like sharp, melodic, breathy[iv], and flat have a more direct meaning for a particular style of speaking [can you add a ref here?]. The information in Table 3.4 was included in the final version of the script order data collection sheet, which is provided in Appendix C.

Additional piloting with mothers from the State College area (900-903) showed marked improvement not only in the length of the appointments but the ease with which more descriptive instructions were followed. Overall, active measurement sessions lasted for 45-75 minutes, and the average number of takes was reduced to less than 80. The next section will detail additional requirements related to the presentation of the stimuli to the decoders and the steps followed to meet those demands.

3.5 Stimuli Post-Processing

3.5.1 fMRI requirements

The neuroimaging procedural design and environment created a variety challenges that either directly or indirectly shaped the characteristics of the set of stimuli. Organized in what is known as a block-related design, the fMRI testing procedure had participants listen to each stimulus in exactly ten second length blocks as depicted in Table 3.5. This put an upper limit on the cumulative length for each of the four utterances within a stimulus.

Table 3.5 This is the order in which the stimuli were presented to the decoder in the MRI scanner. Note that each volume or scan sample takes 2 seconds. The blocks where speech is presented are exactly 10 seconds in length.

Time (s)	**0-4**	**4-14**	**14-20**	**20-30**	**30-36**	**36-46**	**46-52**	**52-62**	**62-68**	**68-74**
# Volumes	2	5	3	5	3	5	3	5	3	5
Stimulus	Silence	Speech	Silence	Speech	Silence	Speech	Silence	Speech	Silence	Speech
Encoder		Unfamiliar		Mom		Unfamiliar		Unfamiliar		Mom
Affect		Neutral		Sad		Angry		Happy		Sad
Script		1a		2a		1a		3a		1a

[iv] The word “breathy” was further qualified as more of a way to relax the voice and avoid tensing the vocal cords as could occur in expressions of desperation and anxiety.

Perhaps the greatest adversary to speech perception, however, was the noise level within the scanner. Measured at more than 95 dBA within the bore of the scanner with a peak frequency of 1.2 kHz, the operational noise level of the scanner is above the permissible exposure limit for an 8 hour day. OSHA's exchange rate of 5 dBA indicates that hearing loss may occur at 4 hours of exposure for scanner noise at 95 dBA [49]. Such high noise levels necessitated the wearing of hearing protection while in the scanner, which provided an estimated 25-30 dB noise reduction [50].

3.5.2 Stimuli extraction

To avoid losing any quality stimuli as a result of neglecting to initiate the recording, the measurements were saved as four continuous waveforms, each corresponding to one affect. Individual takes from each of these long waveforms were assessed for noise intrusion and relative prosodic quality, then extracted into smaller waveforms ranging from 6-15 seconds in duration. Within each extracted waveform, regions where loud breathing and various pops and clicks (usually from saliva bubbles bursting in the mouth) that were not semantically meaningful were manually silenced within the Digital Audio Workstation (DAW) [51]. Although laborious at a first glance, the manual editing actually reduced the work and time spent in later processing steps. The upper plot of Figure 3.12 gives the original waveform in blue, with the manually selected clicks and pops highlighted in red. These sections would have likely been erroneously retained with an automated process. The lower plot of Figure 3.12 represents the speech signals as they were processed for endpoint detection. Amplitudes of either the blue (vocaic energy) or red (consonants) below the threshold (dashed line -45 dB) would be considered part of the utterances. The text of the scripts is provided above the waveforms as reference. From the lower plot of Figure 3.12 on the word "you", it appears that a short, low-amplitude sound occurred followed by a loud sustained sound. This indicates that the mouth made a small popping sound shortly before the start of the vowel.

Acoustically, the noise of the unwanted pops and clicks look very similar to meaningful consonants like the "t" in "last." To avoid loss of meaningful content, all utterance endpoints were assessed through a listening test. Often preserved were small, yet subtle sounds that were actually unvoiced fricatives or breaths located at the end of certain words. The refined waveforms were then exported into separate files and loaded into the MatLab environment [52].

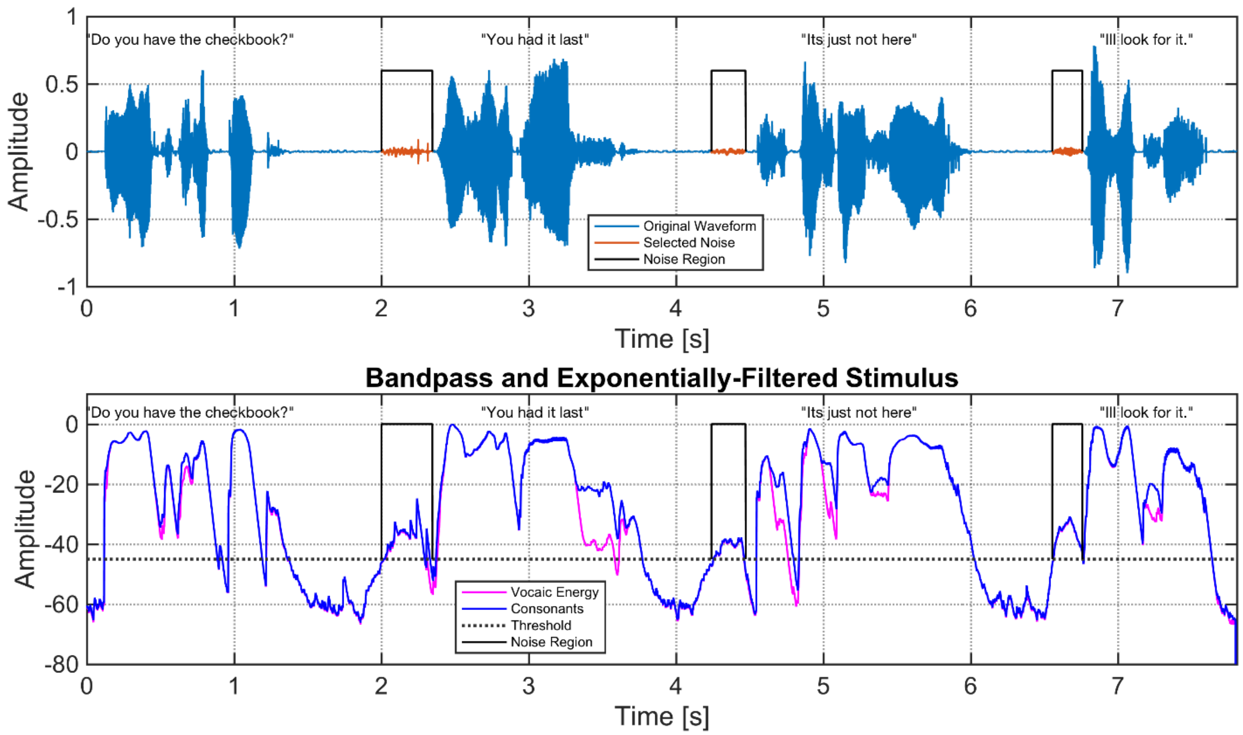


Figure 3.12 This is a plot of a waveform that contains breaths and various pops and clicks. The features that were removed are highlighted in red, as they had no meaningful content.

3.5.3 Signal Processing for parsing and normalizing

Parsing

Once the stimuli were loaded into the MATLAB environment, several steps were taken to ensure that each stimulus could be presented and heard according to the requirements of the MRI environment. The first objective was to determine bounds (in samples) of each of the four utterances. Next, 200 ms and 100 ms of silence were padded to the beginning of the first utterance and end of the last utterance. Finally, equal lengths of silence (3 total) would be created between the four utterances such that the total length of the waveform was exactly 10 seconds * 48kHz or 480,000 samples. Although a script was written to automatically detect the sample bounds of each utterance, the relative size of the inter- and intra-utterance silences was not consistent enough for the algorithm to handle, as shown in Figure 3.13.

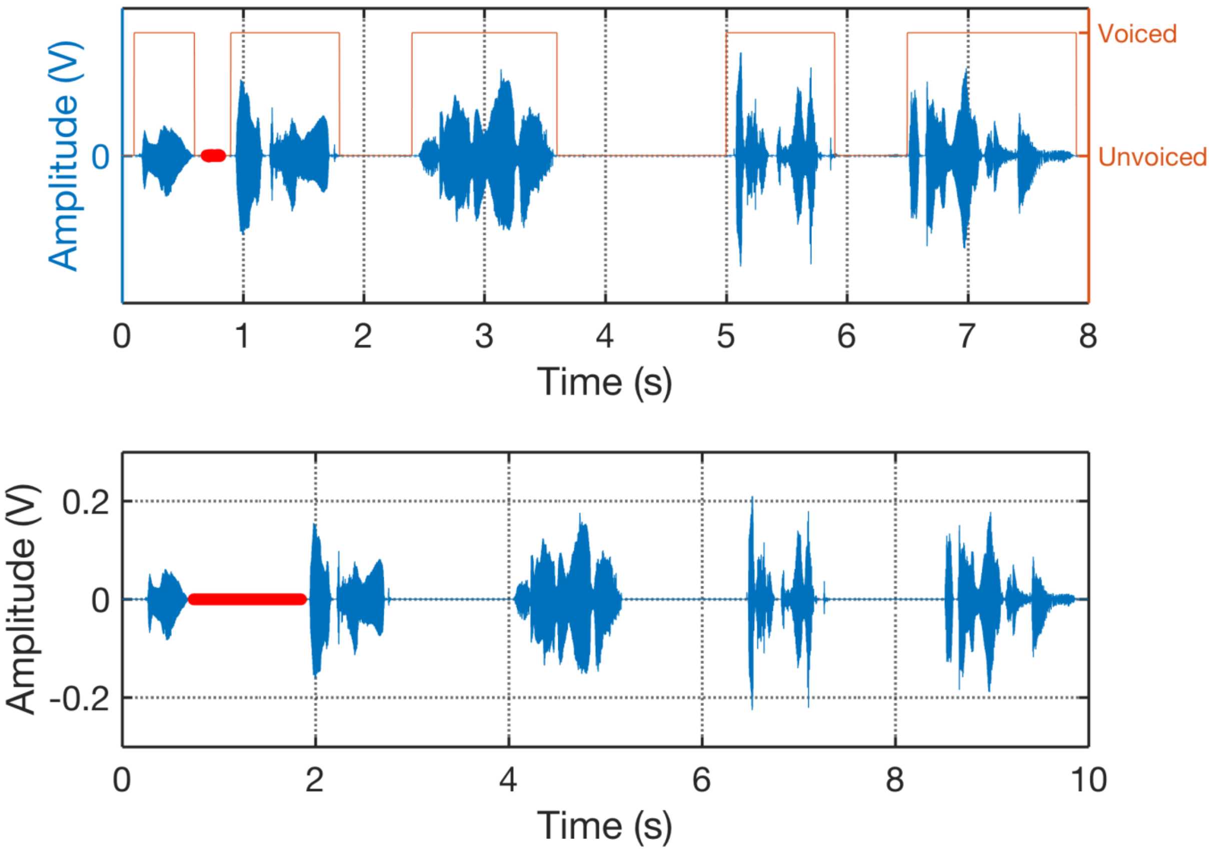


Figure 3.13 The upper plot is a waveform that has a mid-utterance silence and a between utterance silence that caused an error in the program. The lower plot is the output of the algorithm, which has severed one utterance at the location of the long silence (highlighted in red).

Although solutions to this problem are sure to exist, the benefit of speedy automation did not appear to outweigh the cost of stimuli corruption and time allocated to the pursuit. Thus, a small amount of manual input was added to the script whereby the user would click on the between-utterance silent regions of a plotted waveform (see Figure 3.14) and enter these coordinates into the rest of the program. From this location seed, the algorithm expanded the beginning and end point of the silent region sample by sample, until the amplitude of the waveform at both bounds exceeded a predetermined threshold. The splicing code can be made available upon request. With the bounds of each utterance known, their relative amplitudes could then be compared and adjusted as a final step.

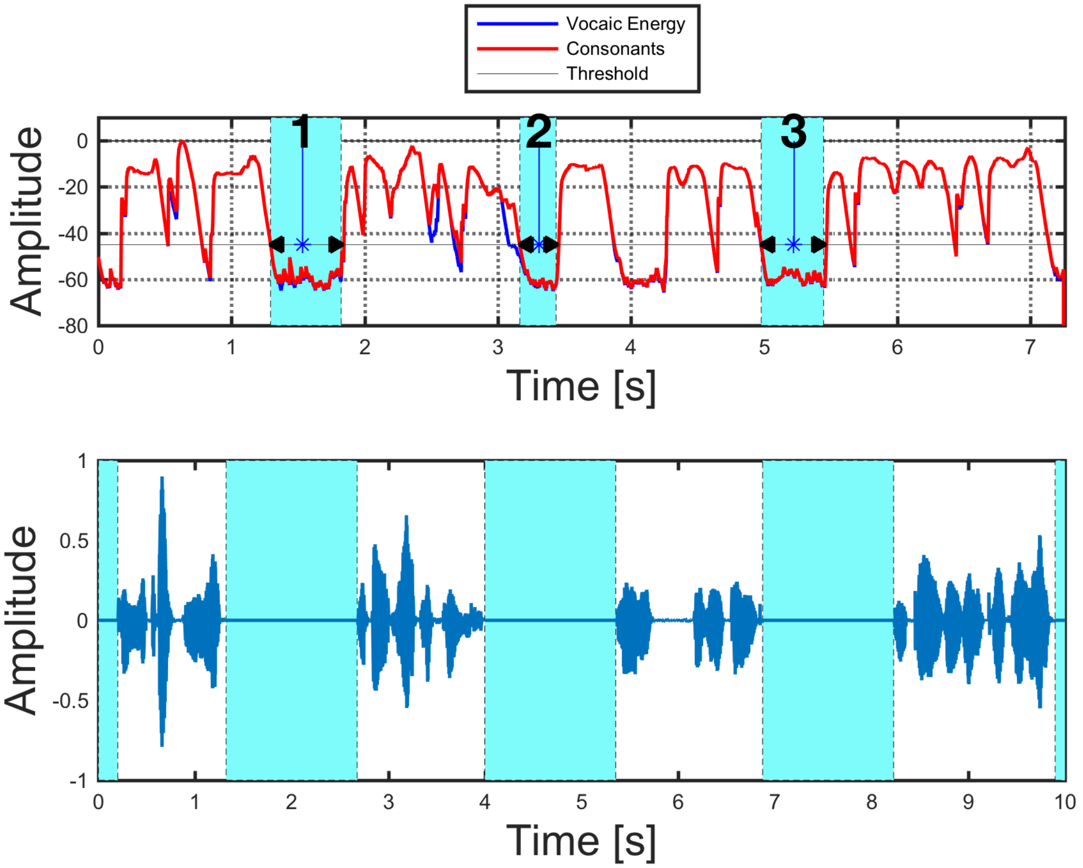
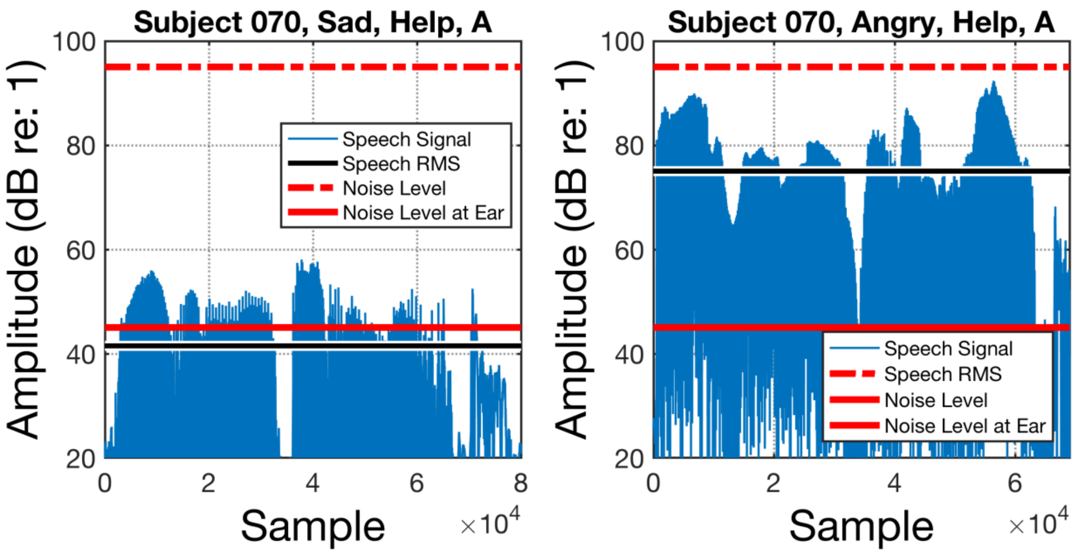


Figure 3.14 The upper plot illustrates a waveform where the user has identified a point within the three silent regions. The lower plot shows the output of the algorithm, which has added the appropriate amount of space between each utterance.

Normalization

As previously mentioned, the loudness of the environment surrounding the listener severely limited the usable dynamic range of any presented signal. The best-case scenario would be one in which each stimulus could be presented at amplitudes relative to the original sound pressure waves incident on the microphone element. This would require an effective dynamic range of almost 60 dB (30-90 dB), and a noise floor lower than 30 dB at the inner ear of the listener. To combat this noise, earmuffs (Restek citation) (30 dB nominal noise reduction) were worn over foam-padded in-ear headphones (Sensimetrics, 10-40 dB noise reduction). Previous work has shown that the effective noise reduction from the combined use of earmuffs and foam plugs ranges from 40 dB to 50 dB for experienced subjects depending on the fit between the plug and the ear canal [53]. Even the optimal 45 dB estimate is nominal, and does not factor in conducted vibration from the structure of the MRI to the ear, which is sure to diminish the noise reduction further.

Figure 3.15 illustrates the danger in calibrating each stimulus to its amplitude at the site of the microphone. The amplitudes of each waveform were calibrated using a custom calibration curve that relates the gain setting at the time of the measurement to RMS pressure of a piston phone (1 kHz, 1 Pa RMS). When both the sad and angry versions of the stimulus are scaled to their appropriate amplitudes, their relative magnitudes are immediately apparent as in Figure 3.15. Here, the louder angry version has been scaled to a loud but safe presentation level of 75 dBA. The waveform for the sad version has been scaled proportionately by the ratio of the relative levels between the two portrayals. Please note that the RMS level of the sadness portrayal hovers almost 5 dB below the at-ear noise level after the most optimal case of noise reduction. One could argue that proportionate scaling could ameliorate the disparity in level, but even then, the effects of frequency masking due to differences in the signal-to-noise-ratio (SNR) would still exist. See Appendix D for the calibration curve.



(a) (b)

Figure 3.15 A comparison of dynamic range of the two affects sad and angry: (a) a waveform of an utterance measured as the speaker expressed sadness quietly, (b). a waveform of the same encoder speaking the same script and utterance with an angry affect. Each have a 45 dB noise reference line.

Figure 3.16 illustrates some of the relative dynamic range between the utterances in an angry affect. From left to right, each utterance's RMS amplitude is -30.5, -23.0, -19.5, -24.8 dB. The greatest difference is the 11 dB between the first utterance and the third utterance. A 11 dB difference here indicates that the third utterance is more than three times (3.55) the RMS amplitude of the first utterance.

Perceptually, listeners would rate the third utterance as more than twice as loud as the first. So some amount of within-stimulus normalization across each utterance is needed to ensure that the noise from the scanner does not overpower utterances of lower amplitudes.

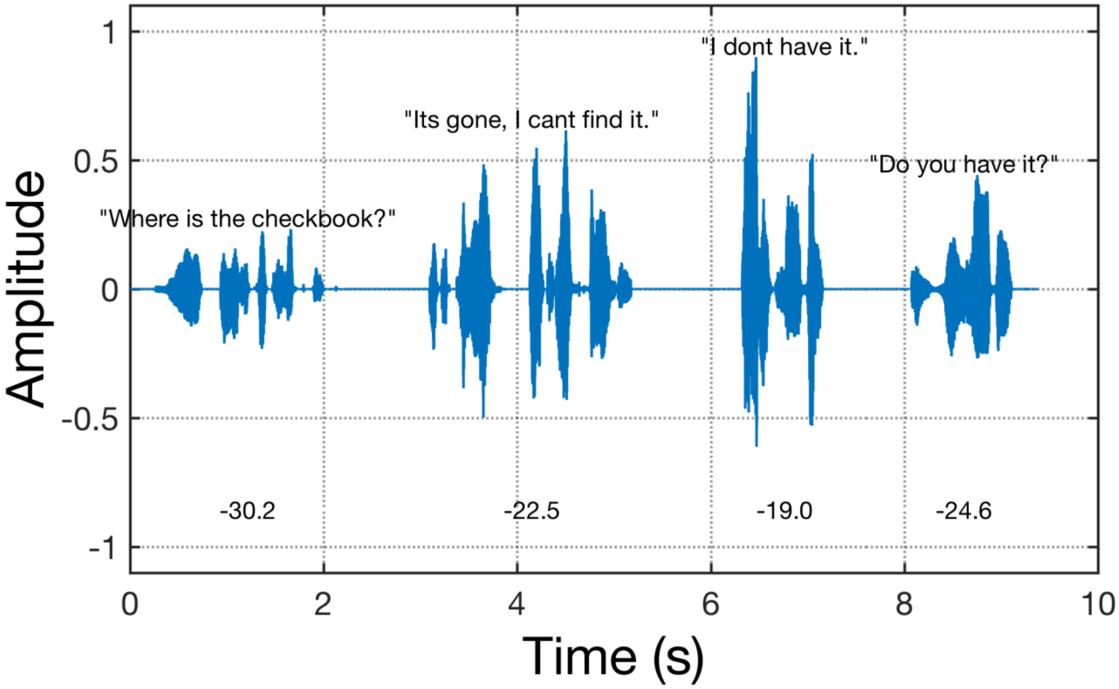


Figure 3.16 Time series of an unprocessed stimulus. The spoken text is given above the waveform and the corresponding A-weighted RMS amplitude of the utterance is given below the waveform.

Several methods of amplitude adjustment were considered, each with their own set of advantages and disadvantages. Figure 3.17 shows two methods of normalizing each utterance. The first method simply scaled each utterance so that their absolute maximum value was 1. This maximum value is necessary for playback over most sound cards, otherwise clipping will occur. The second method scales each utterance so that they all have the same A-weighted RMS amplitude as the utterance with the smallest RMS value. In the improbable case that a resulting peak amplitude is greater than one, each utterance is then multiplied by a constant scale factor to return this peak value to one.

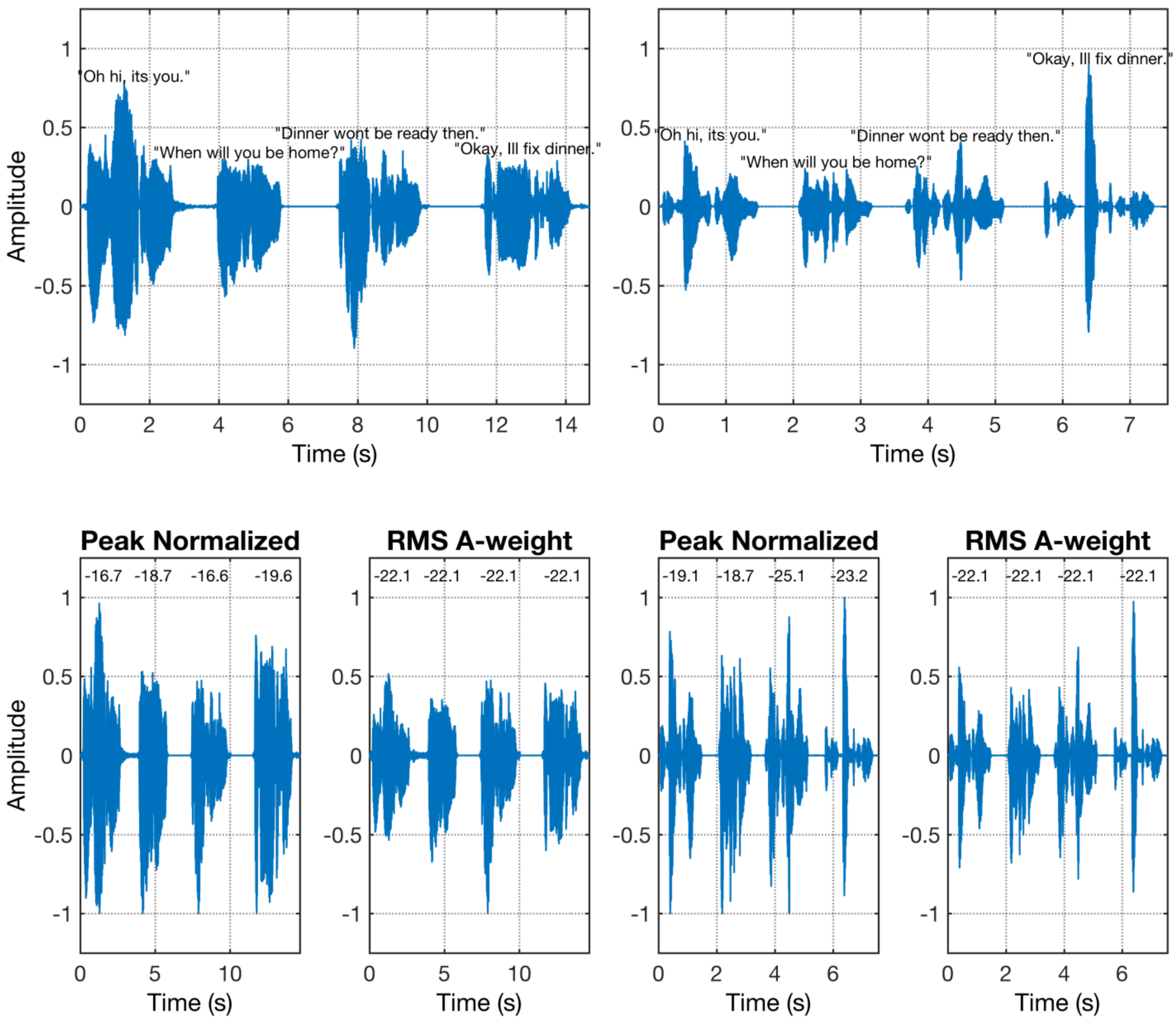


Figure 3.17 The upper two plots show the original waveform for the same script, "Dinner a", spoken in both the sad (left) and angry (right) affects, respectively. Along the bottom row (left to right) are the normalizing methods using absolute peak, RMS, and A-weighted RMS values per utterance. A-weighted RMS values (dB re: 1) are given above each utterance.

Of the two methods presented, normalizing each utterance by its A-weighted RMS amplitude more closely approximates the perception of equivalent loudness across utterance and affect. Regarding the peak normalizing method, the sad utterances were scaled up to 10 dB higher than the angry utterances because of their smaller range of amplitudes. The second method ensures that each utterance within each stimulus can be presented to the decoder at safe and equal loudness.

3.6 Results

Participant Information

Data collection was conducted over the course of 15.5 months from February 24, 2016 to July 10, 2017. Of the seventy-two (72) people that were recruited, fifty-five (55) completed all experimental tasks and produced a full set of stimuli. Forty-three (n=43) of this remaining group of participants granted permission for their voice recordings to be shared for public presentation and scientific use on the online repository Databrary [54]. This set of participants will be the primary focus for the presented work as future analyses on any stimulus depends on the status of each participant's permission to do so. The mean age of the participants as of the day their stimuli was created was 38 years and 9 months, and the standard deviation of their ages is 4 years and 9 months. All participants except one were fluent in English as a first language and were residents of State College, PA or the surrounding counties. Twenty-nine participants reported at least some amount of musical training, and eleven reported having had acting training as of the measurement date.

Summary of Corpus contents

In total, the corpus consists of 43 encoders' portrayals of four affective or non-affective states for eight scripts, each containing four individual phrases. The corpus itself has been archived as it was presented to the decoders in the PEEP study, i.e. each 10 second stimulus file contains all four normalized utterances spaced apart in time as previously discussed. Consequently, there are 1376 of the 10 second stimuli (344 per expressive style), or 5504 utterances (1376 per expressive style). The cumulative length of the corpus is 3 hours 49 minutes and 20 seconds, every sample of which was recorded at a 48 kHz sample rate and 24 bits per sample. The sample rate chosen permits frequency analysis up to 20 kHz and the bit depth provides a signal-to-quantization-noise ratio of 144.49 dB (20*log10(2^24)). The gain setting was recorded for each measurement in addition to qualitative notes regarding the subjective perception of the portrayal or environmental noises detected.

3.7 Discussion

Many valuable lessons were learned in the process of developing the procedure to obtain high quality recordings from the participants. Piloting the recording sessions was an essential step in refining the directions given to the participants. The initial responses from the pilot phase provided a frame of reference to base quality control for successive participation. The most essential pieces of information needed were the degree to which the vocal coach and sound engineer should work with the participant to extract the highest quality emotion portrayals.

The procedure faced many limitations that previous studies specific to vocal expression did not. For example, Banse and Scherer (1996) included only professional actors, and gave each actor an unlimited amount of time and number of attempts to achieve what they felt was their greatest portrayal. These actors also memorized their scripts in order to further naturalize their speech. These factors were likely to increase the accuracy of the vocalizations as they relate to the intended emotion of the participant due to the experience each subject has in acting apart from their true emotional state. None of these details were could have been practically implemented because of the population distribution that was sampled and the time constraints necessary to prevent exhaustion.

Of particular importance to the recording process for this study was the degree of specificity in defining the prompted emotions in both a psychological manner and acoustical manner. Anger was defined as a highly aroused negative emotion that more closely equates to rage, happiness was specified as a highly aroused and positive emotion, such as elation, and sadness was defined as an emotion of very low arousal and of negative valence [6]. Participants did not report having any confusion regarding these specific labels. Many, if not most, studies in vocal expression of emotion do not provide specific instructions on how to achieve the appropriate acoustical form for each emotion [27] [55], although some give direction or suggestions for fine-tuning pronunciation or "naturalness" [43]. The presented works had to walk a fine line between support for prototypical expression and noninvasiveness. Expressions were to be prototypical enough for the decoders (listeners) to at least recognize the affect, yet remain as ecologically valid as possible. Finding the balance of providing feedback and examples for participants to use and modify their own expressions required attention to individual acting abilities. Where there was a conflict between typical changes in pitch or emphasis compared to the natural movements of the participant's voice, priority was given to their natural inclination. In some cases, however, participants were quick to emulate examples rather than reflect their own way of producing an affect. Although great care was taken to ensure that their expressions reportedly felt natural to the participant, there is always the possibility that the measurement process interfered with the stimuli recorded.

4 Acoustic Analysis of the Encoding Process

4.1 Objectives

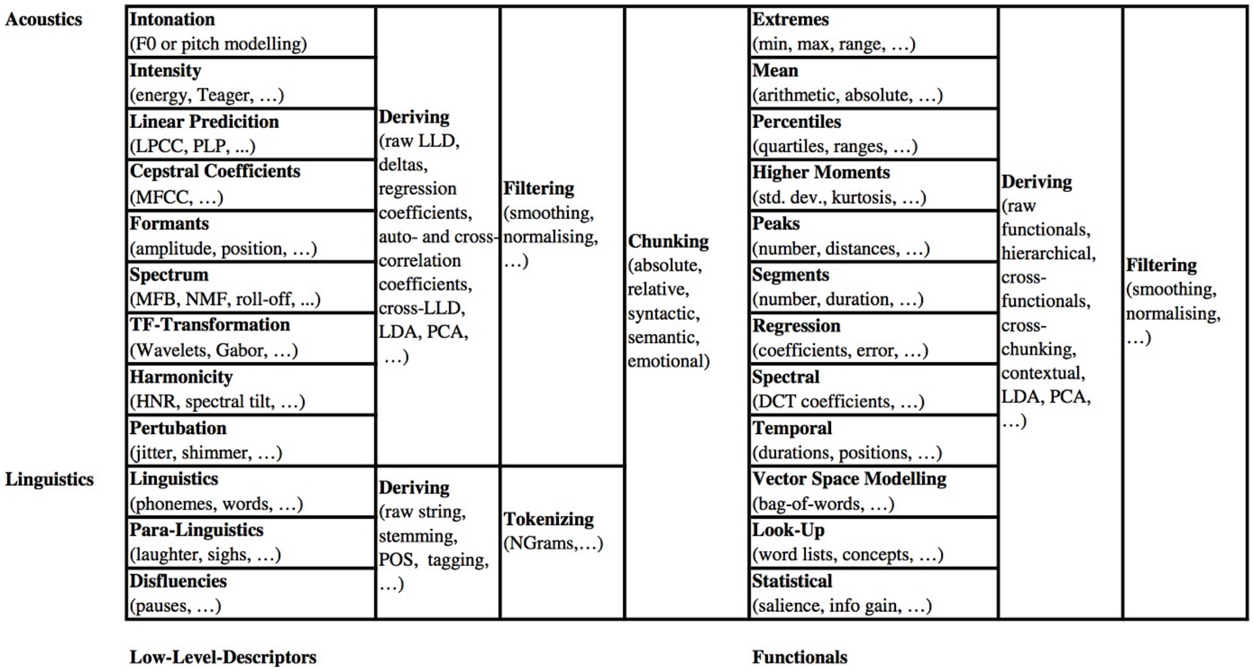
The primary focus of this chapter is to answer the first two questions posed in Chapter 2 regarding this new corpus of emotional speech. These are:

i. Does the presented speech corpus exhibit acoustic cues that agree with the literature?
ii. Do the acoustic cues known to carry emotional information correlate independently of semantic content?

There are, however, a multitude of steps to move from 1,760 audio recordings of 55 people to a clean statistical summary of their acoustical properties. So for each of the acoustic cues that will be considered, there are 1,760 opportunities for assumptions and oversight to elicit erroneous conclusions. A thorough and yet (hopefully) concise account of the methods and algorithms used to process the stimuli database will thus be given.

4.2 Analysis Pipeline

The presented analysis attempts to reduce a large and harmonically rich signal down to a collection of statistically representative values, each representing an acoustical perspective. The need for single value representation stems from categorical subjective validation; for each label or rating given by a listener, there must be one acoustic value to compare to. Each stimulus goes through a set of processes that can be broken down into a hierarchical order. Figure 4.1 illustrates the general categorization of both the feature type (rows) by process type (columns). Signal information moves through this table from left to right. The left column lists various common linguistic and acoustic Low-Level-Descriptors (LLDs). These include contours that represent *F0*, intensity, formants etc.



	Low-Level-Descriptors				Functionals		
Acoustics	**Intonation** (F0 or pitch modelling)	**Deriving** (raw LLD, deltas, regression coefficients, auto- and cross-correlation coefficients, cross-LLD, LDA, PCA, …)	**Filtering** (smoothing, normalising, …)	**Chunking** (absolute, relative, syntactic, semantic, emotional)	**Extremes** (min, max, range, …)	**Deriving** (raw functionals, hierarchical, cross-functionals, cross-chunking, contextual, LDA, PCA, …)	**Filtering** (smoothing, normalising, …)
	Intensity (energy, Teager, …)				**Mean** (arithmetic, absolute, …)		
	Linear Predicition (LPCC, PLP, ...)				**Percentiles** (quartiles, ranges, …)		
	Cepstral Coefficients (MFCC, …)				**Higher Moments** (std. dev., kurtosis, …)		
	Formants (amplitude, position, …)				**Peaks** (number, distances, …)		
	Spectrum (MFB, NMF, roll-off, ...)				**Segments** (number, duration, …)		
	TF-Transformation (Wavelets, Gabor, …)				**Regression** (coefficients, error, …)		
	Harmonicity (HNR, spectral tilt, …)				**Spectral** (DCT coefficients, …)		
	Pertubation (jitter, shimmer, …)				**Temporal** (durations, positions, …)		
Linguistics	**Linguistics** (phonemes, words, …)	**Deriving** (raw string, stemming, POS, tagging, …)	**Tokenizing** (NGrams,…)		**Vector Space Modelling** (bag-of-words, …)		
	Para-Linguistics (laughter, sighs, …)				**Look-Up** (word lists, concepts, …)		
	Disfluencies (pauses, …)				**Statistical** (salience, info gain, …)		

Figure 4.1 Analysis pipeline for processing emotional speech. Stimuli is processed into low-level-descriptors (e.g. F0 contours), upon which functionals can be subsequently applied (e.g. min, max, mean, range) (From Fig. 1 Schuller et al. 2011).

The relative wealth of work on traditionally accepted prosodic features such as pitch, duration, and intensity features shifted focus away from more modern acoustic features (e.g. Cepstrum, Linear Prediction) [21]. A reduced set of LLDs can then enter a second phase of processing whereby *functionals* are applied. *Functionals* consist of any operation that is subsequently applied to the raw LLD contours. These consist of normalizing, filtering, as well as simple statistical operations such as finding the mean or range of a given contour. The end result of these processes is a set of values for each feature (e.g. Mean of *F0*, Standard Deviation of Formant 1) that characterize a given stimulus. From here, these values can be compared to those found in the literature or subjective ratings from a listener.

4.3 Calculating the Low-Level Descriptors (LLDs)

4.3.1 Source considerations

As previously discussed in Chapter 2, the physics and natural tendencies of the source and the receiver necessitate certain spectro-temporal bounds on future analyses. The typical range of frequencies and sound pressure levels (SPL) of both speech and music is illustrated in Figure 4.2.

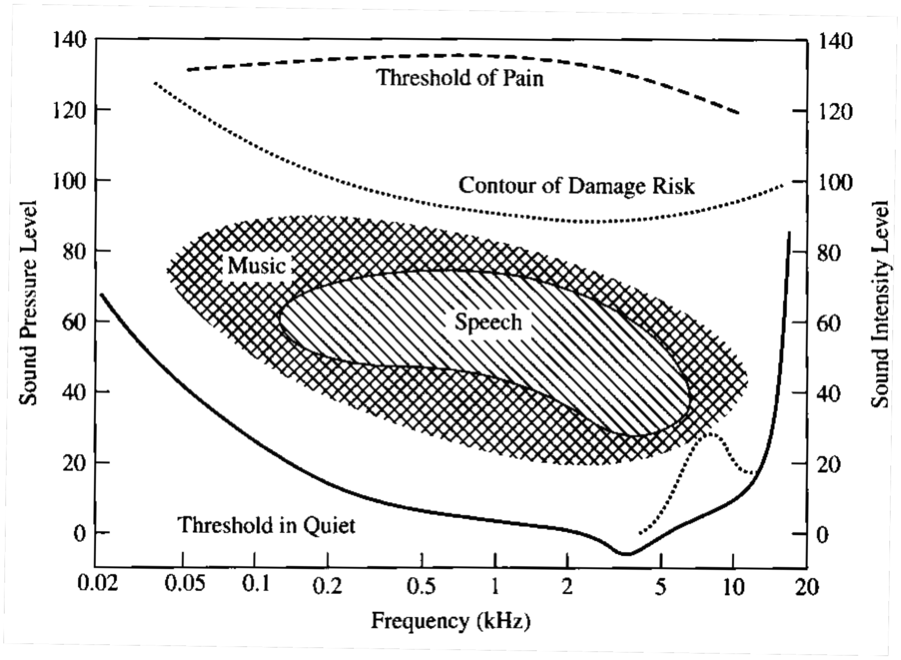


Figure 4.2 Frequency and amplitudes of music and speech. The bottom contour is the threshold of human hearing, and the top (dotted) line indicates where hearing damage can occur (From Fig. 4.13 Rabiner & Schafer 2011).

Of particular importance in this figure is the range of frequencies that speech typically inhabits, which is between 100 Hz and 7 kHz. The bottom solid line indicates the threshold of human hearing for the quietest sounds a person can hear. These values are typically 30-70 dB lower than amplitudes of typical speech. Loud, emotionally aroused speech may occur at levels slightly higher than normal (80 dB). This increase in loudness affects the relative sensitivity of the human ear to various frequencies as is illustrated in Figure 4.3. Tones at 125 Hz and 82 dB SPL would be heard at roughly 70 phons, which is approximately the same amplitude that an 8 kHz tone would be heard at 82 dB SPL.

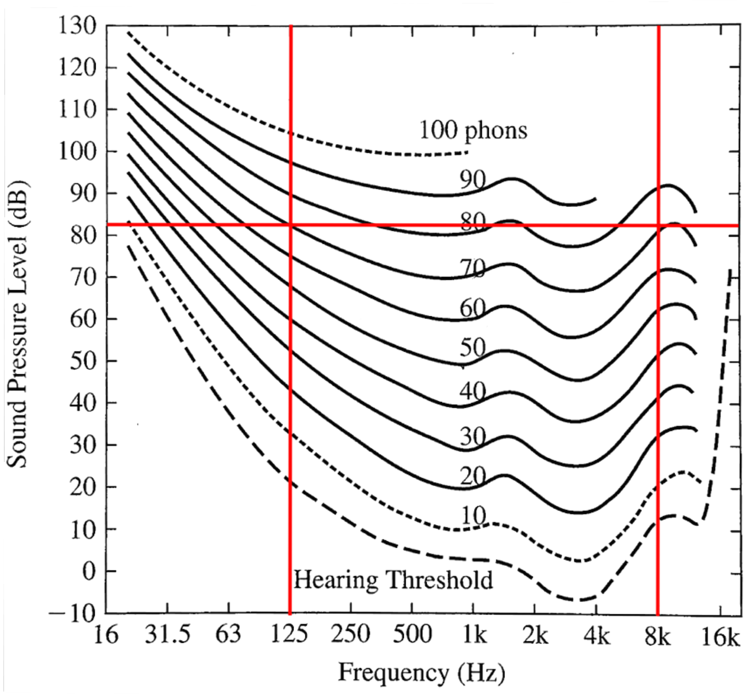


Figure 4.3 Equal loudness curves for pure tones (Adapted from Fig 4.14 Rabiner & Schafer 2011, data source: international standard ISO226).

To account for this expanded range of frequencies, a more conservative upper limit of 8 kHz is used [11]. By the Nyquist criterion, a sample rate of at least 16 kHz should be used for recording or playback of the stimuli. If there is ever a need to reference or compare a stimulus to a model of the vocal tract, there needs to be a certain time over which one can assume that both systems have a constant physical state. For the human vocal tract, this period of time ranges from 10 to 40 ms in length [11]. With these basic details about the voice in mind, a more refined and targeted acoustical analysis was performed.

The following sections have been organized in terms of commonly studied prosodic features. These include duration, intensity, pitch, and voice quality. The later sections will cover features born of modern advancements in DSP algorithms and available computational resources.

4.3.2 Duration

LLDs that fall into the duration category represent the most global and objectively simple characteristics of speech. These features also connect the linguistics of a speech episode to a variety of acoustic measures. As these features concern lengths of time or counting of units of speech, they tend to be processed at the speech episode and utterance level. Prosodic features are suprasegmental by definition; therefore, the units of analysis tend to characterize acoustic cues over the length of an utterance and not just a phoneme. Examples of these features include utterance duration and silence duration. A detailed explanation of how silent and non-silent regions of the stimuli were determined can be found in Chapter 3, section 3.5.2. The proportion of pauses (silences) in a speech episode can be calculated by:

$$P_{paus} = \frac{Duration\ of\ Pauses}{Duration\ of\ Utterance}\left(\frac{s}{s}\right) \qquad (4.1)$$

At the utterance level, the stimulus can be classified as voiced or unvoiced. This simply differentiates the signal into regions where the glottis is actively generating a pulsed excitation at the base of the vocal tract (voiced) or if it is otherwise motionless (unvoiced). This secondary classification affords other LLDs such as duration or number of voiced segments and the duration or number of unvoiced segments. Determining where voiced regions are present is a simple matter of calculating the pitch of the voice in a given frame of analysis, and comparing the strength of the peak correlation to a predetermined threshold. Details on this procedure are provided in subsequent sections. It is still possible to visualize much of what the pitch algorithm attempts to categorize through the use of spectrograms. Figure 4.4 illustrates how a person does not excite their vocal tract with their vocal cords when speaking /s/ or /h/. The regions of the spectrogram where both of these fricatives occur are absent of frequency content close to where *F0* could occur.

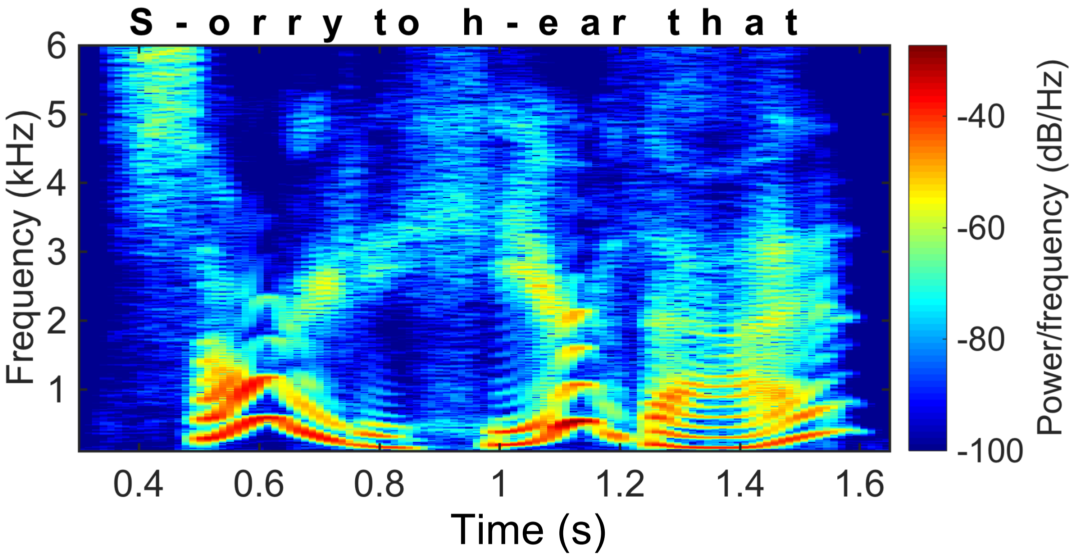


Figure 4.4 This is a spectrogram of subject 001 speaking "Sorry to hear that" in a happy affect. Both the "S" and "h" are aligned with broadband noise and have little low frequency contribution from the source.

The duration of the voiced regions in this case are roughly 1.08 s, and the duration of the unvoiced regions are 0.19 s. One can employ some basic knowledge of language and linguistics to extract more suprasegmental cues in this utterance.

Again referring to Figure 4.4, the number of words and the number of syllables within each word can be compared to the duration of the utterance. The utterance “Sorry to hear that,” consists of four spoken words. The word rate is given in (4.2).

$$Word\ Rate = \frac{\#\ of\ Words}{Duration\ of\ Utterance} = \frac{4}{1.27} = 3.16 \left(\frac{words}{s}\right) \qquad (\ 4.2\ )$$

The word “sorry” has two syllables, while “to”, “hear”, and “that” only have one syllable. Their total results in five syllables, and the syllable rate can then be calculated as in (4.3).

$$F_{syll} = \frac{\#\ of\ Syllables}{Duration\ of\ Utterance} = \frac{5}{1.27} = 3.95 \left(\frac{syllables}{s}\right) \qquad (\ 4.3\ )$$

Previous studies have found that speech rate increases in expressions of anger and happiness, and decreases for sadness.

Intensity

Intensity features reflect the energy and the perceived loudness of the acoustic signal. Although it is possible to calibrate these features so that an absolute pressure or intensity can be reported, relative values to microphone voltage are more often reported [21] [27] [55] .[v]

Measured with a microphone, the effective loudness of speech signal can be characterized on a long and short-term basis. Regarding the former case, the Root-Mean-Square energy of a signal is calculated as:

$$E_{rms} = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x^2(n)} \qquad (\ 4.4\ )$$

Where *N* is the total number of samples within the analysis frame, and *x* is the signal. If a value more representative of the perceptual energy is desired, the signal can be A-weighted prior to computing the RMS. The intensity calculations that were used in the acoustic analysis for comparisons between emotional expressions were performed at the utterance level and divided into both an original “raw” group and a normalized group. These groupings were made in order to differentiate between

[v] The normalization discussed in Chapter 3 still allows for within-utterance intensity analysis. Analysis of the signal intensity as encoded (not as presented) requires adjusting levels to their pre-gain adjusted level.

analysis of relative amplitudes as produced by each encoder at the time of the recording and the amplitudes of the stimuli as they were presented to the listeners (see Chapter 5 for details).

Regarding estimates for the original amplitudes, both the gain settings on the M-Audio M-Track sound card and a calibration curve were necessary for pre-scaling. As was described in Chapter 3, gain settings for each stimulus were documented as part of the recording process on a scale from 0-100 in increments of 5. As it was not known whether the tick marks on the gain dial were linearly related, a calibration curve was created from which the documented gain indices could later refer to. This calibration curve was created by sending a 1 V_{RMS} sinusoid at 1 kHz into the line input of the sound card, and recording the ratio between the input RMS amplitude and the RMS amplitude of the recorded sinusoid (see Appendix D). Three values for each tick mark were recorded and their average served as the reference for that index. A scaling factor for each stimulus was created by using finding the point on the calibration curve referenced by the recorded gain setting. Linear interpolation was used as necessary to determine values that were between tick marks.

To calculate the estimated proportional amplitude of a stimulus at the time of the recording, the signal was first divided by its corresponding scale factor and then A-weighted by using the following pre-computed filter coefficients and the MatLab *filter* function:

b = [1.184072 -2.368144 1.184072]
a = [1.000000 -1.893870 0.895160]

These values were generated for a sample rate of 48 kilosamples per second, and are representative of an approximation of the A-weighting curve at the following characteristic frequency values given by the (exact) ANSI standard S1.4-1983:

f_1 = 20.598997 Hz ; f_2 = 107.65265 Hz ; f_3 = 737.86223 Hz; f_4 = 12194.22 Hz

The intensity was then calculated as the sound pressure level in dB with reference to unity:

$$L_{AI,\mathrm{M}} = 10\log_{10}\left[\frac{1}{T}\int_0^T x_A^2(t)dt\right] \qquad (4.5)$$

Where x_A is the A-weighted time series with or without the scaling factor. The time-dependent intensity contour was calculated in similar fashion to the operation of sound-level meters. This was accomplished by exponentially averaging the square of the A-weighted time series with a time constant of 35 ms. This is the ANSI standard for the impulse setting on sound level meters, and is much shorter than the fast time constant setting of 125 ms. Exponential averaging was done using the MatLab *filter* function, and the coefficients B = [0 α] , A = [1 α-1] where α is given by:

$$\alpha = \frac{1}{f_s T_c} \qquad (\ 4.6\ )$$

Where T_c is the time constant for sound level meters at the impulse setting (35 ms). The intensity contour was then calculated at every point along the time series as:

$$L_{AI}(s) = 10 \log_{10}\left[\frac{\langle x_A^2 \rangle_I}{1}\right] \qquad (\ 4.7\ )$$

Where the subscript *I* indicates that the standard ANSI time constant for impulsive sounds was used. The emulation of the sound level meter at a considerably faster time constant than the fast setting was done in order to maximize the compatibility of the results with other studies that have used a similar set process. The speech processing software PRAAT is frequently used to perform short-term calculations that make up an intensity contour [56]. PRAAT's calculation for the intensity contour starts by convolving the squared time series with a Gaussian window that is 32 ms in length (as a default). PRAAT then computes the intensity in dB with reference to 20 µPa, assuming that the time series *has already been converted to Pascals.* Using this calculation method would have artificially inflated the values reported. It was determined that the best course of action was to use a standardized ANSI sound level estimation method with a standard time constant that was similar to that of many other studies in speech processing. It should be noted that when mean values were specifically of interest, Equation (4.5) was used, while values for the standard deviation used Equation (4.7).

4.3.3 Pitch (F0)

F0, also known as the pitch or fundamental frequency of speech is one of the most widely studied prosodic feature in speech communication [21]. *F0* is useful for both linguistic and affective prosody. Questions in American English are often posed with an upward inflection towards the end of the sentence. In fact, sentences need not be complete or grammatically correct for one to successfully communicate a simple question if the *F0* contour follows the traditional upward trend (e.g. *"dinner?"*). Additionally, people often read off a list of items by starting at a lower frequency and ending at a higher frequency for each item. Although the emotional content is of more interest to the presented work, it is worth being aware of *F0*'s linguistic functions as they may interfere with future contour-dependent analyses.

This section defines *F0* in the context of the voice production system. Such an approach is intended to clarify the intent, strengths, and weaknesses of *F0* estimation algorithms operating on a physical source.

Recall the glottis from the physical model of the speech production from Section 2.2.2. *F0* is defined as the rate of the opening and closing of the glottis and is given in Hertz (Hz). The nature of the excitation source along with the spectral modifications inherent caused by the vocal tract are contrasted between Figure *4.5* (a - c). The motion of the glottis and its spectral contributions are best represented as a train of impulses, not a pure sinusoid. Frequency-domain pitch estimation of a pure sinusoid offers little computational burden (Figure *4.5* (a)), while tracing zero-crossings or peak spacing in the time domain of (Figure *4.5* (b)) and (Figure *4.5* (a)) renders equal ease for the task at hand. A quintessential speech processing problem arises when vocal tract resonances amplify harmonics above the fundamental in Figure *4.5* (c). Such obstacles have been hurdled in a variety of ways, and each method carries with it advantages and disadvantages.

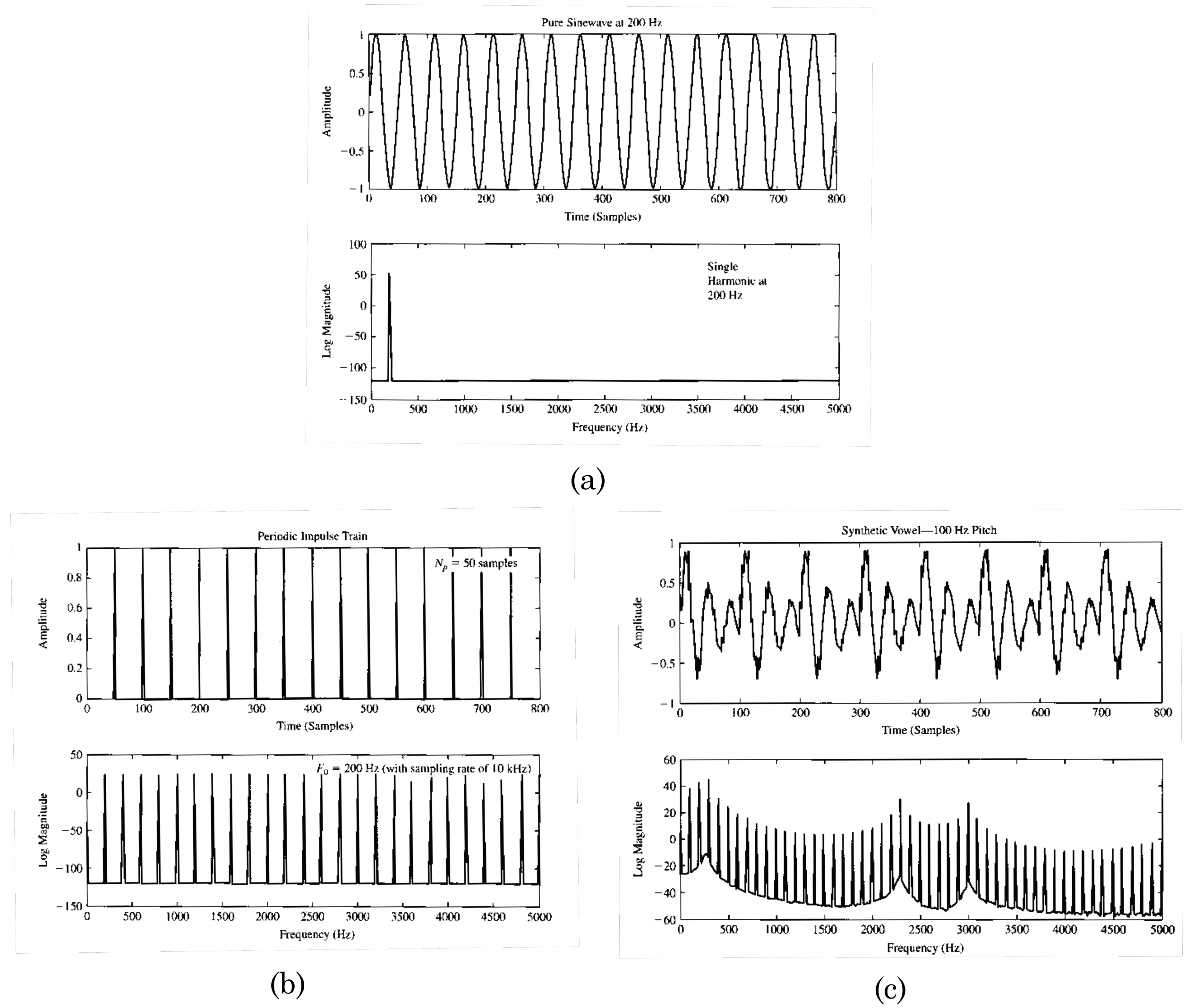


Figure 4.5 (a) a sinusoid and its corresponding magnitude spectrum. (b) an impulse train and its magnitude spectrum. (c) a synthetic vowel at 200 Hz and its corresponding spectrum (From Fig. 10.16, 10.17, 10.18 Rabiner & Schafer 2011).

The first attempts at pitch extraction battled valiantly against computational limitations of the mid-twentieth century. Time-domain methods of categorizing speech waveform amplitudes produced a dense forest of logic trees, the fruits of which rarely outperformed peak extraction by the naked eye (Gold, 1962). Auto-correlation by way of the short-time Fourier Transform (STFT) redefined the search for ***F0*** in terms of time delays. While more computationally taxing, this method advanced the effort into a form currently implemented in modern digital hardware [11]. The following section will present a method that builds upon the auto-correlation procedure for estimating ***F0***.

The autocorrelation function gives a measure of the periodicity of a signal. In the time domain, the auto-correlation function works by shifting a signal down the length of an identical copy, and integrating their product with respect to time. Given a signal x, the autocorrelation function is defined by (4.8).

$$R_{xx}(\tau) = \frac{1}{T}\int_0^T x(t)x(\tau + t)dt \qquad (4.8)$$

A powerful property of this function, as noted by Rabiner is that pitch estimation by way of peak detection is independent of the "time origin of the periodic signal." Figure 4.6 illustrates a perfect sinusoid (a) and it's associated autocorrelation (b)

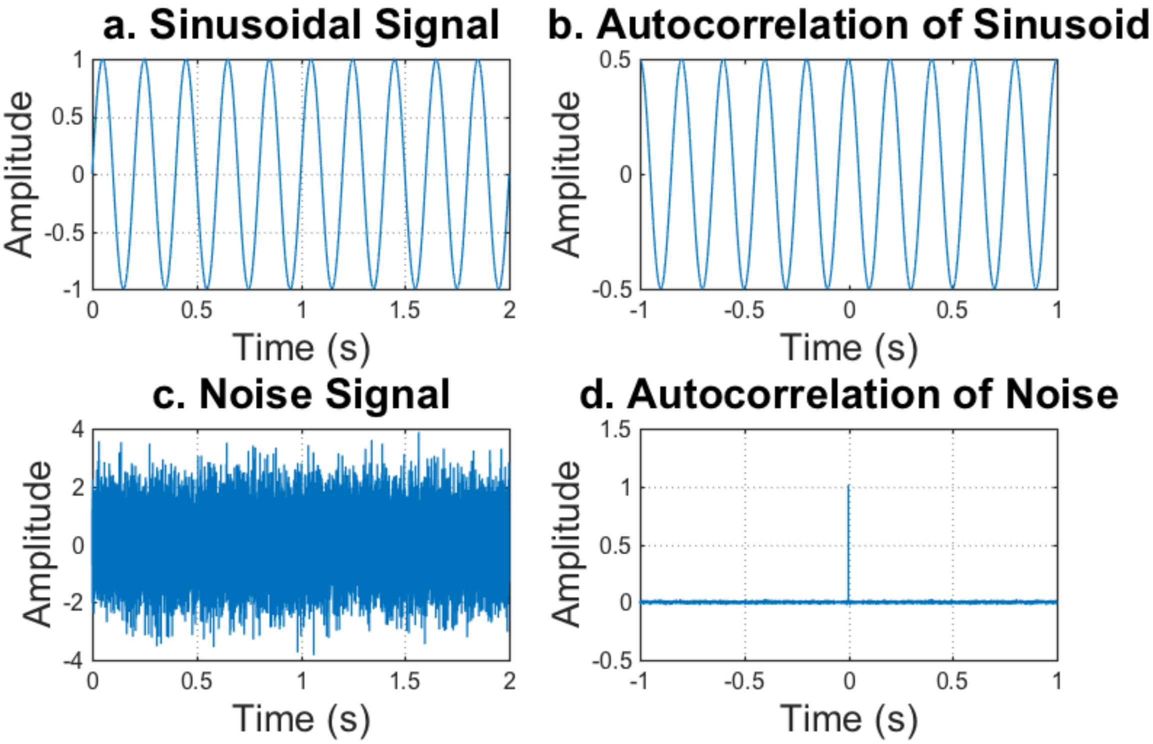


Figure 4.6 These four plots represent a. sinusoidal signal b. the autocorrelation of the sinusoid c. a signal of random noise and d. the autocorrelation of the random noise signal.

Qualitatively, a sinusoid will have a sinusoidal autocorrelation because of periodic alignment as depicted in Figure 4.6 (a). The peaks of the auto-correlation occur at time shifts equal to the period of the sinusoid (b). Conversely, signals comprised of random noise as shown in Figure 4.6 (c) result in a near delta function at zero time shift and much smaller peaks with increasing time shift (d). For a signal with multiple sinusoidal components, the peaks located at larger time shifts correspond to the lower frequency components, while the peaks at shorter time shifts correspond to the higher frequency components.

As mentioned previously, the vocal tract colors the otherwise flat spectrum of an impulse train. The autocorrelation function is affected analogously: peaks in the autocorrelation corresponding to higher harmonics (shorter lag times) are often amplified above the peak at the lag time of the fundamental frequency. Rabiner proposed first low-pass filtering the speech signal with a cut-off frequency of 900 Hz, then center-clipping the resulting waveform at a variable threshold so that just the tips of the largest peaks are included in the analysis. While effective in many cases, this algorithm relies heavily on speaker conformity, especially when considering expressive speech. Several phonetically described formants fall below 900 Hz for both males and females, while expressive vocalized squeals push beyond 900 Hz. Regardless of the situational shortfalls, this algorithm removes potentially useful information within the autocorrelation such as the relative energy of the periodic components. Boersma (1993) addresses these very issues with a surprisingly simple modification.

Considered by many to be the standard reference for pitch estimation, PRAAT software provides a multitude of speech processing tools with explanations for how each object operates [56]. Both Boersma and Rabiner use autocorrelation as the primary tool for estimating the pitch, but where they differ is in their respective solutions to "spectral flattening," or more specifically emphasizing lower frequencies over the formants as previously discussed. The proposed algorithm aims to accurately estimate both the excitation frequency of the glottis and the relative amplitude of this periodic component with respect to the rest of the signal [57]. Thus, filtering and center-clipping methods were not considered. Instead, Boersma proposes dividing the windowed autocorrelation of the voiced segment of speech $r_a(t)$ by the autocorrelation of the window itself $r_w(t)$. This process boosts the amplitudes of peaks in the autocorrelation at longer lag times proportionately to the taper of the window.

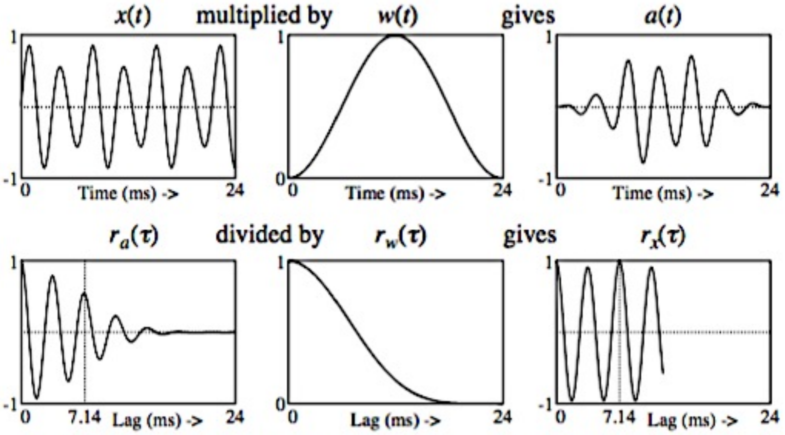


Figure 4.7 This is a graphical representation of the process of weighting the autocorrelation function. Source: Boersma (1993).

Peaks above a threshold correlation value and within a range of time lags are preserved, and the process is repeated to the end of the speech signal. Figure 4.8 gives an example matrix of pitch estimations through which a path finder can be applied to determine a physically reasonable pitch contour. Here, lower frequencies are preferred while octave jumps beyond physical normality are ignored.

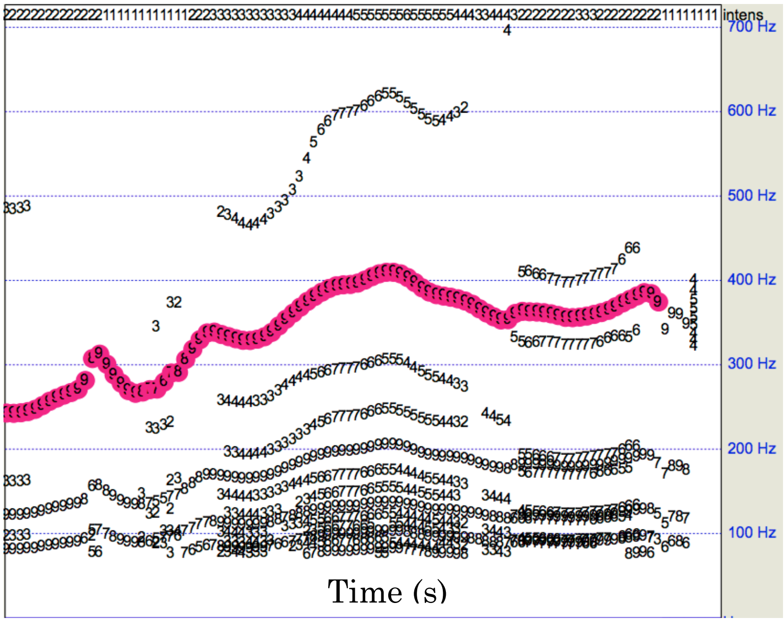


Figure 4.8 This is a screenshot of all of the candidates for the eventual F0 contour output by Praat. Praat's algorithm determines the best possible path based off of the magnitude of the numbers along the contour and jump distances.

The PRAAT "To Pitch (ac)…" was provided the following parameters for the autocorrelation method algorithm: a time step between analysis frames of 2 ms, a minimum possible *F0* value of 75 Hz to extend expected range slightly lower than most female voices, a Gaussian analysis window of 6 times the maximum possible pitch period, a *F0* ceiling of 900 Hz, the maximum number of *F0* candidates to 15, a silence threshold of 0.01, a voicing threshold of 60% of the maximum possible autocorrelation, an octave cost of 0.01, an octave-jump cost of 0.35, and degree of disfavoring voice/unvoiced transitions of 14% relative to the maximum autocorrelation. Additionally, the algorithm was set to remove octave jumps that were likely made of error, and to not conduct any smoothing along the resulting *F0* contour.

4.3.4 Voice quality

The voice quality category includes features that are less frequently associated with the classic prosodic cues of duration, intensity, and *F0* [58]. As such, the conceptual categorization of specific LLDs within this category have also changed over time. For example, short-term deviations in *F0* and intensity have been grouped separately from voice quality (and with *F0* and intensity) [3] [59], while others have left these cues as their own category altogether [21] [58] [60]. The long-term average spectrum has often received its own category [3] [21] [27] apart from voice quality as well. For the sake of simplicity, every acoustic cue that does not fall strictly into the category of duration, intensity, and *F0* as they have classically been organized shall be grouped as a voice quality cue [61].

Broadly speaking, acoustic cues that fall into the voice quality category reflect the differences in the style of excitation and amplification of the voice production system [60]. The proposed categorization of long and short-term spectrotemporal quantities has created a vast range of objective perspectives on the voice which will be further subdivided in the following text. Short term deviations in *F0* and intensity like jitter and shimmer will also be grouped with harmonics-to-noise ratio.

Long-Term-Average Spectrum

Features that belong to this subcategory illustrate various characteristics of the Long-Term-Average Spectum (LTAS). Computed over the entire utterance, the Discrete Fourier Tranform (DFT) is defined as [11]:

$$X(m) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi mn}{N}} \qquad (4.9)$$

where $x(n)$ is real, N is the length of the signal in samples, $n = t/T$ (current time/total time) is the sample index, and the frequency bin index m is an integer multiple of the frequency resolution $(df = fs/N)$ and is defined from 0 to $N/2$ +1.

The Alpha ratio provides a dimensionless measure for the proportion of energy above and below a specified cutoff frequency. This quantity is defined as the ratio of the summed energy from 50 Hz -1000 Hz to the summed energy between 1000 Hz - 5000 Hz [62]. This quantity is calculated as follows:

$$S_\alpha = \frac{\sum_{m_{50}}^{m_{1k}} X(m)}{\sum_{m_{1k}+1}^{m_{5k}} X(m)} \qquad (4.10)$$

where m_i is the frequency bin that corresponds to the range of frequencies provided in the summations.

The Hammarberg index is another dimensionless ratio of spectral energy in the LTAS. Instead of summing the total energy above and below a cutoff frequency, the Hammarberg index is computed by dividing the point of maximum amplitude between 0-2 kHz by the point of maximum amplitude between 2 – 5 kHz [33]:

$$S_{hamm} = \frac{\max\left\{X\left(m|_{m=1}^{m_{2k}}\right)\right\}}{\max\left\{X\left(m|_{m_{2k}+1}^{m_{5k}}\right)\right\}} \qquad (4.11)$$

Similar to the alpha ratio, the ratio of high frequency to low frequency energy has also been examined [55] [59] [63]. Juslin et al. (2001) defines S_{hf500} as the ratio of the total energy of the spectrum above 500 Hz to the total energy below 500 Hz:

$$S_{hf500} = \frac{\sum_{m_{500}+1}^{m_M} X(m)}{\sum_{m=1}^{m_{500}} X(m)} \qquad (4.12)$$

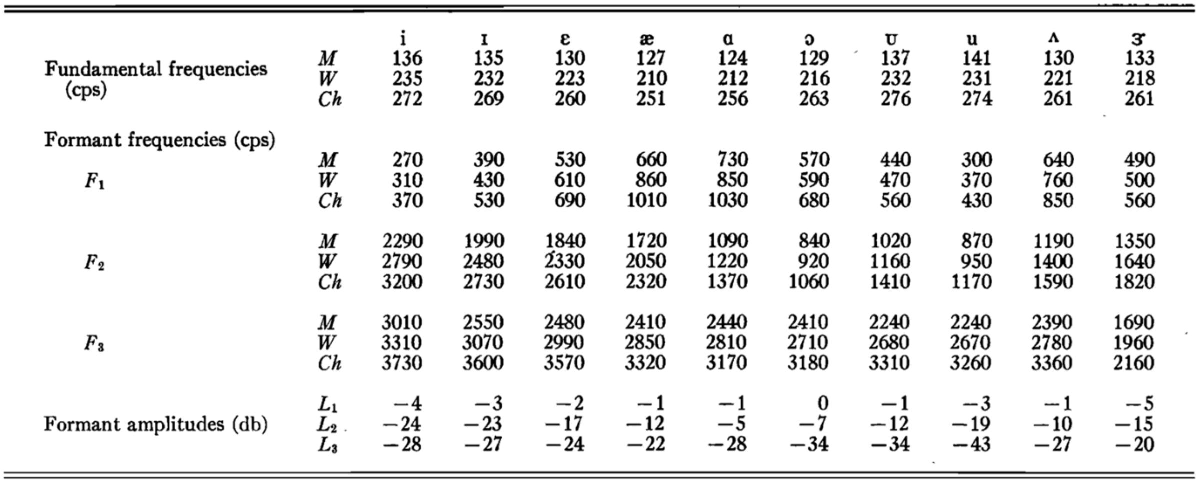
where M is the highest frequency bin in which $X(m)$ is defined. Previous studies have used similar metric definitions except with the numerator and denominator reversed [27] [61]. Correcting for this is a simple matter of inverting the result to match the desired definition. It should be noted that many of the aforementioned parameters such as alpha and S_{hf500} are proportional to the slope of the spectrum if the frequency range is included. Officially, the Spectral Slope is defined as the slope of the linear regression line of the LTAS between 1-5 kHz [64].

The center of gravity of the spectrum treats the spectrum like a distributed mass. If the frequency vector of a spectrum is treated like a measure of distance, the amplitude of the spectrum then becomes the relative quantity of mass at each frequency bin. The frequency at which the spectrum is balanced is calculated as:

$$S_{cog} = \frac{\sum_{m=m_l}^{m_u} F(m)X(m)}{\sum_{m=m_l}^{m_u} X(m)} \qquad (4.13)$$

Formants

Of the LLDs addressed up to this point, few are as directly representative of the physical state of the vocal tract as the formants. Formants are the resonance frequencies of the vocal tract tube [11]. Changes in the length, wall conditions, and area function of vocal tract modify the positions, amplitudes, and bandwidths of the formants. As previously stated, this coloration of the speech spectra enables acoustical differentiation between types of vowels. The relationship between vowel type has been well documented [65], and a small sample of the findings is provided in Figure 4.9.



		i	ɪ	ε	æ	ɑ	ɔ	ʊ	u	ʌ	ɝ
Fundamental frequencies (cps)	*M*	136	135	130	127	124	129	137	141	130	133
	W	235	232	223	210	212	216	232	231	221	218
	Ch	272	269	260	251	256	263	276	274	261	261
Formant frequencies (cps)											
F_1	*M*	270	390	530	660	730	570	440	300	640	490
	W	310	430	610	860	850	590	470	370	760	500
	Ch	370	530	690	1010	1030	680	560	430	850	560
F_2	*M*	2290	1990	1840	1720	1090	840	1020	870	1190	1350
	W	2790	2480	2330	2050	1220	920	1160	950	1400	1640
	Ch	3200	2730	2610	2320	1370	1060	1410	1170	1590	1820
F_3	*M*	3010	2550	2480	2410	2440	2410	2240	2240	2390	1690
	W	3310	3070	2990	2850	2810	2710	2680	2670	2780	1960
	Ch	3730	3600	3570	3320	3170	3180	3310	3260	3360	2160
Formant amplitudes (db)	L_1	−4	−3	−2	−1	−1	0	−1	−3	−1	−5
	L_2	−24	−23	−17	−12	−5	−7	−12	−19	−10	−15
	L_3	−28	−27	−24	−22	−28	−34	−34	−43	−27	−20

Figure 4.9 Vowels of American English and their corresponding formant frequency and amplitude. Source: Peterson and Barney (1952)

Consider for a moment the values for F_1 between IH and EH. For women, a small change in the pronunciation of IH to a more open EH results in an average delta of 200 Hz. The same can be seen for F_1 between AA and AO (850 Hz to 590 Hz). So although vowels traditionally prescribe a relatively consistent harmonic structure in spoken vowels, small deviations from normal pronunciation present a viable pathway towards emotion discrimination. In similar fashion, results from previous studies suggest that formants are in fact modulated by emotion [27] [55] [66] [67].

The task of calculating the formant frequencies and amplitudes is still quite laborious and prone to error. Rather than steer the text of this written work into the weeds of linear predictive coding (LPC), details on the Burg algorithm can be found from the reference section [68]. The parameters input to the formant estimating function include: 6500 Hz for the maximum possible frequency of the highest formant, the maximum number of formants equal to five, a time step between analysis frames of 5 ms, the analysis window equal to 25 ms, and finally the pre-emphasis frequency set to 50 Hz. The pre-emphasis parameter indicates the cutoff frequency of the high-pass filter applied to the signal, which is intended to correct for low frequency weighting that occurs during the processing.

Short Term Variability

Acoustic cues related to short-term variability include Jitter, Shimmer, and the Harmonics-to-Noise Ratio (HNR). All three of these parameters are offered by PRAAT [56]. PRAAT offers several methods for calculating jitter, ranging from absolute estimations (local, absolute) to five-point period perturbation quotients (ppq5). The default calculation is the relative local method "(local)" and is defined as the average absolute difference between consecutive periods, divided by the average period. The local relative method is the version used here. As this parameter compares *F0* values to one-another, the *F0* contour must first be calculated (see previous section on *F0* for details). Once the *F0* contour is known, the absolute jitter values are first calculated by:

$$F0_{jitt}(s) = jitter(s) = \sum_{i=1}^{N} \left| \frac{T_i - T_{i-1}}{N-1} \right| \qquad (\ 4.14\ )$$

where T is the inverse of *F0* or the pitch period, and T_i is duration of the i^{th} period, measured in seconds. Because *F0* values across an utterance may be missing due to noise or weakness in the estimation, only pitch period values within a physically reasonable range are considered. The minimum or period floor has been set to 0.1 ms or 10kHz, and the maximum period or period ceiling has been set to 10 ms or 100 Hz. For male speakers, this period ceiling should likely be increased to accommodate lower voices. Next, the mean period is calculated as:

$$\bar{T}(s) = \sum_{i}^{N} \frac{T_i}{N} \qquad (\ 4.15\ )$$

Finally, the relative jitter is given by:

$$F0_{jitt,\mathrm{M}} = jitter = \frac{jitter(s)}{\bar{T}(s)} \qquad (\ 4.16\ )$$

where *jitter* ranges between 0% and 200%. The Multi-Dimensional Voice Program (MDVP) reports that values of 1.04% from this method are a threshold for pathology [69]. Shimmer also measures short-term deviations of a speech signal. Instead of temporal deviations of the provided period, shimmer determines the relative or absolute deviations in amplitude. The chosen method for calculating shimmer is the relative local option “(local)”. The equations for calculating shimmer are exactly the same as (4.8)(4.14)(4.15), but with absolute amplitude in place of the period *T*. Values for shimmer are given in percent, and the MDVP uses 3.810% [69].

As discussed in section 1.2., the speech production system works to both filter the spectrum of the impulse train and contribute broadband noise into the speech signal. As a result, speech contains a combination of periodic components and broadband noise. Anger often moves the voice to produce stronger fricatives by tightening up the constrictions in the vocal tract that create vortex shedding or by increasing Reynold’s number.

The noise present in speech is rarely ever constant. Consonants located throughout an utterance produce varying degrees of broadband noise depending on the state of voicing, placement in a word, and even state of physical stress. Bitter conversations uttered through gritted teeth muffle the periodic contributions from the fundamental and formant frequencies while tightened constrictions at the oral cavity produce additional noise.

Boersma’s algorithm for the Harmonics-To-Noise Ratio (HNR) refines the accuracy in determining the strength of a signal’s periodic content relative to the power of the noise component. The basis for the analysis is the normalized autocorrelation in (4.17).

$$r_x'(\tau) = \frac{r_x(\tau)}{r_x(0)} \tag{4.17}$$

Where the numerator is the autocorrelation of the signal and the denominator is the value of the autocorrelation at zero time-lag. By definition, the value in the denominator is the mean square of the signal, and it represents the power contributions from both the noise N(t) and periodic components H(t). The power of the strongest periodic component of the signal is defined in Equation (4.18). This value is taken at the time-lag equal to one period and then normalized by the value of the autocorrelation at zero time-lag.

$$r_x'(\tau_{max}) = \frac{r_x(\tau_{max})}{r_x(0)} = \frac{r_H(0)}{r_x(0)} \tag{4.18}$$

Normalization has limited the range of possible values from 0 to 1. With this in mind, subtracting the normalized periodic peak value from one results in a measure of the noise power in the signal (4.19).

$$HNR = 1 - r_x'(\tau_{max}) = \frac{r_N(0)}{r_x(0)} \tag{4.19}$$

Although this equation neglects the contributions of other periodic components in the signal, the lack of filtering and short-term viability of the analysis provides a repeatable and speaker-independent metric.

Modern Advancements

The LLDs covered so far reflect but a small portion of the advancements made in both digital signal processing and computational statistics. A combination of linear predictive coding and cepstral analysis has boosted the accuracy of digital representations of speech to the point where telephone companies can fully reconstruct speech signals from a small set of features [11]. The mel frequency cepstral coefficients (MFCC) and their first derivative provide enough information to conduct system identification and perceptually meaningful and smooth reconstruction of the original signal. Although robust to noise and certainly useful for automatic speech recognition, application of these features requires exceptionally advanced time-alignment and statistical modeling to be useful beyond signal reconstruction. Recent work by Zhou et al. (2001) found that MFCCs perform poorly in emotion classification.

The trend of more modern studies in automatic emotion classification is the use of machine learning techniques such as artificial neural networks (ANN), linear discriminant analysis (LDA), and hidden Markov models (HMM) on acoustic (and sometimes semantic) data sets [34] [44] [58].

4.4 Results

The manner of recording the stimuli limited the length of the possible utterances to a cumulative sum of 10 seconds per measured stimulus. This limitation also affects the possible range and variance of utterance lengths as observed from the script-level. Within each script, however, words with varying numbers of syllables, lists, questions, and more could potentially influence various duration LLDs. This compelled the LLD extraction to be conducted on the utterance level for all of the stimuli. This produced a total of 7040 values for each LLD (55 encoders X 4 affects X 8 scripts X 4 utterances). Table 4.1 gives a statistical summary of each LLD with a functional typically used to describe a salient emotional cue.

The average utterance length (T_{utt}) was 1.29 seconds with a standard deviation of less than 30% of the mean. The average duration of the voiced segments was 0.82 seconds ($\sigma = 0.26$), which makes up about 64% of the total utterances. This indicates that the vast majority of the utterances are comprised of speech where the glottis is actively exciting the vocal tract. The closest LLD to the description of speech rate is the number of syllables per second (F_{syll}), which had a mean value of 4 Syl./s (σ = 1.06). It should be noted that syllabic rate and script may correlate as this quantity was calculated by dividing the integer number of phonetic units per utterance by the length of the utterance.

Although the word intensity has been used to describe an acoustic phenomena in previous works, here the term sound level ($L_{AI,\mathrm{M}}$) will be used in order to avoid confusion in future analyses of the subjective intensities. The subscript "A" indicates that the digital signals were A-weighted prior to integration. Strictly speaking, the functionals denoted by "M" or "SD" in the subscript for $L_{AI,\mathrm{M}}$ do not apply to the same contour of values. $L_{AI,\mathrm{M}}$ was calculated by integrating the A-weighted, gain adjusted digital signal, and dividing by the duration of the signal in seconds. $L_{AI,\mathrm{R}}$ and $L_{AI,\mathrm{SD}}$ took the range and standard deviation respectively of an exponentially averaged sound level contour. This calculation of this contour follows the standards for sound level meters set to an impulse time constant of 35 ms and A-weighted spectra. The signals fed to this process were adjusted to match the relative amplitude before the gain was applied to the signal entering the sound card. So although the absolute values of the $L_{AI,\mathrm{M}}$ are not an accurate representation of the sound pressure level of the encoder, the relative differences are still viable. All sound level standard deviations were within roughly 35% of their respective means.

F0 LLDs exhibited slightly more variance than the duration or intensity cues in Table 4.1. $F0_{\mathrm{R}}$ (μ =173 Hz) and $F0_{\mathrm{SD}}$ (μ = 46 Hz) had standard deviations of approximately 80% of their respective means. This suggests that more information may be present in these features than others.

The alpha ratio S_{a} (μ = 6.45 dB, σ = 6.06 dB) had variance near 90% of its mean, as did the ratio of high frequency energy above 500 Hz to the energy below 500 Hz (S_{hf500}). The generally large amount of variance in the spectral ratios necessitate closer examination.

Regarding the formant-based LLDs, $F1_{\mathrm{M}}$ (μ = 513.08 Hz, σ = 74.9 Hz), $F1_{\mathrm{SD}}$ (μ = 169.52 Hz, σ = 33.36 Hz), and $F1_{bw,\mathrm{M}}$ (μ = 172.5 Hz, σ = 58.21 Hz) all fell within typical ranges for female speakers. The values for the second formant $F2_{\mathrm{M}}$ (μ = 2012.16 Hz, σ = 176.89 Hz), $F2_{\mathrm{SD}}$ (μ = 499.99 Hz, σ = 83.3 Hz), and $F2_{bw,\mathrm{M}}$ (μ = 307.64 Hz, σ = 47.89 Hz) have lower variance than expected. The standard deviation of $F2_{\mathrm{M}}$ across the stimuli is only 9% of the mean. This may be indicative of excessive outlier reduction and a loss of information due to the extraction methods. $F1_{prec,\mathrm{M}}$ (μ = 1092%, σ = 925%), the absolute percent difference of the first formant frequency's deviation from the neutral affect counterpart exhibited variance across the corpus on the order of 85%. $F2_{prec,\mathrm{M}}$ (μ = 484%, σ = 430%) followed similarly.

Short-term voice quality features like the harmonics-to-noise ratio HNR_{M} (μ = 7.49 dB, σ = 3.29 dB) indicate that the stimuli was in fact comprised of significantly periodic content. The mean autocorrelation $R_{x,\mathrm{M}}$ (μ = -164.01 dB, σ = 19.46 dB) showed very little deviation across the body of stimuli. Jitter ($F0_{jitt,\mathrm{M}}$, μ = 2.65%, σ = 1.16%) and shimmer ($L_{AI,shi\mathrm{mm}}$, μ = 8.0%, σ = 2.47 %) exhibited relatively high values according to the PRAAT reference of 1.04% and 3.810% for voice pathology respectively.

Table 4.1 Statistical summary of LLD-functional pairs. All values are calculated across all encoders, affects, scripts, and utterances.

f(LLD)	Units	Mean	Min	Max	SD
T_{utt}	s	1.29	0.44	3.24	0.35
T_{vo}	s	0.82	0.10	2.52	0.26
P_{paus}	s/s	35.18	0.23	87.24	14.61
F_{syll}	Syl./s	4.00	1.10	9.00	1.06
$L_{AI,\mathrm{M}}$	dB re: 1	-29.14	-49.64	-5.54	7.60
$L_{AI,\mathrm{R}}$	dB re: 1	33.78	10.59	107.20	11.75
$L_{AI,\mathrm{SD}}$	dB re: 1	7.30	2.32	33.25	2.48
$F0_{\mathrm{M}}$	Hz	229.23	101.96	597.01	68.78
$F0_{\mathrm{R}}$	Hz	173.05	3.87	792.37	136.64
$F0_{\mathrm{SD}}$	Hz	46.25	1.35	246.46	38.07
S_{a}	dB re: 1	6.45	-12.56	25.33	6.06
$S_{ha\mathrm{mm}}$	dB re: 1	18.72	-3.22	42.52	7.73
S_{hf500}	dB re: 1	1.29	0.09	11.30	1.15
S_{hf1k}	dB re: 1	0.60	0.05	4.25	0.44
S_{cog}	dB/Hz	1004.77	194.28	6246.53	581.30
S_{slope}	Hz	-0.004	-0.013	0.006	0.002
$F1_{\mathrm{M}}$	Hz	513.08	253.80	779.69	74.90
$F1_{\mathrm{SD}}$	Hz	169.52	25.22	273.62	33.36
$F1_{\mathrm{bw,M}}$	Hz	172.50	5.14	427.61	58.21
$F2_{\mathrm{M}}$	Hz	2012.16	1395.97	2473.71	176.89
$F2_{\mathrm{SD}}$	Hz	499.99	156.78	798.32	83.30
$F2_{bw,\mathrm{M}}$	Hz	307.64	82.73	494.85	47.89
$F1_{prec,\mathrm{M}}$	%	1092.18	0.29	9753.28	925.41
$F2_{prec,\mathrm{M}}$	%	484.44	0.77	4326.91	430.31
HNR_{M}	dB re: 1	7.49	-0.53	22.04	3.29
$R_{x,\mathrm{M}}$	dB re: 1	-164.01	-318.72	-67.17	19.46
$F0_{jitt,\mathrm{M}}$	%	2.65	0.45	10.41	1.16
$L_{A,shi\mathrm{m}}$	%	8.00	2.41	31.66	2.47

Note: T_{utt} *= duration of the utterances (s),* T_{vo} *= duration of the voiced sections (s),* P_{paus} *= proportion of pauses,* F_{syll} *= syllabic rate (Syl./s),* $L_{AF,\mathrm{M}}$ *= mean intensity (dB),* $L_{AF,R}$ *= range of intensity,* $L_{AF,\mathrm{SD}}$ *= standard deviation of intensity,* $F0_{\mathrm{M}}$ *= mean fundamental frequency,* $F0_{R}$ *= range of fundamental frequency,* $F0_{\mathrm{SD}}$ *= standard deviation of fundamental frequency,* S_{alpha} *= spectrum alpha ratio,* $S_{ha\mathrm{MM}}$ *= spectrum Hammarberg Index,* S_{hf500} *= spectrum high frequency energy ratio (*f_c *= 500 Hz),* S_{hf1k} *= spectrum high frequency energy ratio (*f_c *= 1 kHz),* S_{cog} *= spectrum center of gravity,* S_{slope} *= spectral slope,* $F1_{\mathrm{M}}$ *= mean first formant frequency (Hz),* $F1_{\mathrm{SD}}$ *= standard deviation of the first formant frequency,* $F1_{bw,\mathrm{M}}$ *= mean bandwidth of first formant,* $F1_{prec,\mathrm{M}}$ *= first formant frequency precision,* HNR_{M} *= mean harmonics-to-noise ratio,* $R_{x,\mathrm{M}}$ *= mean autocorrelation,* $F0_{jitt,\mathrm{M}}$ *= mean jitter of the fundamental frequency,* $L_{A,shi\mathrm{M}}$ *= mean shimmer of the intensity.*

The next stage of the analysis aims to identify how each of the LLDs correlate with one another. Before running between-LLD correlations, values for each LLD were normalized to each speaker's mean value and standard deviation. This maneuver follows the recommendations made by Banse & Scherer (1996) who note that speaker-specific baseline values for many of the acoustic cues may mask emotion's differential effect. Table 4.2 gives Pearson's linear correlation coefficient for each LLD after normalizing to within subject mean and variance across all four affect.

Beginning from the left column labeled "1", as expected, T_{utt} was most highly correlated with T_{vo} ($r = 0.94$, $p < 0.05$) and P_{paus} ($r = 0.83$, $p < 0.05$). The overall trend in the matrix results show high correlations for the same LLDs with different functionals applied to them such as $F0_{\mathrm{R}}$ and $F0_{\mathrm{SD}}$ ($r = 0.87$, $p < 0.05$). Spectral measures such as S_{a} and S_{hf1k} ($r= -0.68$, $p < 0.05$) also correlated as expected. This is likely due to the similarity of their precise definitions as ratios of summed spectral energy (the numerator and denominator of S_{a} and S_{hf1k} are similar if reversed*)*. HNR_{M} and $F0_{\mathrm{M}}$ were moderately correlated ($r = 0.26$, $p < 0.001$). The sorted sum of the absolute values in this matrix (excluding duration quantities) S_{a} , $S_{ha\mathrm{mm}}$, $F1_{\mathrm{M}}$, $L_{AF,\mathrm{SD}}$, $F2_{\mathrm{M}}$, and $F0_{\mathrm{M}}$, as the top six most highly correlated LLDs with the set.

Table 4.2 This is a matrix of Pearson's linear correlation coefficients for each of the LLD-functional pairs. Bold face text indicates statistical significance of $p < 0.05$. The numbers next to the LLD symbols indicate their relative position on the perpendicular

ID	Cue	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	T_{utt}	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2	T_{vo}	**.94**	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	P_{paus}	**.83**	**-.87**	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	F_{syll}	**-.11**	**-.12**	**-.14**	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
5	$L_{AI,\mathrm{M}}$	.00	-.01	**-.07**	.00	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	$L_{AI,\mathrm{R}}$	-.02	**.04**	.01	.02	.02	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
7	$L_{AI,\mathrm{SD}}$	**.12**	**-.12**	**.05**	**-.1**	**.1**	**.75**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
8	$F0_{\mathrm{M}}$	**-.05**	**.03**	**.09**	**.06**	**.18**	.02	**-.11**	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
9	$F0_{\mathrm{R}}$	**.06**	-.02	**-.06**	.00	**.05**	.00	-.02	**.05**	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
10	$F0_{\mathrm{SD}}$	**-.03**	.00	.00	.00	**-.03**	.00	**.05**	**.31**	**.87**	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
11	S_{a}	.00	**.03**	-.02	**.13**	.00	**-.07**	**.08**	**-.11**	**.03**	.00	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
12	S_{hamm}	**.08**	**-.1**	-.01	**-.09**	**-.04**	**.06**	**-.11**	**-.05**	.00	-.02	**.51**	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
13	S_{hf500}	**-.04**	**.05**	**.1**	**.03**	**.18**	**-.05**	**.07**	**-.05**	.02	-.02	.02	**.07**	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
14	S_{hf1k}	.02	-.02	**-.07**	.02	**-.03**	**-.04**	**.05**	**.03**	.02	**-.03**	**-.68**	**.15**	**.53**	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.
15	S_{slope}	**.05**	**-.06**	.00	-.02	.00	.02	**-.04**	.00	.02	-.02	**-.09**	**-.12**	**-.06**	**.09**	-	.	.	.	.	.	.	.	.	.	.	.	.	.
16	S_{cog}	**-.08**	**.1**	**.15**	.00	.00	**.04**	**-.1**	**.05**	-.02	**-.02**	**.12**	**-.16**	**-.06**	**-.02**	**.3**	-	.	.	.	.	.	.	.	.	.	.	.	.
17	$F1_{\mathrm{M}}$	**.1**	**-.08**	**-.03**	.00	**-.11**	.00	**-.05**	**.17**	.00	**-.03**	**-.11**	**-.05**	**.21**	**-.22**	-.02	**-.11**	-	.	.	.	.	.	.	.	.	.	.	.
18	$F1_{\mathrm{SD}}$	.00	**.03**	**-.04**	.00	**.05**	**-.04**	**.1**	**-.07**	.00	.00	**-.16**	**.03**	**-.08**	**-.07**	**.03**	**-.1**	**.18**	-	.	.	.	.	.	.	.	.	.	.
19	$F1_{\mathrm{bw,M}}$	**.09**	**-.04**	**-.07**	**.15**	.01	.00	.02	**-.06**	**.03**	**-.04**	**-.04**	.00	**-.09**	.00	.00	**.03**	**.06**	**.02**	-	.	.	.	.	.	.	.	.	.
20	$F2_{\mathrm{M}}$	**.14**	**-.14**	**-.09**	**-.07**	-.02	.00	**-.05**	**.04**	**.06**	**-.06**	**.1**	**-.32**	-.01	**.06**	**-.1**	**.26**	**-.33**	**.09**	**-.11**	-	.	.	.	.	.	.	.	.
21	$F2_{\mathrm{SD}}$	.01	**-.03**	**-.02**	**-.19**	-.02	.00	.00	**-.03**	**.04**	-.02	**-.09**	**.11**	.00	**-.04**	**-.05**	**.2**	**-.16**	**.11**	-.02	**-.3**	-	.	.	.	.	.	.	.
22	$F2_{bw,\mathrm{M}}$	.00	-.01	**.06**	**-.07**	**-.06**	**.03**	**-.07**	**.05**	**-.03**	**.04**	.02	**.15**	**.06**	.00	**.11**	**.07**	**-.05**	**.11**	**.03**	.00	**-.12**	-	.	.	.	.	.	.
23	$F1_{prec,\mathrm{M}}$	.02	.00	**-.02**	**-.08**	**.13**	.02	.00	**.16**	-.02	**.02**	**.03**	.00	.01	.00	.01	**.07**	.00	**-.05**	**.05**	.02	.00	-.02	-	.	.	.	.	.
24	$F2_{prec,\mathrm{M}}$	-.02	.01	-.02	**-.08**	**.12**	**-.02**	**.04**	**-.05**	.00	**.04**	**.05**	.00	**.06**	**.04**	**.02**	.00	**.04**	.00	.00	**.04**	**.1**	.00	**.18**	-	.	.	.	.
25	HNR_{M}	.01	**.02**	**-.18**	-.02	**-.17**	-.01	**.1**	**.26**	**-.09**	**-.05**	**.1**	**.27**	**-.02**	.01	**-.07**	**.1**	**-.24**	.00	**.04**	**.03**	**-.05**	**.04**	**.03**	-.01	-	.	.	.
26	$R_{x,\mathrm{M}}$	.00	.01	**-.03**	.02	**-.07**	**.1**	**-.03**	**-.11**	-.01	.00	.00	.01	**-.03**	**-.04**	.00	.00	.02	.00	**.09**	**.03**	**-.04**	**.04**	**.05**	.00	**.07**	-	.	.
27	$F0_{jitt,\mathrm{M}}$	.00	.00	**.18**	**-.08**	**-.12**	.02	**-.13**	**.04**	.01	.00	**.07**	**.1**	**-.1**	**.05**	**-.09**	**.04**	**-.14**	**.04**	.02	-.01	**-.02**	**.03**	.02	.00	**-.24**	**.14**	-	.
28	$L_{A,shim,\mathrm{M}}$	.02	-.02	**-.04**	**.16**	**-.09**	**.03**	**.03**	**-.15**	**.03**	-.02	**-.06**	.01	**.06**	**-.05**	**.06**	.00	**-.02**	**.09**	.00	.00	.00	**.03**	**.05**	**.04**	**-.15**	**-.08**	**.49**	-

The predictions made by Scherer (1986) indicate the direction that an LLD is likely to move in relative to its value in an unemotional state. Magnitude of changes in each LLD have also been provided where applicable. These predictions provide a preliminary set of references to compare the general acoustical trends of the speech corpus. Because Scherer's definitions are more broad than many of the exact definitions of each LLD equation, some comparisons between prediction and LLD will require some minor interpretation. Scherer's predictions indicate changes in acoustics resulting from the range of applicable stimulus evaluation checks (SEC) [6]. As a reminder, the hierarchical performance of these five SECs is the essential basis for emotional differentiation according to Scherer (see Table 4.3). The "novelty check" is the first SEC that the encoder will perform. Here, the encoder determines whether any patterns of external or internal stimuli have changed in a noticeable or abrupt manner. Next, the encoder performs an "Internal pleasantness check" to determine if this stimulus is enjoyable or not. The three following SECs include: "Goal/need significance", "Coping potential", and "Norm/self compatibility" checks of the stimulus.

Referred to as anger, happiness, and sadness, these three emotional states as intended are most similar to equivalent to Scherer's definitions of rage or hot anger, elation or joy, and sadness or dejection. Stimuli producing an angry emotional state (as hypothesized by Scherer) would be interpreted as very novel, [pleasant = open], extremely significant and obstructing of the encoder's goals, and in immediate need of a behavioral response. Additionally, the stimulus would be evaluated as having a high coping potential, and that the cause of the stimulus incompatible with both societal and personal norms.

Table 4.3 Summary of the stimulus evaluation checks in their order of processing (adapted from Scherer 1986).

Category	SEC	Anger	Happiness	Sadness
1	Novelty	high	high	low
2	Pleasantness	open	high	low
3.1	Relevance	high	high	high
3.2	Expectation	discrepant	discrepant	discrepant
3.3	Conduciveness	obstruct	high	obstruct
3.4	Urgency	high	low	low
4.1	Control	high	-	none
4.2	Power	high	-	-
4.3	Adjustment	high	medium	medium
5.1	External	low	high	-
5.2	Internal	low	high	-

Scherer continues to define groupings of these SECs in terms valence, activation, and power dimensions. The term "Hedonic Valence" refers to the generally positive or negative nature of the pleasantness SEC (#2) and the conduciveness SEC (#3.3). The "Activation" grouping refers to the combined influence of the outcomes from the relevance (see Table 4.4), expectation (3.2), and urgency (3.4), subchecks. These are thought to influence the dominance of the sympathetic nervous system (ergotropic arousal). Recall that the sympathetic nervous system is a branch of the autonomic nervous system (ANS) that determines when to expend lots of energy depending in an emergency. Lastly, the power group refers to the fourth SEC or "Coping potential."

Table 4.4 Groupings for the SECs by valence, activation, and power. Adapted from Scherer (1986).

Emotion	Hedonic Valence	Activation	Power
Anger	narrow	very tense	extremely full
Happiness	wide	medium-tense	medium-full
Sadness	narrow	lax	thin

The scope of this study is limited to three emotions. This merits a hierarchical organization in favor of emotion rather than LLD category. Based on the overall outcomes of the valence, activation, and power SEC groupings, anger would be characterized by a narrow, very tense, and extremely full voice. Figure 4.10 illustrates the specific source of many of these outcomes in sequential order.

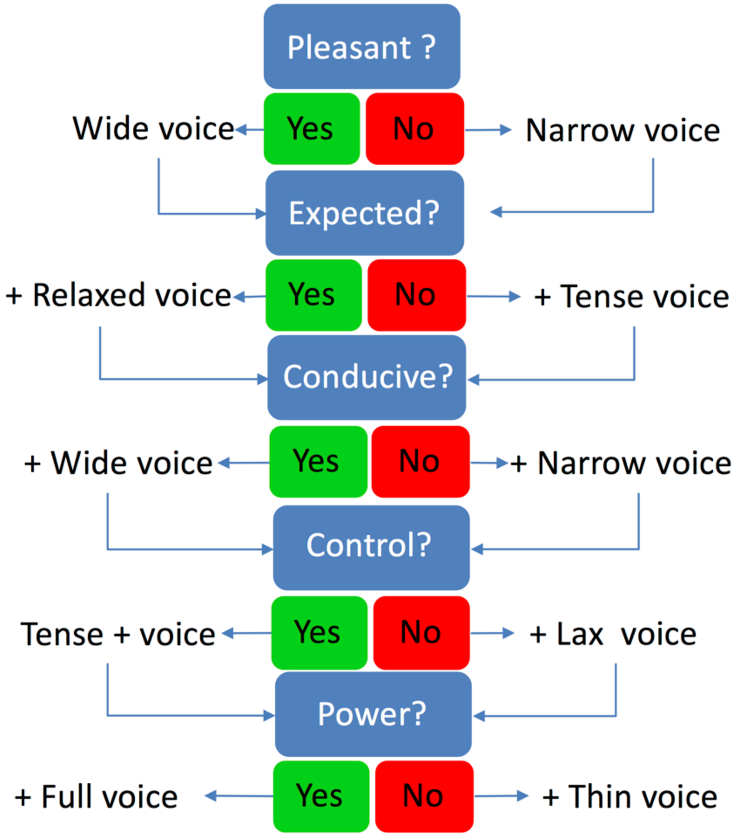


Figure 4.10 Chronological flow of SECs and the hypothesized voice type each outcome may produce.

A narrow voice would be characterized by faucal and pharyngeal constriction, a tensing of the tract walls, shortening of the vocal tract and corners of the mouth retracted downward. These physical changes would increase the high frequency energy, while increasing the frequency of the first formant and sharpening the first formant bandwidth. The second and third formant would likely decrease in frequency. A tense voice would likely consist of increased tension in the entire system: constrictions of the upper larynx and pharynx, and tensing of the muscles surrounding the larynx. A decrease in salivation would likely occur. These changes would likely raise the *F0* of the glottis, further boost both high frequency energy, decrease the bandwidth of the first formant, and possibly increase jitter and shimmer. Finally, the fullness of the angry voice would work to further relax the vocal apparatus, deepen the amount of respiration, and move the phonation register toward the chest. This would lower the *F0* and boost the overall loudness of the voice.

Table 4.5 gives the predicted and observed changes in LLDs from a non-expressive state to anger. Of the measured acoustic cues with available predictions, nine followed the predicted direction of change, and three changed in the opposite direction. Nine out of the twelve or 75% of the results for anger agreed with predictions. The prediction for $F0_{\mathrm{M}}$ indicates that increases or decreases should be expected depending on which voice type dominates the expression. For anger, the opposing effects may be the power SEC decreasing $F0$ and SECs within the activation category that increase $F0$ (see Figure 4.10).

Table 4.5 Comparison of predictions and outcomes for differences in LLDs between a non-expressive state and anger. + indicates an increase, - indicates a decrease, and = indicates no change.

		Emotion			
		Anger (rage)			
Category	**Acoustic Cue**	**Prediction**	**Result**	**Agreement**	**Symbol**
F0	Perturbation	+	-	no	$F0_{jitt,\mathrm{M}}$
	Mean	±	+	-	$F0_{\mathrm{M}}$
	Range	+ +	+	yes	$F0_{R}$
	Variability	+ +	+	yes	$F0_{\mathrm{SD}}$
Intensity	Mean	+ +	+	yes	$L_{AI,\mathrm{M}}$
	Range	+	+	yes	$L_{AI,R}$
	Variability	+	+	yes	$L_{AI,\mathrm{SD}}$
Voice Quality	High-Frequency Energy	+ +	+	yes	S_{hf1k}
	Spectral Noise		+		$(HNR_{\mathrm{M}})^{-1}$
	F1 mean	+	+	yes	$F1_{\mathrm{M}}$
	F2 mean	-	+	no	$F2_{\mathrm{M}}$
	F1 bandwidth	- -	-	yes	$F1_{bw,\mathrm{M}}$
	Formant precision	+	+	yes	$F1_{prec,\mathrm{M}}$
Duration	Speech Rate	+	-	no	F_{syll}

The results for anger align well with predictions for LLDs in the intensity category, however differences can be seen in the $F0$ category, voice quality, and duration category LLDs. Anger produced significantly lower $F0$ perturbations ($F0_{jitt,\mathrm{M}}$), higher $F0_{\mathrm{M}}$ than neutral expressions, and both the second formant mean frequency ($F1_{\mathrm{M}}$) and speech rate (F_{syll}) moved in opposite directions from predicted. Scherer's predictions for the movement of the second and third formant frequencies is largely dependent on a combination of facial expressions and laryngeal constrictions. Many possible factors such as the text of the words spoken could have influenced $F2_{\mathrm{M}}$ to increase consistently from neutral. The decrease in speech rate from a non-expressive state to anger was an unexpected result and one that opposes the general consensus of the literature to date [10] [22]. Further discussion of these findings is provided in the following sections.

Table 4.6 gives the predicted change in LLD for expressions of happiness and the observed delta. Nine out of the twelve or 75% of the results for happiness agreed with predictions. Strong agreement was found for all *F0* and intensity LLDs except for *F0* perturbations ($F0_{jitt,\mathrm{M}}$), which were lower than predicted. Only the formant precision ($F1_{prec,\mathrm{M}}$) results agreed with predicted changes with unambiguous directions. As seen in anger, happiness was also marked by significantly lower speech rates than the non-expressive state. This too counters the consensus found in the literature.

Table 4.6 Comparison of predictions and outcomes for differences in LLDs between a non-expressive state and happiness. + indicates an increase, - indicates a decrease, and = indicates no change.

		Emotion			
		Happiness (joy)			
Category	**Acoustic Cue**	**Prediction**	**Result**	**Agreement**	**Symbol**
F0	Perturbation	+	-	no	$F0_{jitt,\mathrm{M}}$
	Mean	+	+	yes	$F0_{\mathrm{M}}$
	Range	+	+	yes	$F0_{R}$
	Variability	+	+	yes	$F0_{\mathrm{SD}}$
Intensity	Mean	+	+	yes	$L_{AI,\mathrm{M}}$
	Range	+	+	yes	$L_{AI,R}$
	Variability	+	+	yes	$L_{AI,\mathrm{SD}}$
Voice Quality	High-Frequency Energy	±	+	yes	S_{hf1k}
	Spectral Noise		+		$(HNR_{\mathrm{M}})^{-1}$
	F1 mean	-	+	no	$F1_{\mathrm{M}}$
	F2 mean		=		$F2_{\mathrm{M}}$
	F1 bandwidth	±	-	Yes	$F1_{bw,\mathrm{M}}$
	Formant precision	+	+	Yes	$F1_{prec,\mathrm{M}}$
Duration	Speech Rate	+	-	No	F_{syll}

Fewer rows of agreement between predicted and observed trends for sadness are found in Table 4.7. Just three out of the fourteen or 21% of the results for sadness agreed with predictions. None of the *F0* LLD predictions aligned perfectly with the observed trends. Sadness had statistically equivalent values for $F0_{\mathrm{M}}$ compared to neutral. Where the range and variation of the *F0* values were predicted to decrease, the opposite was also observed. The voice quality category and duration categories in Table 4.7 indicate some limited agreement however. High frequency energy (S_{hf1k}) decreased and the mean of the first formant's bandwidth ($F1_{bw,\mathrm{M}}$) increased as predicted.

Table 4.7 Comparison of predictions and outcomes for differences in LLDs between a non-expressive state and sadness. + indicates an increase, - indicates a decrease, and = indicates no change.

		Emotion			
		Sadness (dejection)			
Category	**Acoustic Cue**	**Prediction**	**Result**	**Agreement**	**Symbol**
F0	Perturbation	+	-	no	$F0_{jitt,\mathrm{M}}$
	Mean	±	=	no	$F0_{\mathrm{M}}$
	Range	-	+	no	$F0_{\mathrm{R}}$
	Variability	-	+	no	$F0_{\mathrm{SD}}$
Intensity	Mean	- -	=	no	$L_{AI,\mathrm{M}}$
	Range	-	+	no	$L_{AI,R}$
	Variability	-	+	no	$L_{AI,\mathrm{SD}}$
Voice Quality	High-Frequency Energy	±	-	yes	HF_{1000}
	Spectral Noise	+	-	no	$(HNR_{\mathrm{M}})^{-1}$
	F1 mean	+	-	no	$F1_{\mathrm{M}}$
	F2 mean	-	=	no	$F2_{\mathrm{M}}$
	F1 bandwidth	±	+	yes	$F1_{bw,\mathrm{M}}$
	Formant precision	-	+	No	$F1_{prec,\mathrm{M}}$
Duration	Speech Rate	-	-	yes	F_{syll}

Formant precision has been defined as “the degree to which formant frequencies attain values prescribed by phonological system of a language.” The prediction for a low arousal and negatively valenced emotion is for this “degree” to decrease rather than increase. Previous studies have also measured the precision of articulation, or the absolute difference between the first formant frequency and its value in a non-expressive state [22]. Higher values of this metric indicate an increase in articulatory effort and are therefore more precise. By this definition, only the frequency values of the first formants for anger, happiness and sadness can be compared with one-another. In this case, happiness exhibited the greatest precision of articulation for the first formant, followed by anger and sadness (statistically equal). The second formant’s precision trended differently with the greatest precision given to anger, followed by happiness and sadness. $F2_{prec,\mathrm{M}}$ values for happiness and sadness were statistically equivalent. When comparing formant bandwidths it should be noted that the first and second formant’s bandwidths are significantly affected by wall loss and the rounded quality of the glottal pulses [11]. The results very clearly show that sadness has significantly broader bandwidths compared to any other expression, which may be indicative more relaxation in the vocal tract compared to the neutral affect.

LLDs in the duration category exhibit significant trends that oppose Scherer's predictions. Defined as the "number of speech segments per time unit," "Speech Rate" can be compared to both the syllabic rate and the inverse of the utterance length. The results for syllabic rate indicate statistically significant differences between all four modes of expression. Expressions of both anger and happiness exhibited much lower speech rates than the neutral expressions. It is unclear why emotions that typically fit into a highly aroused dimension would be marked by such trends.

The difference between the presented results and the consensus of the related works on speech rate may also be caused by the inconsistency of the reference frame. Juslin & Laukka (2003) made very clear the difficulty with which fair comparison can be conducted between studies take the average LLD value exclusively across emotional expressions versus studies that report deviations from a neutral affect. Such a difference would account for nearly all of the gross deviations for sadness if the neutral expressions were excluded. $F0_{\mathrm{M}}$, $L_{AI,\mathrm{M}}$, $L_{AI,\mathrm{SD}}$ all share relatively low values compared the means of the emotional expressions but higher than the non-emotional expressions.

Banse & Scherer (1996) also tested these predictions in a study with twelve professional German speaking actors (6 male, 6 female). Fourteen emotions were studied, and the stimuli consisted of scrambled phonemes of Indo-European origin. The values presented in Table 4.8 are extracted from Banse & Scherer's labels for HAn (hot anger), Sad, and Ela (elation) in the columns "B & S (1996)". Adjacent to these columns are those for the presented work "PEEP (2017)." Note that it appears as though the definition used for the S_{hamm} were difference rather than a ratio of spectral power. The later values of PEEP (2017) have been adjusted to reflect this proportional difference and sign. Mean-squared-error (MSE) terms were calculated between the mean values given and not the standard deviations. Smaller MSE values indicate better agreement between these sets of values. Results were more consistent for happiness (MSE = 0.24) than anger (MSE = 0.37) or sadness (MSE = 0.47).

Table 4.8 Means of the normalized LLDs from Banse and Scherer (1996) and the present study. Neutral LLDs for PEEP (2017) were included in the normalization to better represent 25% of the corpus.

	Anger		**Happiness**		**Sadness**	
Cue	B & S (1996)	PEEP (2017)	B & S (1996)	PEEP (2017)	B & S (1996)	PEEP (2017)
T_{vo}	-0.45 ± 0.66	0.04	-0.34	0.02	1.25	0.56
$L_{AI,\mathrm{M}}$	1.19	0.98	1.05	0.58	-1.16	-0.80
$F0_{\mathrm{M}}$	1.13	0.01	1.24	1.39	-0.32	-0.69
$F0_{\mathrm{SD}}$	0.50	0.11	0.21	1.28	0.43	-0.64
$S_{h\mathrm{amm}}$	1.13	0.41	0.58	0.44	-0.43	0.62
S_{hf500}	-0.55	-0.79	-0.29	-0.29	1.23	0.67
S_{hf1k}	-1.34	-0.77	-0.05	-0.38	0.90	0.71
	MSE:	0.37	**MSE:**	0.24	**MSE:**	0.47

The analysis up to this point has only examined general, surface level trends found in the acoustics of the speech corpus. Still keeping to the utterance level as a unit of analysis, this section will shed light on how the full list of acoustic cues particularly in the voice quality category compare to previous work. A review by Juslin & Laukka (2003) compiled a broad list of acoustic cues in 77 studies of vocal expression. Although predisposed to the objective use of figures and tables, trends were also given in terms of each author's overall categorization of the cues as affected by the emotions considered ("high", "medium", "low"). Unfortunately, this does not solve the issue of reference frame, however the body of work does lay significant groundwork for creating general consensus apart from predictions over a larger number of LLDs.

In order to compare to these trends, a one-way repeated measures ANOVA was performed on the LLDs with emotion (anger, happiness, and sadness), script (1-8), and utterance (1-4) as effects [add stat reference]. The analysis was repeated over encoder (1-55), which effectively incorporates the covariance structure of each encoder's LLD vector such that preliminary within-subject normalization was not necessary. That being said, the salience of acoustical changes that can be easily compared to analytically do not include a neutral affect. A perfectly normal and emotionless expression treatment within the effect of emotion can be thought of as a relatively dissimilar variable. Anger, happiness, and sadness all offer the presence of emotion, whereas "the neutral" expressions offer by definition zero emotion at all. So running analyses where the emotion type is of concern may risk definitional dissonance when one treatment such as anger is compared to no emotion whatsoever.

It was concluded that the optimal approach to ensuring a relatable and comparable analysis of variance was to normalize the LLD vectors across (not within) all encoders with the non-affective (neutral) expressions included, then enter just the matrices for expressions of anger, happiness and sadness into the repeated measures ANOVA design. A mixed procedure ANOVA was conducted on the LLDs using the statistical software SAS, which included Type 3 tests of mixed effects [70]. For each LLD, the mixed procedure process was performed using a compound symmetry covariance structure, encoder as a subject effect, and a Kenward-Roger method for the fixed effect squared error and degrees of freedom. The type three tests of fixed effects reported the significance of each of the hypothesis tests for the fixed effects (emotion, script, utterance) and their interactions. F statistics and their corresponding p-values for each of these tests are provided in Table 4.9.

The LLDs with the greatest F-statistics from Table 4.9 are as follows: $F0_{\mathrm{M}}$, $L_{AI,\mathrm{M}}$, S_{a}, $S_{ha\mathrm{mm}}$, $F0_{\mathrm{SD}}$, $F0_{\mathrm{R}}$, HNR_{M}, S_{hf1k}, S_{hf500}, and $F1_{\mathrm{M}}$ ($p < 0.0001$). The effect of script was significant across all LLDs except for $L_{AI,\mathrm{M}}$, the most significant being voice duration T_{vo}, $F2_{\mathrm{M}}$, and F_{syll}. Generally speaking, cues in the time-sensitive category tended to be more sensitive to script effects. This result was expected given the limited range of values that the utterances could have. While $L_{AI,\mathrm{M}}$ ranked near the top of the most significant cues to emotion, it was the least significant for script effects. The effect of utterance was less pronounced on the LLDs overall. This observation is probably indicative of the weak relative meaning of the utterance number as it was chronologically recorded. Areas of significance in this set of analyses could suggest that specific utterances within each script caused greater change in the LLDs, or that encoders favored the second or third phrase as spoken. With 5280 degrees of freedom, these results are almost sure to be significant even with 99.999% confidence level. Also, the significance reported does not necessarily apply to all emotions equally. To get at the specific level differences within each affect, Tukey least-squares means comparisons were run between modes of expression. Here, the neutral affect has been included as a comparison, even though it was not entered into the initial mixed data procedure.

Table 4.9 Results from the repeated measures ANOVA for each of the 28 LLDs and their functionals.

	Variable[vi]					
	Emotion (E)		**Script (S)**		**Utterance (U)**	
f(LLD)	***F***	X^2	***F***	X^2	***F***	X^2
T_{utt}	1038.48***	2076.96	161.88***	1133.15	20.52***	61.57
T_{vo}	431.69***	863.37	347.46***	2432.21	11.98***	35.95
P_{paus}	210.26***	420.53	247.***	1728.97	39.58***	118.73
F_{syll}	1060.39***	2120.78	331.37***	2319.56	298.92***	896.76
$L_{AI,\mathrm{M}}$	4897.98***	9795.96	2.59	18.12	7.53***	22.58
$L_{AI,\mathrm{R}}$	92.89***	185.77	18.03***	126.23	3.19	9.56
$L_{AI,\mathrm{SD}}$	277.46***	554.91	43.6***	305.23	9.25***	27.76
$F0_{\mathrm{M}}$	6330.59***	12661.20	36.45***	255.12	17.13***	51.38
$F0_{\mathrm{R}}$	3179.91***	6359.83	25.71***	179.96	14.58***	43.73
$F0_{\mathrm{SD}}$	4126.6***	8253.20	40.51***	283.54	11.33***	33.98
S_{alpha}	4764.97***	9529.94	193.23***	1352.60	89.99***	269.97
$S_{h\mathrm{amm}}$	4733.41***	9466.82	59.52***	416.66	109.02***	327.05
S_{hf500}	1872.39***	3744.79	148.1***	1036.73	5.66*	16.99
S_{hf1k}	2020.81***	4041.61	119.14***	833.95	39.53***	118.58
S_{cog}	598.63***	1197.26	35.28***	246.93	58.47***	175.42
S_{slope}	469.16***	938.32	257.29***	1801.06	118.98***	356.93
$F1_{\mathrm{M}}$	1075.84***	2151.68	109.69***	767.82	83.64***	250.93
$F1_{\mathrm{SD}}$	190.3***	380.60	26.23***	183.60	63.44***	190.32
$F1_{bw,\mathrm{M}}$	117.05***	234.09	31.35***	219.45	79.92***	239.76
$F2_{\mathrm{M}}$	237.73***	475.47	346.59***	2426.14	57.16***	171.49
$F2_{\mathrm{SD}}$	118.48***	236.97	223.41***	1563.89	76.91***	230.74
$F2_{bw,\mathrm{M}}$	401.52***	803.04	44.01***	308.06	28.2***	84.60
$F1_{prec,\mathrm{M}}$	25.37***	50.74	5.75***	40.28	22.83***	68.50
$F2_{prec,\mathrm{M}}$	60.62***	121.24	15.07***	105.50	2.38	7.13
HNR_{M}	2464.61***	4929.22	169.46***	1186.23	42.21***	126.63
$R_{x,\mathrm{M}}$	682.43***	1364.86	22.39***	156.76	19.26***	57.79
$F0_{jitt,\mathrm{M}}$	406.11***	812.22	63.45***	444.18	12.78***	38.33
$L_{A,shim}$	189.05***	378.10	27.99***	195.95	1.54	4.61

[vi] *F statistics with three stars next to them are significant to $p < 0.0001$. These results indicate a highly significant effect of emotion type (without the no expression values). T_{utt} = duration of the utterances (s), T_{vo} = duration of the voiced sections (s), P_{paus} = proportion of pauses, F_{syll} = syllabic rate (Syl./s), $L_{AI,\mathrm{M}}$ = mean intensity (dB), $L_{AI,R}$ = range of intensity, $L_{AI,\mathrm{SD}}$ = standard deviation of intensity, $F0_{\mathrm{M}}$ = mean fundamental frequency, $F0_R$ = range of fundamental frequency, $F0_{\mathrm{SD}}$ = standard deviation of fundamental frequency, S_{alpha} = spectrum alpha ratio, $S_{ha\mathrm{MM}}$ = spectrum Hammarberg Index, S_{hf500} = spectrum high frequency energy ratio (f_c = 500 Hz), S_{hf1k} = spectrum high frequency energy ratio (f_c = 1 kHz), S_{cog} = spectrum center of gravity, S_{slope} = spectral slope, $F1_{\mathrm{M}}$ = mean first formant frequency (Hz), $F1_{\mathrm{SD}}$ = standard deviation of the first formant frequency, $F1_{bw,\mathrm{M}}$ = mean bandwidth of first formant, $F1_{prec,\mathrm{M}}$ = first formant frequency precision, HNR_{M} = mean harmonics-to-noise ratio, $R_{x,\mathrm{M}}$ = mean autocorrelation, $F0_{jitt,\mathrm{M}}$ = mean jitter of the fundamental frequency, $L_{A,shim\mathrm{M}}$ = mean shimmer of the intensity.*

Figure 4.11 represents the Tukey least-squares means comparisons performed on four of the LLD categories. The mean values and standard errors have been adjusted to agree with the compound-symmetry structure of the repeated measures model. All differences in the means of the populations were significant except for Neutral X Sadness for $L_{AI,\mathrm{M}}$ and $F0_{\mathrm{M}}$, which is evident in the upper left and right plots of Figure 4.11. The remaining means comparisons are located in Appendix G.

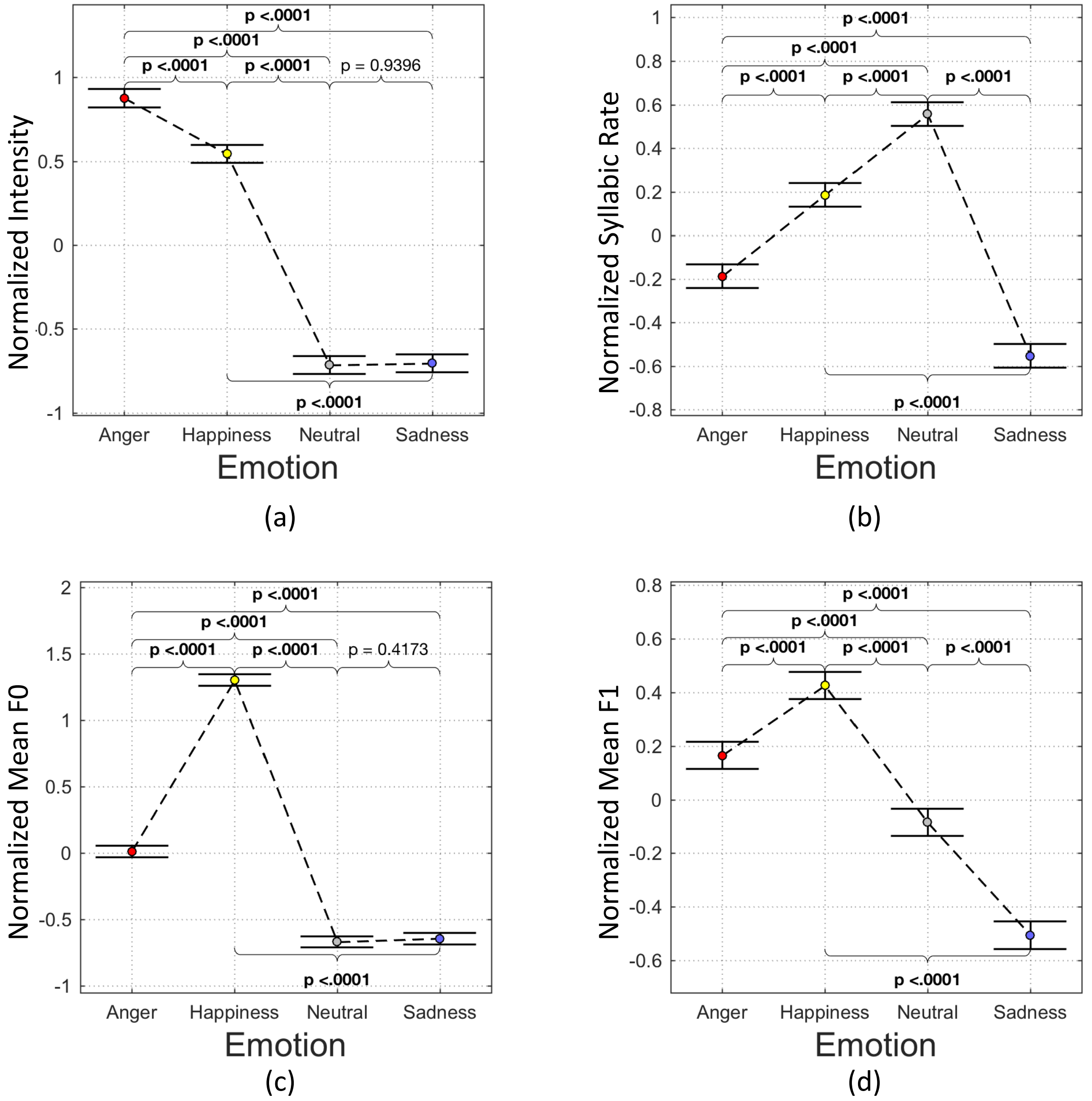


Figure 4.11 Mean estimates and standard errors for the LLDs between all modes of expression including neutral. (a) mean intensity (b) syllabic rate, (c) mean F0, (d) mean F1, All values are zscores taken across all encoders (not within).

These results largely support trends both predicted by Scherer (1986) and consensus determined by Juslin & Laukka (2003). Generally speaking, anger and happiness had significantly greater intensity and *F0* values than did sadness.

Compared across the three emotions, syllabic rate trended was predicted if neutral is ignored, although recent studies have found similar results [24]. Predictions and consensus for the first formant differ. Scherer predicts that happiness should see a lowering of the first formant, while Juslin and Laukka determined that five other studies have found the opposite trend.

In addition to a mixed data model, a general linear two-way ANOVA was performed on the data. This analysis was done in an effort to extract more concrete numerical characteristics that speak to the degree to which variance is explained by certain LLDs. This procedure mirrors much of the work by Juslin and Laukka (2001), except instead of splitting the design across emotion and intensity, the factors used here were emotion, script, and utterance. The obvious disadvantages of this method are the linear assumptions of the spacing of emotions. The LLDs targeted for this next step were focused on those shared by Juslin and Laukka (2001). The results for this analysis are provided in Table 4.10.

Table 4.10 Linear ANOVA results from PEEP and Juslin & Laukka (2001). From left to right are columns for the acoustic cue label, F-statistic, effect size, and p-value.

	PEEP (2017)				**J & L (2001)**		
	Emotion (E)				Emotion (E)		
cue	***F***	η^2	***p***	**cue**	***F***	***p***	η^2
P_{paus}	87.17	0.03	***	Pause Prop	3.36	*	0.22
F_{syll}	353.52	0.12	***	Speech Rate	17.73	***	0.6
$L_{AI,\mathrm{M}}$	2818.06	0.52	***	VoInt (M)	25.93	***	0.68
$L_{AI,\mathrm{SD}}$	203.65	0.07	***	VoInt (SD)	16.82	***	0.58
$F0_{\mathrm{M}}$	4083.77	0.61	***	*F0* (M)	14.02	***	0.54
$F0_{\mathrm{SD}}$	3393.09	0.56	***	*F0* (SD)	14.27	***	0.54
S_{hf500}	1010.89	0.28	***	HF 500	26.63	***	0.69
S_{hf1k}	1060.69	0.29	***	Hf 1000	8.84	***	0.42
$F1_{\mathrm{M}}$	482.16	0.16	***	F1	8.73	***	0.42
$F1_{bw,\mathrm{M}}$	81.35	0.03	***	F1bw	7.99	***	0.4
$F1_{prec,\mathrm{M}}$	20.91	0.01	***	F1prec	7.43	***	0.38
$F2_{\mathrm{M}}$	83.16	0.03	***	F2	0.35		0.3
$F2_{bw,\mathrm{M}}$	263.79	0.09	***	F2bw	0.15		0.01
$F0_{jitt,\mathrm{M}}$	235.75	0.08	***	Jitter	0.98		0.08

Although the sample sizes are very different, there appears to be come agreement between the variance explained by the LLDs (η2). Values for mean sound level $L_{AI,\mathrm{M}}$ (η2 = 0.52, p < 0.0001 ; η2 = 0.68, p < 0.001), $F0_{\mathrm{M}}$ (η2 = 0.61, p < 0.0001 ; η2 = 0.54, p < 0.001), $F0_{\mathrm{SD}}$ (η2 = 0.56, p < 0.0001 ; η2 = 0.54, p < 0.001) all appear to be significant and responsible for some portion of the variance created by emotion. The areas of greatest divergence seem to be between the duration quantities such as F_{syll} and Speech Rate, where the results meet just 20% of the variance explained by Juslin & Laukka (η2 = 0.12, p < 0.0001) vs. (η2 = 0.6, p < 0.001). While still significant, cues such as $F1_{bw,\mathrm{M}}$ and $F1_{prec,\mathrm{M}}$ were much lower in effect size than $F1_{bw}$ and $F1_{prec}$. This could be due to the difference in how this cue was calculated. According to Juslin & Laukka, a single frequency value was used to set what was considered to be the "neutral" formant frequency based on a lossless tube model of a certain length. Presently, $F1_{prec,\mathrm{M}}$ was calculated as a difference between the formant frequencies of an actual attempt at speaking without emotion.

4.5 Discussion

Conclusive remarks regarding the fundamental questions of this chapter rely on a frame of reference that appears to change depending on the study. Results of the acoustic analysis were compared to Scherer's (1986) predictions based on a framework of honest, yet inconclusive hypotheticals. It was found that absolute deviations of acoustic parameters shared very little in common with predictions for displays of sadness, particularly in the case of sound level, *F0*, and formant frequency movements. Overall, the error between predictions and the results largely followed the valence dimension. The biases of the formants towards upper or lower frequencies only once matched for the case of $F1_{\mathrm{M}}$ in anger. There are many possible reasons for this variance including, but not limited to, consistency in the facial features of the encoder, phonetic differences of the scripts, and possibly the history of the encoder as it influences phonetic stylization of specific words. The only way to rule out effects of the personal phonetic stylization would be to sample expressions wide and far enough to average out individual differences. This has been the attempt in the present work, however the set of scripts only contains 20-30 unique words in the English language– many of which may not appear in common vernacular.

The primary question of whether the acoustics of the corpus align with those of previous works requires conditional answers. At best, comparative consensus offers a vague quality assurance tool.

i. Does the presented speech corpus exhibit acoustic cues that agree with the literature?
ii. Do the acoustic cues known to carry emotional information correlate independently of semantic content?

Answering the first question requires a frame of reference for comparison. The effect of emotion on a specific LLD value can assessed as a difference from a non-expressive state or from a different emotional state. Unfortunately, much of the consensus reported in the literature lacks a firm reference frame as is noted by Juslin and Laukka (2003), and so assumptions must be made to qualify a response. LLD deltas were assumed to be calculated in reference to the total set of emotions studied in the case of Juslin and Laukka (2003) or with the neutral version as a reference value.

Compared between *emotions* it appears that the corpus has strong agreement with the literature in almost all cases. When compared to a neutral or non-expressive state, many of the trends oppose those found in the literature, specifically for duration LLDs. Table 4.11 gives a comparison of general rends found by Juslin and Laukka (2003) and the results of the acoustic analysis. Areas where results agreed with the Juslin and Laukka literature review are highlighted in gray. Direction of change for the columns labeled PEEP (2017) are strictly referencing differences from neutral. Columns labeled J & L (2003) are the relative labels given by Juslin and Laukka (2003) and are representative of categorical observations, not necessarily referencing a difference from neutral. Regarding F_{syll}, relative to one-another the results agree with the consensus, however in absolute terms, both anger and happiness had slower syllabic rates than the neutral reference. The proportion of pauses (P_{paus}) did not deviate significantly for portrayals of anger. Sadness trended against the consensus of the literature by decreasing rather than increasing from the neutral portrayals. Compared to LLDs for neutral portrayals, LLDs for anger and happiness changed in directions supported by the literature. Although $F0_{jitt}$ did move in the opposite direction from previous studies, this LLD had the least number of contributors in the review. Overall, 22/33 LLDs or 67% of the results were in favor of the consensus determined by Juslin and Laukka (2003).

Table 4.11 Comparison of trends in LLDs between the results of the acoustic analysis and consensus from the review by Juslin & Laukka (2003).

	F_{syll} (Syl./s)		P_{paus} (s/s)	
Emotion	**J & L (2003)**	**PEEP (2017)**	**J & L (2003)**	**PEEP (2017)**
Anger	Fast	Slower	Low	Same
Happiness	Fast	Slow	Low	Lower
Sadness	Slow	Slowest	High	Low

	$L_{AI,\mathrm{M}}$ (dB)		$L_{AI,\mathrm{SD}}$ (dB)	
Emotion	**J & L (2003)**	**PEEP (2017)**	**J & L (2003)**	**PEEP (2017)**
Anger	High	Higher	High	Higher
Happiness	High	High	High	High
Sadness	Low	Same	Low	High

	$F0_{\mathrm{M}}$ (Hz)		$F0_{\mathrm{SD}}$ (Hz)	
Emotion	**J & L (2003)**	**PEEP (2017)**	**J & L (2003)**	**PEEP (2017)**
Anger	High	High	High	Higher
Happiness	High	Higher	High	Highest
Sadness	Low	Same	Low	High

	S_{hf1k} (dB)		$F1_{prec,\mathrm{M}}$ (%)	
Emotion	**J & L (2003)**	**PEEP (2017)**	**J & L (2003)**	**PEEP (2017)**
Anger	High	Higher	High	Low*
Happiness	High	High	High	High
Sadness	Low	Low	Low	Low

	$F1_{\mathrm{M}}$ (Hz)		$F1_{bw,\mathrm{M}}$ (Hz)	
Emotion	**J & L (2003)**	**PEEP (2017)**	**J & L (2003)**	**PEEP (2017)**
Anger	High	High	Low	Low
Happiness	High	Higher	Low	Lower
Sadness	Low	Low	High	High

	$F0_{jitt,\mathrm{M}}$ (%)	
Emotion	**J & L (2003)**	**PEEP (2017)**
Anger	High	Lower
Happiness	High	Lowest
Sadness	Low	Low

Regarding the effect of semantic content, the results indicate that the main effects of script and the interactions of utterance and script were significant in contributing to the variance among the emotion categories. One can look to the interaction of intended emotion (E) and script (S) on the variance, or the interactions of emotion, script and utterance (U) given in Table 4.12. The interaction of intended emotion and script caused significant differences in all but two out of the twenty-eight LLDs ($L_{AI,\mathrm{R}}$ and $F2_{prec,M}$). The interactions between emotion, script, and utterance was significant for 100% of the LLDs included in the analysis. These results indicate that the type of script affected the acoustical properties of the portrayals of different emotions. Furthermore, the utterance within the script also significantly altered the acoustics of the speech corpus. Although these interactions appear to be significant, the degree to which they affect the total variance remains to be seen. As a coarse form of comparison, the average F-statistics for the main effects and interactions can be compared. The main effect of emotion had an average F-statistic of 1502.7, which is an order of magnitude greater than main effect of script ($F = 112.6$), emotion x script ($F = 8.9$), and emotion x script x utterance ($F = 7.3$). As a result, one can further qualify the conclusions of this chapter. The results indicate that although the trends in acoustics were significantly affected by the semantic content of the scripts, the magnitude of these appear to be small.

Table 4.12 Repeated measures ANOVA results for the interactions between intended emotion, script, and utterance.

	E x S			**E x S x U**		
	F	*p*	χ^2	*F*	*p*	χ^2
T_{utt}	6.87	***	96.16	4.97	***	208.64
T_{vo}	7.01	***	98.1	6.1	***	256.25
P_{paus}	7.39	***	103.5	5.36	***	225.24
F_{syll}	3.54	***	49.61	2.98	***	125.13
$L_{AI,\mathrm{M}}$	6.05	***	84.71	3.68	***	154.54
$L_{AI,\mathrm{R}}$	1.72		24.07	1.96	*	82.27
$L_{AI,\mathrm{SD}}$	2.94	*	41.22	4.45	***	187.03
$F0_{\mathrm{M}}$	17.54	***	245.58	7.19	***	301.85
$F0_{\mathrm{R}}$	12.14	***	169.91	2.88	***	120.99
$F0_{\mathrm{SD}}$	16.39	***	229.49	4.55	***	191.16
S_{a}	13.12	***	183.71	11.29	***	473.99
$S_{ha\mathrm{mm}}$	10.41	***	145.74	15.51	***	651.49
S_{hf500}	21.02	***	294.33	17.88	***	750.8
S_{hf1k}	17.81	***	249.29	12.78	***	536.6
S_{cog}	2.53	*	35.36	9.28	***	389.94
S_{slope}	10.38	***	145.26	9.36	***	393.19
$F1_{\mathrm{M}}$	11.29	***	158.07	11.44	***	480.38
$F1_{\mathrm{SD}}$	7.72	***	108.13	11.18	***	469.4
$F1_{bw,\mathrm{M}}$	11.68	***	163.48	19.26	***	808.97
$F2_{\mathrm{M}}$	3.18	***	44.58	6.59	***	276.98
$F2_{\mathrm{SD}}$	11.01	***	154.15	3.54	***	148.53
$F2_{bw,\mathrm{M}}$	7.58	***	106.06	4.32	***	181.37
$F1_{prec,\mathrm{M}}$	3.37	***	47.21	4.79	***	201.36
$F2_{prec,\mathrm{M}}$	2.15		30.07	2.23	***	93.64
HNR_{M}	15.19	***	212.67	7.81	***	328.01
$R_{x,\mathrm{M}}$	3.19	***	44.59	2.69	***	112.98
$F0_{jitt,\mathrm{M}}$	10.29	***	144.02	4.77	***	200.33
$L_{A,shi\mathrm{m}}$	6.56	***	91.83	5.27	***	221.2

5 Subjective Evaluation of the Corpus

5.1 Background and Objectives

The encoding process has been briefly examined in Chapter 4. Stimuli, acoustics cues, and the trends between them have been obtained, but that's only one side of the story. Recall the model of Brunskwikian lens model of emotion communication from Chapter 2: at the moment only the objective distal cues have been obtained [6]. The assumption up to this point is that these distal cues perfectly correlate with the state of the encoder. It is possible that the stimuli lack ecological validity, and it is also equally possible that the acoustic trends have been confounded by the transmission of information throughout the digital sequence of processing. It may also be the case that the right set of acoustic features has not been analyzed. Studying the decoding process can help direct and validate acoustic analysis [55]. Decoding studies aim to assess many qualities of the emotion content in speech such as the perceived emotional intensity, the authenticity of the portrayal, clarity, and purity. By and large, the vast majority of decoding studies use a set of listeners to determine which stimulus is most easily discriminated or recognized, in addition to various qualities that DSP cannot otherwise quantify [3] [22].

5.1.1 Present Goals

The main goals of this study were to:

1. Assess how well the encoded emotions in the corpus match the perception of emotion in a recognition task.
2. Assess the perception of emotional intensity, and determine what acoustical cues correlate with the perception of emotional intensity.
3. Assess how presentation of emotional stimuli in and out of context affects the perception of emotion.

To meet these goals, this test was divided into three main tasks: a recognition task, an intensity task, and an open response composition task.

Recognition

The primary aim of the recognition task is to determine whether the intended emotion of the encoders (talkers) matches the general attribution of decoders (participants). A total of 30 participants were recruited to investigate how well each recorded stimulus performs in a recognition task. High recognition rates were expected for this set of stimuli given the small list of four "basic" or modal emotions portrayed (anger, happiness, sadness, and no expression). It is also possible that nearly every treatment (e.g. encoder, script, emotion) varies independently of the classification response. In addition to assessing a general indication of agreement or disagreement, the relative magnitude of changes in the classification performance will be examined. Measuring the duration of time that a decoder takes to perform the classification task will be particularly helpful in the case of small variance in the recognition rates. Differences in response times would likely indicate a change in task difficulty, and these values could potentially serve as continuous regressors in future analyses.

Intensity

Although recognition rates and perceived emotional intensity may certainly correlate the qualitative perception of the strength of the emotional portrayal will be considered as a separate subjective variable of interest. The reason for this separation partly stems from the possibility of confusing that which promotes a strong feeling (e.g. an emotional response) and what may simply aid in the identification of an acted emotion. It was therefore of interest to obtain intensity ratings from the decoders as a separate task. Scherer (2003) notes that decoding studies often report recognition rates while their methods actually test discrimination from a set of fixed responses. This has been acknowledged and an effort to account for forced-choice bias from the discrimination task will be made.

Contextual Significance

The stimuli were encoded originally in clusters of four chronologically ordered utterances. Each successive utterance follows fairly logically from its predecessor in the form of a "halfalogue", or one side of an over-heard phone conversation. Randomizing the order of utterance presentation eliminates the construction of a meaningful storyline and the cumulative connection to the conversation is minimized. Presenting the utterances that make up each script in the order that they were written and encoded offers way of comparing how contextual synthesis affects the decoding process.

5.2 Methods

5.2.1 Selection of stimuli

The stimuli were selected entirely from the corpus of speech recorded for PEEP. Because this study was technically separate from PEEP, only the participants from PEEP that had provided consent to share their stimuli on the Databrary website could be selected [71], which was 42 out of the 55 original number of mothers recorded All of the 42 participants were native English speakers and they each generated four prosodic versions (3 emotional, 1 no expression) of eight scripts, with four utterances per script (42 X 4 X 8 X 4 = 5376 possible stimuli). Ideally, all of 5376 utterances, or even 1344 script clusters would be played back to a large pool of listeners, but the limits of human focus and exhaustion demanded some form of stimulus sampling. The target testing time was 1.5 hours maximum to avoid participant fatigue and also keep the testing to a single session per participant. Of this total time, a maximum of approximate 60 minutes was set for evaluating the stimuli with the rest of the time for paperwork.

As it was not known yet which prosody would be perceived, all three emotions (anger, happiness, sadness) and non-expressive versions were selected. This left a total of 336 stimuli for each prosody to choose from:

$$\frac{\text{5376 utterances}}{\text{corpus}} * \frac{\text{1 script}}{\text{4 utterances} * \text{4 prosodies}} = \frac{\text{336 scripts \& encoders}}{\text{1 prosody}} \qquad (\,5.1\,)$$

At 10 seconds each in length, this leaves 56 minutes of listening per prosody. If the target testing time were to be less than 1.5 hours per listener, a selection of one or both of these remaining pools needed to be made. To make claims regarding the effect of language and context, at least two scripts were necessary. If these claims were to be generalized or if encoder abnormalities accounted for, at least two encoders would be needed as well.

Further practical measures regarding the testing time influenced the process of stimuli selection. Because response times were being collected for the discrimination task, it was determined that the subjective intensity ratings should be collected as a separate line of successive questions rather than combine the two. The composition task aimed at investigating contextual effects of the utterances as encoded with an open response format. It was expected that participants would take significantly longer with this task due to its format. The number of stimuli selected for this last task would be proportional to the discrimination task by a factor of one quarter, but the total length of stimuli presented would be no less. Leaving a half hour for questionnaires, consent forms, and tutorials before each task, a more realistic target would be roughly twenty minutes for the discrimination task, twenty minutes for the intensity task, and twenty minutes for the composition task. With added time between questions and delays from response times, finding an optimal set of encoders and scripts to represent the corpus was still necessary.

A combination of acoustic analysis and piloting the study was employed to determine the appropriate combinations of encoders and scripts. First, the corpus was analyzed and encoders ranked in terms of the magnitude of variance in their LLDs between prosodies. The matrix of LLD values was first normalized across all encoders for each cue to eliminate the effects of the range of LLD values skewing the analysis. The results of this analysis are illustrated in Figure 5.1. The two encoders with the greatest summed variance are 024, followed by 015. These two are markedly higher than the rest of the encoders by almost two standard deviations, and were selected as candidates for inclusion in the study.

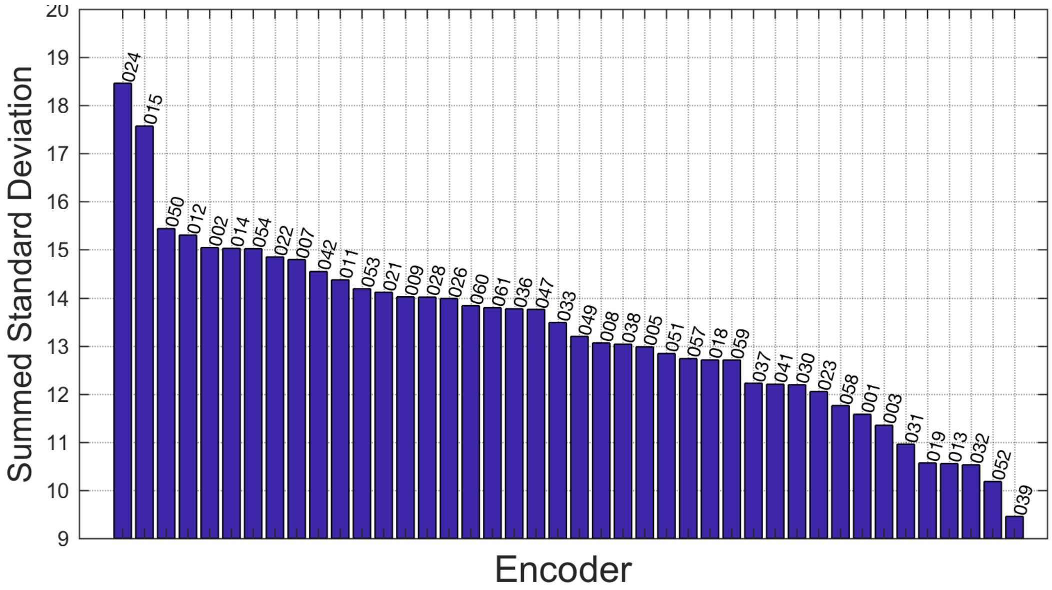


Figure 5.1 The summed standard deviation of the LLDs between emotion categories for each encoder, sorted in descending order, i.e. from encoders who produced recordings with the largest amount of variance between prosodies to those who had the least amount of variance.

Next, scripts were analyzed for their effects on the variance of the LLDs for each encoder. For each script, the standard deviation over prosody across all encoders was ranked in similar fashion the encoder analysis. Figure 5.2 gives the ranking of each script's total contribution to the variance across emotion. "chk-a" and "din-a" have the greatest total variance across emotion for all encoders. These two scripts were contributed to the pool of candidates for subjective validation.

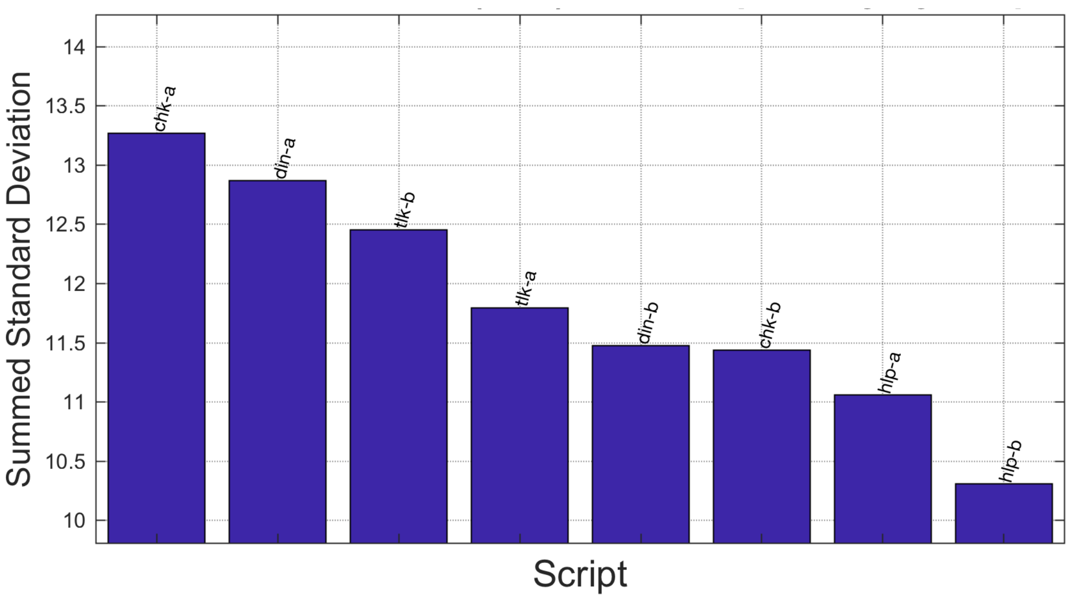


Figure 5.2 This is a bar plot of the summed standard deviation of LLD values over emotion across all encoders, ranked by script.

It should be noted that in comparison to the encoder rankings, the range of summed variance for the eight scripts are all within three standard deviations. This result suggests that all of the scripts contribute relatively evenly to the encoding process, any therefore differences across script may not matter after all. The results of the one-way ANOVA given in Table 5.1 for the cues with the highest effect sizes suggest that script may not have as great of an impact on the perception of emotion. Regarding the largest effect sizes for emotion, $F0_{\mathrm{M}}$ has an effect size $\eta^2 = 0.61$, whereas script effects on $F0_{\mathrm{M}}$ are more than an order of magnitude less ($\eta^2 = 0.03$, $p < 0.0001$). The same is true for $F0_{\mathrm{SD}}$ ($\eta^2 = 0.56$, $p < 0.0001$) and $L_{AI,\mathrm{M}}$ ($\eta^2 = 0.52$, $p < 0.0001$) in comparison to the script effects ($\eta^2 = 0.04$, $p < 0.0001$) and ($\eta^2 = 0.002$, $p = 0.166$).

Table 5.1 F-statistics and effect sizes for linear one way ANOVA tests for the effects of emotion and script for the top 5 effects for emotion. Note the relatively low F-statistics and effect sizes for script in comparison to emotion.

	Emotion			Script		
cue	*F*	η^2	*p*	*F*	η^2	*p*
$L_{AI,\mathrm{M}}$	2818.06	0.52	***	1.49	0.002	0.166
$F0_{\mathrm{M}}$	4083.77	0.61	***	23.51	0.03	***
$F0_{\mathrm{SD}}$	3393.09	0.56	***	33.31	0.04	***
S_{hf500}	1010.89	0.28	***	79.96	0.10	***
S_{hf1k}	1060.69	0.29	***	62.53	0.08	***
$F1_{\mathrm{M}}$	482.16	0.16	***	49.16	0.06	***

Regardless, the top three encoders were selected (024, 015, and 050) plus a sampling of encoders with lower acoustic scores. The top script "chk-a" along with a sample from the middle "din-b", "tlk-a" were selected for rounds of piloting with members of the research team. Scripts such as "din-a" and "tlk-b" were not considered due to the nature of their content. "din-a" features several commas and colloquial interjections that frequently corrected in the recording process:

"Oh, hi, it's you. When will you be home? Dinner won't be ready then. Okay, I'll fix dinner."

Script "tlk-b" was not considered as it features a list of words that also had to be corrected during the recording process:

"Can you talk now? About lots of things. Money, the weekend. Okay, we won't talk now."

The verb "correction" is not intended to imply synthetic modification. The pauses introduced by the presence of commas as well as listed words often increased the total length of the utterance beyond what was acceptable for ten second presentation for PEEP. During these occasions, the speaker would be asked to ignore the commas and speak more quickly than previously done, thereby artificially modifying the encoder's original expressive style.

Recognition rates for the pilot tests were not taken, as the group of participants were fairly familiar with the kind of stimuli being presented and the acoustic cues examined. The goal of the piloting was to determine whether the encoders selected by the acoustic analysis were in fact rated highly before a larger study was conducted. Also of interest was the subjective perception of variance in the collection of expressions. The stimuli were presented over a set of Sony MDR-7506 headphones, and the listeners were asked to rate how intense they felt the emotion being expressed was on a scale from 0 (no intensity at all) to 9 (extremely high intensity). The results of the piloting largely solidified the results of the acoustic analysis. Encoders 015 and 024 on average had the greatest intensity ratings for the expressed emotions. When averaged over encoder, the intensity ratings deviated by less than 10%, which further substantiates the claim that script may not in fact be of consequence. Although the ratings indicated that “chk-a” was given low ratings overall, comments from the pilot participants indicated that the variety in expression for this particular script was particularly low. Total consistency with the intensity ratings would not necessarily benefit future correlation analyses, so a range of intensities was permissible as long as the perception of variety amongst the sample population was there.

Results from a final round of piloting with a subset of the encoders (015, 024, and 059) and scripts (“din-b” and “tlk-a”) exhibited an acceptable range and average ratings (greater than the 4.5/9) for both script and encoder. Three encoders, two scripts, and four prosodies produces twenty-four stimuli for the composition task, and ninety-six stimuli (24 x 4 utterances) for the recognition and intensity tasks.

5.2.2 Participant information

Advertisements for participation were entirely conducted on Penn State’s StudyFinder website and through department listservs [72]. Participants were required to have hearing thresholds of 15 dB HL or lower for the 250 Hz to 8 kHz octave bands which was tested using a Welch Allyn AM282 manual audiometer at the time of the appointment. All participants were fluent in English as a first language, were at least 18 years of age at the time of the testing, and none reported experiencing or having been diagnosed with emotion-related psychological disorders (e.g. autism, depression etc.).

The diagnostic analysis of nonverbal accuracy II (DANVA) was administered as a final test for inclusion in the study [45]. This exam was used in part to ensure that participants all met a baseline for emotion discrimination, and to enable future comparisons between the results of this test and those of the PEEP study. DANVA is divided into a visual and auditory task. The visual task (called “Adult Faces”) presents a photograph of an adult making a certain facial expression. The participant is instructed to choose the emotion (anger, happiness, sadness, fear, surprise) that they think the subject in the photo is expressing, and a new photograph appears. The auditory task (called “Adult Paralanguage”) runs in a similar manner; participants are presented a recording of a person speaking in with a specific affect and asked to identify from a list of five categories which emotion they believe is being expressed. There are 24 questions per task, and each task takes approximately three minutes to complete. Inclusion was based on achieving correct response rates (p) significantly higher than chance (p_0). This was determined through a Z test for proportions with $H_0 : p = p_0$ vs. $H_a: p > p_0$ at alpha = 0.05. This meant that participants must get at least eleven out of the twenty-four questions correct (no more than 13 errors). All participants passed at rates higher than chance. Participants that met the inclusion requirements were awarded $10 compensation, and were entered into a raffle for $50 with a chance of winning roughly of 1/35. A minimum compensation of $5 was provided if the participant did not meet the hearing sensitivity requirements.

In similar form to Lima and Castro (2011), the Autism Spectrum Quotient (ASQ) was administered as a test for possible presence of social or communication abnormalities typically exhibited by individuals on the autism spectrum [73]. This test was fairly simple to administer, and comprises 50 questions regarding personal preference for social function and related topics. The ASQ is scored in points out of 50, and the higher the score the more likely the participant may exhibit autism-related traits. Results differed slightly for males than females ($\bar{x} = 17.4$, $\sigma = 6.58$ vs. $\bar{x} = 12.13$, $\sigma = 5.74$), with six males and two females in the intermediate category (+20/50). No participants scored high enough to be considered having high functioning autism (HFA).

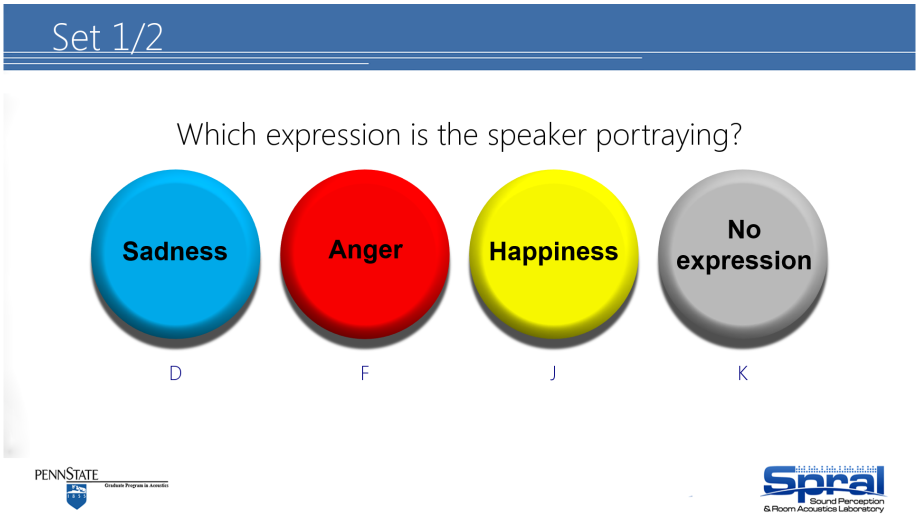
Thirty people (n=30, 15 male, 15 female) with an average age of 24.4 years from the State College area participated in the study. Of these 30, ten reported having some acting experience, and 26 reported having some form of musical training. Participant occupation included 9 graduate students, 12 undergraduate students, and 9 professionals working in the State College area. One additional person volunteered for the study, but didn't met the hearing requirements. All participation was voluntary, and written consent was obtained at the beginning of each appointment and the testing followed the approved Penn State Human Subjects IRB protocol #6980.

Participants were given three main tasks during the experiment. The first and the second tasks involved collecting identification rates and subjective perception of emotional intensity from the larger pool of short utterances (the individual sentences within the two scripts). The third task gave the participants the freedom to compose the relative presence of emotions in the 10-second stimuli comprised of the four utterances in context. Because of the vast differences in stimuli, the first two tasks were labeled Set 1a and Set 1b respectively, and the third task labeled Set 2. Participants were given a short tutorial before both Set 1 and Set 2 that provided details about the test format, the types of questions, and the preside definitions of possible responses. In the tutorial, participants were told that at the start of each question a recording of speech would automatically begin playing and would be played once.

5.2.3 Set 1a – Emotion recognition task and response times

The purpose of Set 1a was to test whether participants could identify the intended emotion of the encoder from a selection of possible responses. Participants were instructed to "identify as quickly as possible which emotion the speaker is portraying. The speaker may not be expressing an emotion, and you may choose to select 'No Expression.' You can only submit one answer." All 96 utterances were included in this set and were presented separately.

Participants submitted their responses by pressing the letter on the keyboard corresponding to the letter beneath the button on the screen. Figure 5.3 is a screenshot of user interface during the presentation of a stimulus. The letters "D", "F", "J", and "K", are positioned comfortably in the middle of the keyboard, and the surfaces of the keys for "F" and "J" feature small tactile locators for the left and right index fingers respectively. Participants were instructed to keep their left and right hands over these four letters at all times during Set 1a. To mitigate possible effects of how the expression labels were ordered relative to one another, each participant received a new randomized order (e.g. Happiness | No expression | Sadness | Anger) that would remain in that order for the rest of Set 1a. Anticipation bias due to timing of the question presentation was also accounted for by pseudo-randomizing the time between questions from 200 ms to 600 ms. The average delay between questions was 400 ms, and the total duration of the set fixed to 13 minutes and 12 seconds. Response times were also collected, measured from the moment the stimuli began to play to the time the response was submitted. If a response was not submitted within 4 seconds (the preset duration of each stimulus) + 400 ms, the response was labeled as timed out and the next question presented.


Set 1/2

Which expression is the speaker portraying?

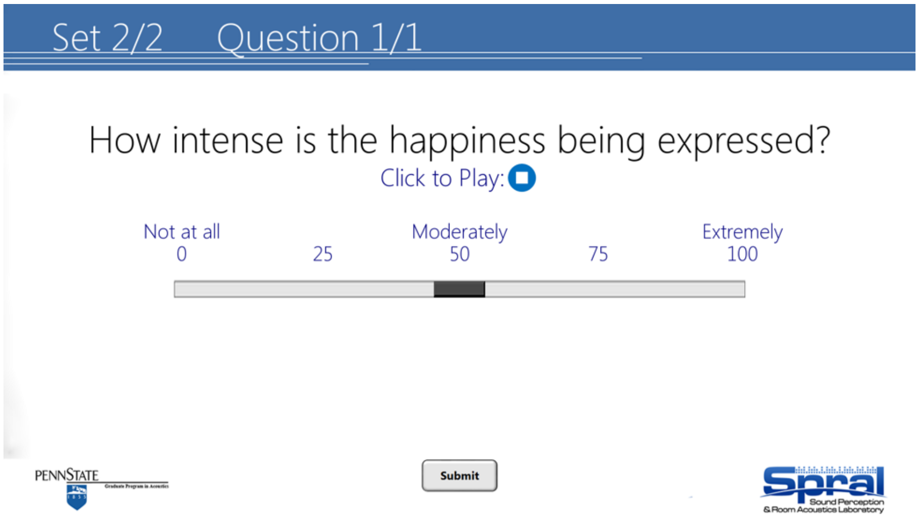
Sadness
Anger
Happiness
No expression
D
F
J
K
PENNSTATE
Graduate Program in Acoustics
Spral
Sound Perception & Room Acoustics Laboratory

Figure 5.3 Graphical user interface for Set 1a – the recognition task. The letters beneath each expression category correspond to letters on a standard "QWERTY" keyboard layout.

5.2.4 Set 1b – Emotional Intensity Ratings for the utterances

The purpose of Set 1b was to investigate how intensely participants perceived the emotions portrayed in the stimuli. Only excerpts from Set 1a that were identified as having a prosody (anger, happy, and/or sad), whether correct or not, were included in this set. This type of design allowed for separation between question sets while minimizing the gross time of engagement in the stimuli. It was decided that there was little to be gained in the inquiry of a neutral expression's intensity when it was previously identified as expressionless. Note that in Figure 5.4, only one emotion is offered as an option for intensity judgement. This classification was also predetermined by the responses of the previous set (Set 1a). In summary, stimuli identified as emotional in Set 1a would later be presented in a context of their previous response in Set 1b. Participants were not informed of this design as it may have affected their responses for Set 1a.

Participants were told that at the beginning of each question a recording of speech would begin playing. They were told that they could repeat the recording as many times as desired. Participants were instructed to rate on a scale of 0-100 the intensity of the speaker's portrayal of the emotion label offered. The full script of the rating scale definitions are as follows "if you think that the speaker is expressing very intense emotion, you should give a rating of 100 on the emotion intensity scale. If, on the other hand, you think the speaker is expressing emotion with no intensity at all, you should give a rating of 0." After the presentation of each stimulus, a horizontal slider would appear on the screen as is shown in Figure 5.4, offering a continuous scale to select the desired intensity.



Set 2/2 Question 1/1

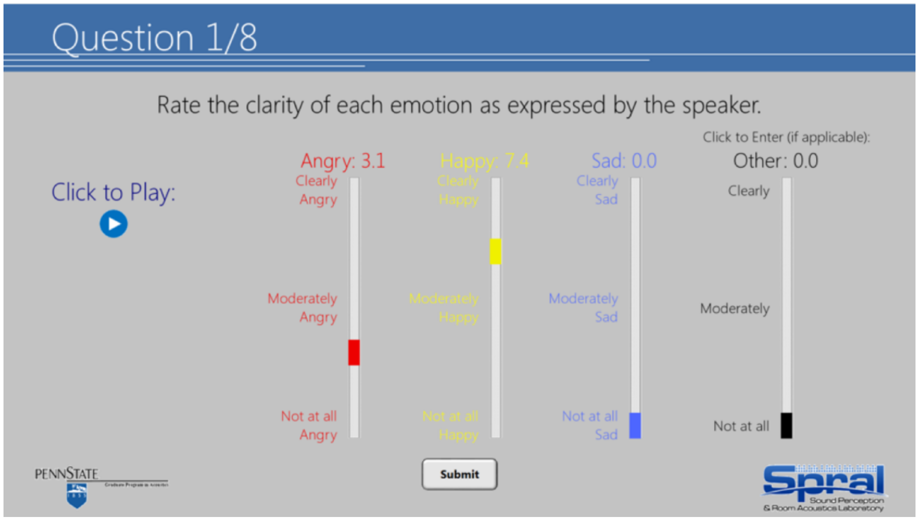
How intense is the happiness being expressed?

Figure 5.4 This is a screenshot of Set 1b after the stimulus had finished playing.

At the end of the tutorial, participants were given a practice set to familiarize themselves with the process for both Sets 1a and 1b. This practice set contained twenty-four utterances from a different set of encoders and scripts, and lasted less than four minutes. For Set 1a, all 96 utterances from each of the three encoders, four prosodies, and three scripts were presented. Participants were given the option to take a two-minute break between Set 1a and Set 1b, however none opted to take this small break.

5.2.5 Set 2 - Composition and Context of Full Scripts

For Set 2, participants were told that a speech recording much longer than the previous recordings would begin playing at the beginning of each question. Each stimulus had to play at least once to the very end before participants could submit a response, and each recording could be repeated indefinitely. The instructions were to "consider if the speaker is clearly expressing one or more emotions. If one or more emotions is detected, then rate how clearly the speaker is expressing each of those emotions on their respective scale." The scale ran from 0 (not at all e.g. Angry) to 100 (clearly e.g. Angry) as shown in Figure 5.5. They were specifically instructed to leave the sliders at 0 (the default position) if they felt that the speaker did not express that particular emotion whatsoever. If multiple emotions were detected, then each were to be evaluated as independent questions of presence. Participants were instructed to write in an emotion in the "other emotion" placeholder if they felt that the expressed emotion belonged to a different category that they could identify. The hope with this design was to mitigate potential bias caused by the forced choice design of Set 1a. In the case where no expression (neutral) was detected, participants were instructed to leave all of the sliders at 0.



Question 1/8

Rate the clarity of each emotion as expressed by the speaker.

Angry: 3.1

Happy: 7.4

Sad: 0.0

Other: 0.0

Figure 5.5 This is a screenshot of the training set for Set 1b.

Upon completion of all three tasks, participants were asked to submit evaluations of the testing procedure and impressions of the stimuli apart from the specific questions previously posed. Details on the procedural impressions are provided in Appendix H. The average test length across all participants as approximately 90 minutes.

5.2.6 Stimuli Playback

All participants were presented stimuli at a fixed amplitude of 72 dBA re: 20 µPa over a pair of Sony MDR-7506 studio monitor headphones. After each practice set, participants were asked if the levels were too loud or too quiet as a precaution. Gains on the sound card were left unchanged for all participants. Custom filters were created for the left and right ear using a Brüel and Kjær Type 4100 Sound Quality Head and Torso Simulator. The filters were 4096 point long finite impulse response (FIR) filters built with the MatLab command "fir2." To increase coherence and reduce noise, sine sweeps were averaged over twenty cycles, and their mean spectra frequency vectors and magnitudes fed to the function. The magnitude and frequency response of the filters are provided in Figure 5.6.

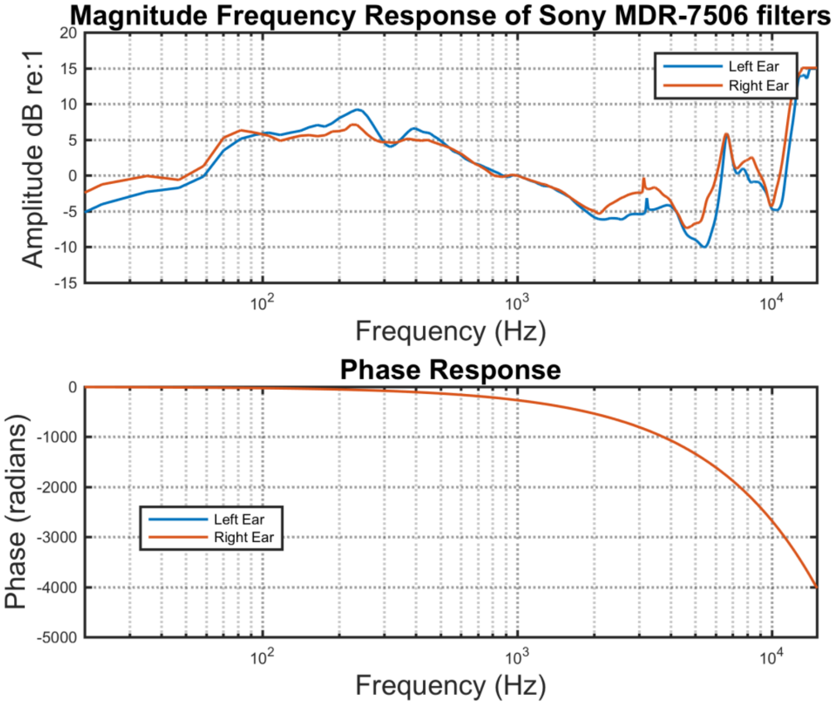


Figure 5.6 These are plots of the magnitude of the frequency response for the filters for each headphone ear (top) and the corresponding phase response (bottom).

5.3 Results

5.3.1 Set 1a - Recognition

Average rates for correct responses across prosodies for Set 1a were 92.0%. The rates broken down by prosody were 94.9% for anger, 93.2% for happiness, 92.4% for sadness and 87.5% for no expression. Response times on average were 1.9 seconds from the start of the stimulus to the moment of entry. By prosody they were: 1.783s for anger, 1.778s for happiness, 1.953s for sadness, and 2.1s for no expression. These rates are all significantly higher than the 60% rate average reported by Scherer (2003) for over thirty studies of emotion recognition. He notes that the 60% correct response rate is more than five times random chance, which comes out to nearly 10%. This implies that the 30 studies on average considered ten emotions (1/10 random chance) which is much more than double the number of expressions considered here. Still, the baseline correct response rate was 87.5% for no expression- a healthy four times random chance. It is safe to say that the listeners were significantly capable of discriminating the intended expression from the four options given. Effects of emotion on both error rates and response times are given after the summary.

To understand what may have affected recognition rates or response times, a one-way repeated measures mixed model ANOVA was performed on the subset of data not including the intensity ratings for both the identification rates and response times. With the dependent variable set to recognition rate, the main effects of encoder (*F* = 18.77, p < 0.0001), and rated emotion (*F* = 23.66, p < 0.0001) were both significant as is shown in Table 5.2.All tests hypotheses tests were conducted at an alpha of 0.05 with a Bonferroni correction that brings the effective alpha to 0.0167. Neither script or utterance effects had any significant influence on the discrimination rates. Interactions between the main effects did cause significant differences in the discrimination rates. Encoder x rated emotion (*F* = 35.11, p < 0.0001) had the second highest F-statistic of the ANOVA list. The interactions between script and rated emotion nor the interactions between script x utterance x rated emotion were significant to the identification rates.

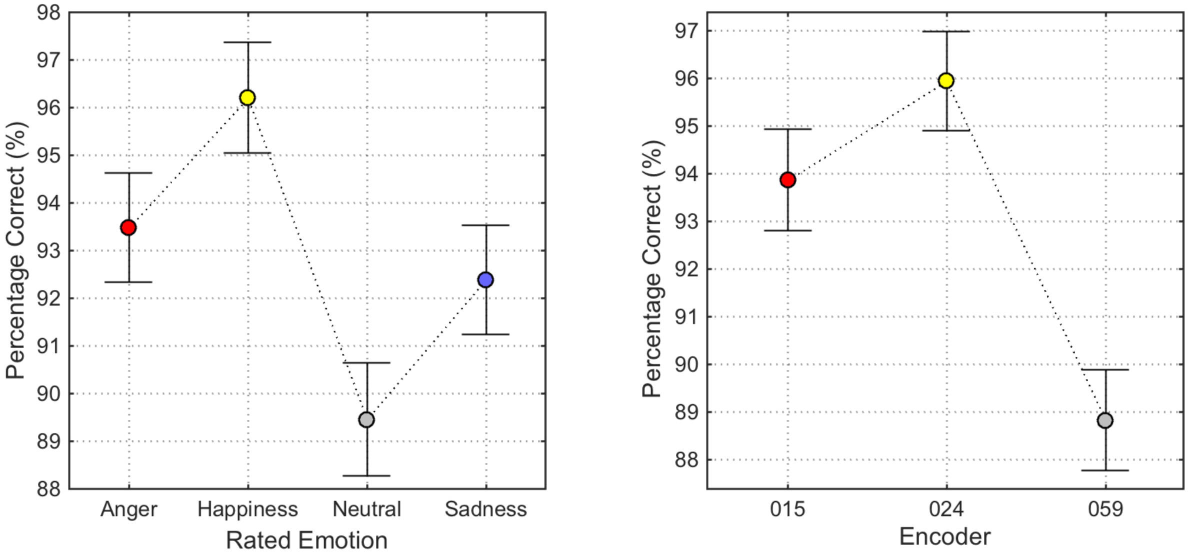
The main effect of script was also not found have a significant effect on the response times, as is shown in the fourth row of Table 5.2. The greatest main effect was the rated emotion (*F* = 68.43, p < 0.0001), followed by encoder. Interestingly enough, the interaction of script X rated emotion (*F* = 4.93, p < 0.0001) and script X utterance (*F* = 12.65, p < 0.0001) were both significant. Overall, the ANOVA results for response time gave significant effects for 10 out of the 15 main and interaction effects, while identification rates only showed significance for 6 out of the 15.

Table 5.2 F-statistic and p-values from the one-way repeated measures ANOVA for the main effects and interactions for discrimination rates.

Repeated Measures ANOVA table for Correct Responses

Effect	Num DF	Den DF	χ^2	*F*	Pr > χ^2	Pr > *F*
Encoder (E)	2	2755	37.54	**18.77**	<.0001	<.0001
Rated_Emotion (RE)	3	2761	23.66	**7.89**	<.0001	<.0001
E x RE	6	2760	35.11	**5.85**	<.0001	<.0001
Script (S)	1	2756	0.42	0.42	0.5188	0.5188
E x S	2	2755	6.32	3.16	0.0425	0.0426
S x RE	3	2759	2.66	0.89	0.4467	0.4469
E x S x RE	6	2758	5.87	0.98	0.4384	0.4386
Utterance (U)	3	2755	2.37	0.79	0.4992	0.4994
E x U	6	2755	26.23	**4.37**	0.0002	0.0002
U x RE	9	2759	45.74	**5.08**	<.0001	<.0001
E x U x RE	18	2759	50.29	**2.79**	<.0001	<.0001
S x U	3	2755	9.27	3.09	0.0259	0.0261
E x S x U	6	2755	5.68	0.95	0.4596	0.4598
S x U x RE	9	2759	10.27	1.14	0.3287	0.3292
E x S x U x RE	18	2759	32.4	1.8	0.0197	0.0202

To determine how each of these variables affected the recognition rates, a least-squares means comparison was performed on the recognition rates. The results of these analyses are illustrated in Figure 5.7 (a) and (b). Recognition rates given in Figure 5.7 (a) for the effect of rated emotion indicate significant differences between many emotion categories. Happiness had the highest overall recognition rate ($\bar{x}$ = 96.20%, σ = 1.16%), followed by anger ($\bar{x}$ = 93.47%, σ = 1.14%), sadness ($\bar{x}$ = 92.38%, σ = 1.15%), and neutral ($\bar{x}$ = 89.45%, σ = 1.18%). Recognition rates between happiness and anger were not significantly different ($p = 0.19$). Significant differences were also found between happiness and sadness ($p < 0.05$), happiness and neutral ($p < 0.0001$), and anger and neutral ($p < 0.05$). Overall these results indicate that recognition rates were significantly affected by the rated emotion. The limited significance (and overall high recognition rates) found between just happiness and sadness for the emotional portrayals indicates that these differences may be limited in extent.



(a) (b)

Figure 5.7 (a) Recognition rates as a function of the effect of rated emotion and (b) recognition rates as a function of the effect of encoder.

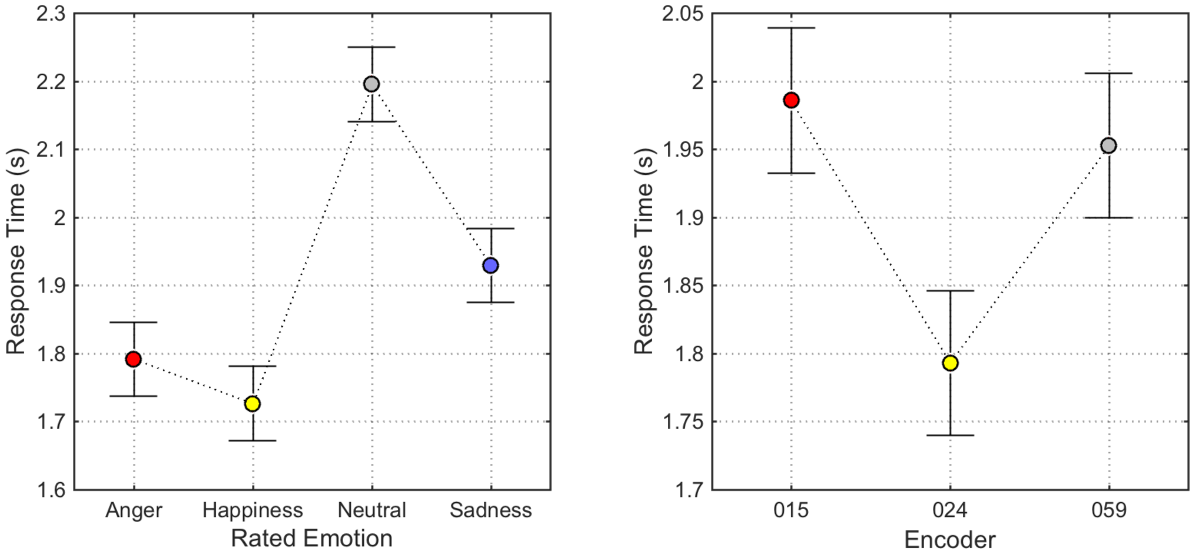
Regarding the effect of encoder on the recognition rates in Figure 5.7 (b), encoder 059 ($\bar{x} = 88.82\%$, $\sigma = 1.05\%$) had significantly lower rates than both encoders 015 ($\bar{x} = 93.86\%$, $\sigma = 1.06\%$) and 024 ($\bar{x} = 95.94\%$, $\sigma = 1.04\%$) at $p < 0.0001$. Portrayals from encoders 015 and 024 were rated with statistically equivalent accuracy. These results affirm the acoustical analysis that was conducted in the stimuli selection; 024 had the greatest summed variance between emotion portrayals, followed closely by 015, and from a much greater distance by 059.

Table 5.3 Results for the repeated measures ANOVA with response time as a dependent variable.

Response Time

Effect	**Num DF**	**Den DF**	χ^2	***F***	**Pr >** χ^2	**Pr >** ***F***
Encoder (E)	2	2755	47.1	**23.55**	<.0001	<.0001
Rated_Emotion (RE)	3	2756	205.3	**68.43**	<.0001	<.0001
E x RE	6	2756	15.14	2.52	0.0192	0.0194
Script (S)	1	2755	0.36	0.36	0.5475	0.5476
E x S	2	2755	1.56	0.78	0.4578	0.4579
S x RE	3	2756	14.8	4.93	0.002	0.002
E x S x RE	6	2756	14.83	2.47	0.0216	0.0218
Utterance (U)	3	2755	10.99	3.66	0.0118	0.0119
E x U	6	2755	45.5	**7.58**	<.0001	<.0001
U x RE	9	2756	36.68	**4.08**	<.0001	<.0001
E x U x RE	18	2756	58.78	**3.27**	<.0001	<.0001
S x U	3	2755	37.95	**12.65**	<.0001	<.0001
E x S x U	6	2755	17.74	2.96	0.0069	0.007
S x U x RE	9	2756	16.96	1.88	0.0494	0.0499
E x S x U x RE	18	2756	41.93	2.33	0.0011	0.0012

To determine the nature of the relationship between response times and the effects, a least-squares means comparison was performed. The results of these analyses are illustrated in Figure 5.8. Looking to the effect of response time (rT) in Figure 5.8 (a), neutral had the slowest response time ($\bar{x}$ = 2.20 s, σ = 0.055 s), followed by sadness ($\bar{x}$ = 1.93 s, σ = 0.054 s), and anger ($\bar{x}$ =1.79 s , σ = 0.054 s). These differences were all significant at $p < 0.001$, except for anger andhappiness which were not statistically different. Regarding the effect of encoder in Figure 5.8 (b), encoder 024 was the most quickly rated overall ($\bar{x}$ = 1.79 s, σ = 0.052 s), followed by encoder 059 ($\bar{x}$ = 1.95 s, σ = 0.053 s) and 015 ($\bar{x}$ = 1.98 s, σ = 0.054 s). Response times for 024 were significantly lower than both 015 and 059 at $p < 0.0001$, but rTs for 015 and 059 were statistically equivalent to each other. It is interesting to note that while recognition rates for 015 and 024 were not statistically different, their respective response times were significantly different (10%) from one another. This might be explained by average differences in utterance lengths. Future work would benefit from a multivariate approach to quantifying these measured differences in response times using a range of acoustic variables.



(a) (b)

Figure 5.8 (a) Average response times as a function of the effect of rated emotion and (b) average response times as a function of the effect of encoder.

5.3.2 Set 1b- Intensity Ratings

The average intensity rating across all responses was 65.4 out of 100. Anger had the greatest intensity rating at 68.7, followed by sadness (66.5), and happiness (65.35). No expression was occasionally labeled as emotional (82 times out of 2880 responses = 2.85%) but was subsequently rated on average as having the lowest overall intensity of 31.5 out of 100. Pearson's linear correlation coefficient for response time (RT) and correct responses was significant ($r = -0.16$, $p < 0.001$) as well as RT and intensity rating (I) ($r = -0.09$, $p < 0.001$). Intensity ratings were most significant and highly correlated with correct responses ($r = 0.36$, $p < 0.001$). The results indicate that correctly discriminated stimuli were significantly more likely to be given higher intensity ratings. Furthermore, the inverse relationship between RT and both correct responses and intensity ratings indicates that stimuli that received faster responses (lower RT) were also rated with significantly greater accuracy and higher intensity.

To understand what may have affected the perceived emotional intensity of the stimuli, a one-way repeated measures mixed model ANOVA was performed on the data with intensity ratings as the dependent variable. Intensity ratings were the most significantly affected by the main effects and their interactions (see the bolded F-statistics in Table 5.4. Encoder was the greatest main effect ($F = 96.95$ $p < 0.0001$) followed by script ($F = 29.46$, $p < 0.0001$). No significance was found for the main effect of utterance.

Table 5.4 Results from the repeated measures ANOVA with intensity ratings as the dependent variable. F-statistics are given in the column third to the right, and their corresponding p-values are provided in the right-most column.

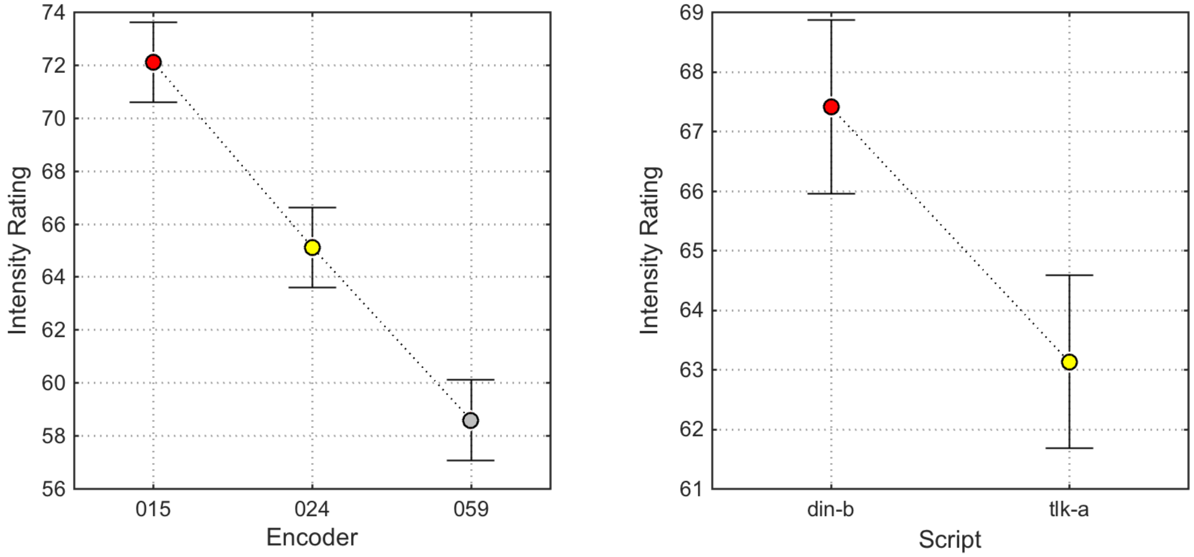
Intensity Ratings

Effect	Num DF	Den DF	χ^2	*F*	Pr > χ^2	Pr > *F*
Encoder (E)	2	2063	193.89	**96.95**	<.0001	<.0001
Rated_Emotion (RE)	2	2064	3.27	1.64	0.195	0.1952
E x RE	4	2064	188.23	**47.06**	<.0001	<.0001
Script (S)	1	2063	29.46	**29.46**	<.0001	<.0001
E x S	2	2063	2.21	1.11	0.3306	0.3308
S x RE	2	2063	18.49	**9.24**	<.0001	0.0001
E x S x RE	4	2063	59.26	**14.81**	<.0001	<.0001
Utterance (U)	3	2063	1.1	0.37	0.7768	0.7768
E x U	6	2063	96.97	**16.16**	<.0001	<.0001
U x RE	6	2064	15.42	2.57	0.0172	0.0175
E x U x RE	12	2064	41.84	**3.49**	<.0001	<.0001
S x U	3	2063	11.6	3.87	0.0089	0.009
E x S x U	6	2063	28.52	**4.75**	<.0001	<.0001
S x U x RE	6	2064	22.58	**3.76**	0.0009	0.001
E x S x U x RE	12	2064	58.29	**4.86**	<.0001	<.0001

To determine how each effect changed the intensity ratings, a least-squares means comparison was performed on the intensity ratings. The results of these analyses are illustrated in Figure 5.9. The effect of rated emotion on the intensity rating was not analyzed here because the repeated measures ANOVA did not show this effect to be significant. Regarding the effect of encoder on the intensity ratings in Figure 5.9 (a), the encoder performance followed this ascending order: 059 < 024 < 015 ($p < 0.0001$). Significant differences were found for the effect of script (see Figure 5.9 (b)), with din-b ($\bar{x} = 67.40$, $\sigma = 1.46$) having a higher overall intensity rating than tlk-a ($\bar{x} = 64.12$, $\sigma = 1.46$) at $p < 0.0001$.

These results, to a certain extent, affirm some of the acoustic rankings performed for the stimuli selection. Both 015 and 024 had higher summed acoustic variance between intended emotion portrayals than 059. While there were no significant differences between recognition rates for 015 and 024, here there are large differences in intensity ratings between the two. Also of note is the large difference between intensity ratings for script that was not found in previous subjective analyses. Acoustically, tlk-a had a greater summed variance between emotion portrayals than din-b – a result that opposes the trend in intensity ratings here.

It is possible that these differences could be explained by biases introduced by word choice or societal rules and code usage for each listener. For example, some words in din-b may favor greater acoustic differentiation between emotion categories compared to tlk-a. Scripts containing questions could further structure the acoustic markers for emotion, which is the case in tlk-a. Future work would also benefit from an utterance-level approach to quantifying these differences. Given that the first utterance in tlk-a has the question "Oh, you're tired?", one might expect this utterance to have the least amount of acoustic differentiation and possibly intensity ratings.



(a) (b)

Figure 5.9 (a) Average intensity ratings as a function of the effect of encoder and (b) average intensity ratings as a function of the effect of script.

Figure 5.10 illustrates how the intensity ratings changed as a function of the interaction between rated emotion and script. Within din-b, rated emotion was not significantly different across emotion Tlk-a, however, resulted in significantly lower intensity ratings between sadness and anger than for din-b. Comparing within emotion category, anger was not significantly affected by changing the script, while both happiness and sadness intensity ratings were significantly lower for tlk-a compared to din-b.

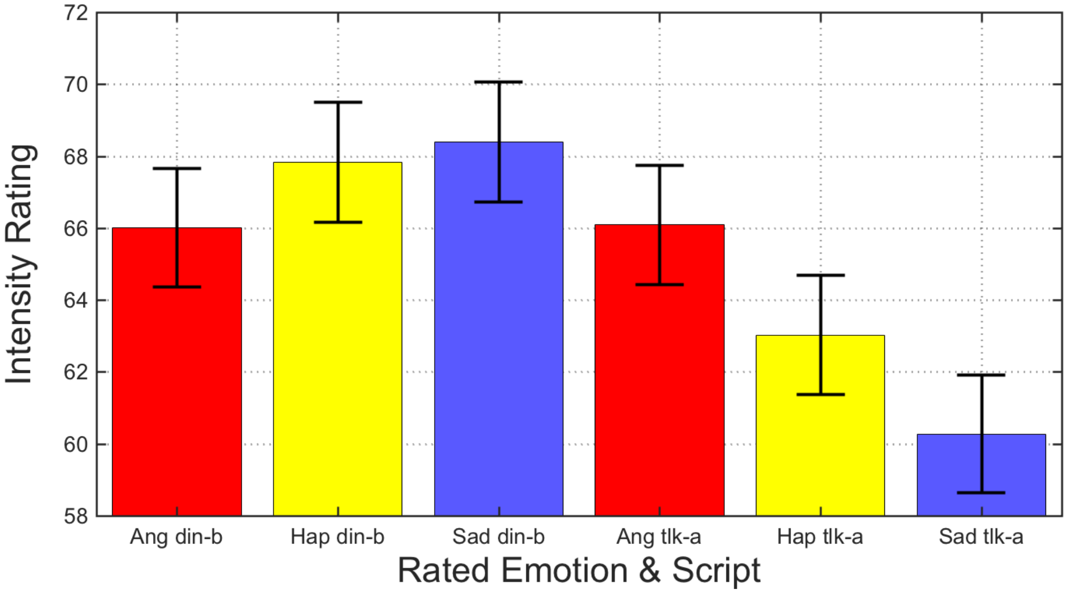


Figure 5.10 Average intensity ratings and standard errors as a function of the interaction between rated emotion and script.

5.3.3 Set 2 – Composition and Clarity

Clarity ratings from Set 2 differed slightly from what was observed in Set 1b intensity ratings. Happiness had the highest overall clarity ($\bar{x}$ = 8.1 , σ = 2.28), followed by anger ($\bar{x}$ = 7.77 , σ = 2.41), and followed by sadness ($\bar{x}$ = 6.86 σ = 2.85). Not only are the means greatest for happiness, but the standard deviations are smallest too. For the write-in category, the most frequently submitted responses were “annoyed” (n = 15), “tired” (n = 15), “depressed” (n = 5), “excited” (n = 5), “bored” (n = 4), “fear” (n = 4), “overwhelmed” (n = 4), “disgust” (n = 3), “optimistic” (n = 3), and “sarcastic” (n = 3). Write-in responses for annoyed and tired received clarity ratings of ($\bar{x}$ = 6.4 , σ = 2.53) and ($\bar{x}$ = 7.67 , σ = 2.41), respectively. Eight out of these top ten greatest response counts were negatively focused, and two of them were more positive. Sarcasm is a layered expression of deceit with an underlying message usually of a negative connotation. As such, expressions that were considered happy at the surface level yet negative at the ground truth could feasibly include both happy and angry intended portrayals.

Comparison of the results of Set 1 and Set 2 required a targeted and clearly defined approach. For Set 1a, the independent variable extracted was the total number of times each stimulus was identified in one of the three emotion categories. This count was converted to a percent of the maximum count, which is 30. This method mirrors that of Lima & Castro (2011) and Banse & Scherer (1996). To compare to Set 2, within which 18 of the 24 stimuli were intended as carriers of emotion, values across each of the four utterances in Set 1a were averaged. Values for the two matricies were averaged across the 30 decoders and plotted by emotion label, encoder, and script in Figure 5.11. Pearson's linear correlation coefficient showed no significant linear trend across stimuli for average clarity values from Set 2 and percentages from Set 1 (r = 0.26, p = 0.29). t-tests for the mean indicated that on average, clarity of emotion in Set 2 ($\bar{x} = 0.75$) was lower than the percentages of Set 1a ($\bar{x} = 0.93$) with $p < 0.0001$. It is recognized that the transformation of information from the responses of one line of questions may not perfectly withstand semantic scrutiny; the goal of this analysis is to simply perform one possible comparison.

The results show that the ordered presentation of each utterance as originally created by the encoder weakened the perception of the overall emotional accuracy. Although each paired comparison results in higher responses from the out-of-context presentations, there are several places where the means are less than 5% from equal (see Figure 5.11 happiness 015 din-b, 059 din-b, sadness 015 din-b).

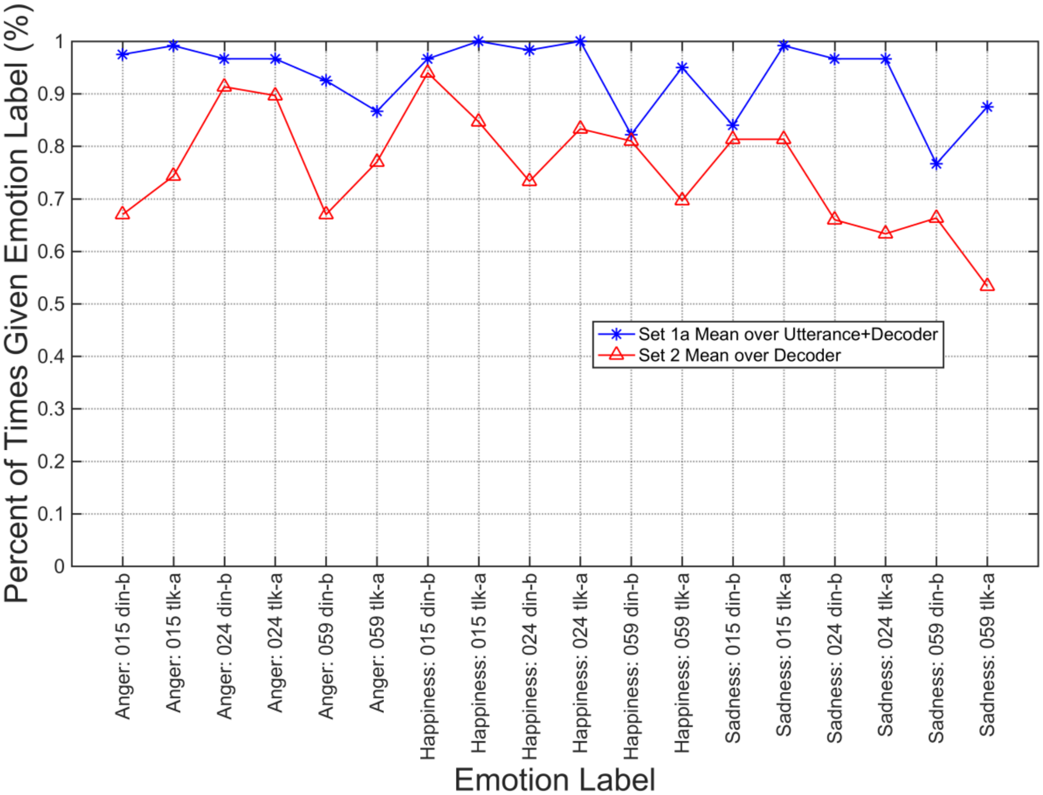
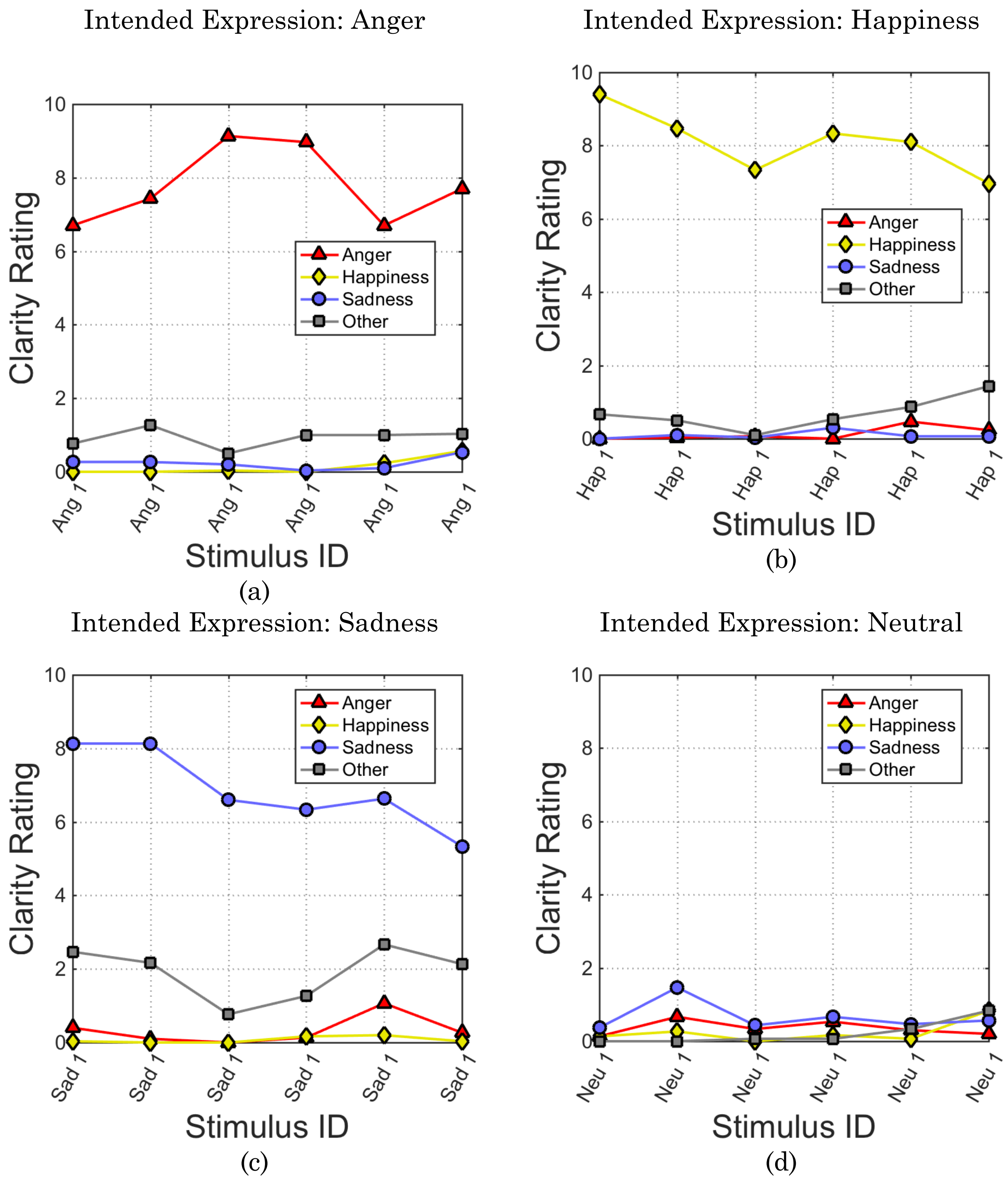


Figure 5.11 The blue line represents the percentage of time each stimulus was identified in Set 1a as having the emotion label given on horizontal axis. The red line is the average response for the clarity (Set 2) of the emotion for each of the respective labels.

Figure 5.12 provides a graphical summary of the clarity ratings for each of the four intended vocal expressions. The ratings illustrated in Figure 5.12 (a) and (b) indicate that the highest clarity was given to the category that matched the intended emotion, which were anger and happiness, respectively. Regarding the intended emotion of sadness given in Figure 5.12 (c), the greatest overall clarity ratings were awarded to the correct category, but the write-in category received higher clarity ratings than was found for the other three intended expressions. This observation reaffirms the general trend found in the write-in analysis, where the descriptors such as tired, depressed, bored, and overwhelmed were all among the most frequently submitted responses. Generally speaking, these results indicate that the word "sadness" was the least fitting of the decoder's perception of the vocal expressions portrayed.


Intended Expression: Anger

Intended Expression: Happiness

Intended Expression: Sadness

Intended Expression: Neutral

(a)

(b)

(c)

(d)

Figure 5.12 Clarity ratings of the stimuli for the intended emotions (a) anger, (b) happiness, (c) sadness, and (d) no expression.

5.3.4 Comparisons to Acoustics

The following analysis aimed to examine how well the acoustics of the stimuli explain the results of the subjective listening tests. The presented framework follows closely that of Lima & Castro (2011) and Banse & Scherer (1996). Multiple regression analyses were performed on the data collected from Set 1, and a set of the acoustic variables discussed in Chapter 4. The independent variables were 15 acoustic cues that had previously demonstrated significant differences between emotion categories. The independent variable was the number of times that each stimulus had been rated as one of each of the four provided categories (anger, happiness, sadness, neutral). This linear regression process was repeated for each of the four expression categories and the results are provided in Table 5.5.

Sadness had the lowest error ($R^2 = 0.711$) and correlated with HNR_{M}, $F1_{bw,\mathrm{M}}$, $F0_{\mathrm{SD,}}$ and F_{syll}. Happiness ($R^2 = 0.642$) had a strong positive beta weight for $F0_{\mathrm{SD}}$, and anger had an even stronger, but negative relationship with HNR_{M}. These results generally indicate that the lower the harmonicity (less periodic the signal), the more likely the participants were to identify the stimulus as angry. Stimuli with greater variance in $F0$ were significantly more likely to be identified as happy. The results for happiness $F0_{\mathrm{SD}}$ agree with Juslin & Laukka (2001). Spectral cues like S_{hf500} did not exhibit significant differences- a result that is not shared by Lima & Castro (2011) and Juslin & Laukka (2001). Speech rate was highly significant for both sadness ($ß = -3.10$, $p < 0.05$) and the neutral expressions ($ß = 5.19$, $p < 0.01$). Also of interest is the sign besides each of these beta weights. A higher speech rate seems to correlate with neutral expressions, while variance in the recognition of sadness appears to share a negative relationship with speech rate. This result is in line with the predictions of Scherer (1986), but was not identified as significant by Juslin & Laukka (2001) ($ß = 0.05$, $p > 0.05$).

Table 5.5 Beta weights and proportion of variance explained (bottom row) for the four categories of expression. Bolded numbers are significant at $p < 0.05$.

Cue	Anger	Happiness	Sadness	Neutral
T_{utt}	-4.07	2.04	6.01	-3.98
T_{vo}	4.53	-2.68	-4.89	3.04
P_{paus}	-2.32	-0.72	0.09	2.95
F_{syll}	-2.95	0.85	**-3.10**	**5.19**
$L_{AI,\mathrm{M}}$	-2.19	-1.77	1.50	**2.46**
$F0_{\mathrm{M}}$	-0.75	2.75	2.38	**-4.38**
$F0_{\mathrm{SD}}$	-4.11	**7.19**	**-4.47**	1.39
S_{alpha}	-1.44	0.14	-0.22	1.52
S_{hf500}	2.28	-0.87	-0.96	-0.45
$F1_{\mathrm{M}}$	-2.07	0.61	1.65	-0.19
$F1_{bw,\mathrm{M}}$	2.16	-0.44	**1.81**	**-3.54**
$F2_{\mathrm{M}}$	**2.61**	-1.15	0.32	-1.78
$F2_{bw,\mathrm{M}}$	0.41	-0.86	-0.18	0.63
HNR_{M}	**-8.41**	2.25	**5.88**	0.27
$F0_{jitt,\mathrm{M}}$	-2.23	-0.22	1.73	0.73
Adj. R^2	0.471	0.642	0.711	0.521

The beta weights from the multiple regression analysis of the intensity ratings and the recognition rates is provided in Table 5.6. For each intended emotion, rows where either intensity or recognition rates correlated with the acoustic cues are highlighted in gray. For anger, just 20% of the cue correlations for intensity were significant for recognition. Significance was shared for 25% of the significant cues for happiness. None of the cues that were found to have a significant effect on recognition rates were found to be significantly correlated with the rated emotional intensity. These results indicate that overall, the acoustic cues that correlated significantly with the recognition of emotion did not necessarily correlate with the perception of emotional intensity.

Table 5.6 Beta weights from the multiple regression analysis for intensity ratings and recognition rates. Bolded text indicates significance at $p < 0.05$.

	Anger		**Happiness**		**Sadness**	
Cue	Intensity	Recognition	Intensity	Recognition	Intensity	Recognition
T_{vo}	32.00	4.53	**20.63**	-2.68	**11.58**	-4.89
P_{paus}	7.85	-2.32	**11.95**	-0.72	2.78	0.09
F_{syll}	0.06	-2.95	3.84	0.85	3.95	**-3.10**
$F0_{\mathrm{M}}$	**16.07**	-0.75	2.33	2.75	9.52	2.38
$F0_{\mathrm{SD}}$	**-12.82**	-4.11	**5.48**	**7.19**	-2.80	**-4.47**
S_{alpha}	4.26	-1.44	0.30	0.14	**-10.78**	-0.22
$F1_{\mathrm{M}}$	-3.53	-2.07	1.30	0.61	**-3.47**	1.65
$F1_{bw,\mathrm{M}}$	-2.88	2.16	-0.11	-0.44	-1.75	**1.81**
$F2_{\mathrm{M}}$	1.62	**2.61**	1.02	-1.15	0.89	0.32
$F2_{bw,\mathrm{M}}$	-0.18	0.41	**-2.83**	-0.86	0.01	-0.18
HNR_{M}	**-12.40**	**-8.41**	-2.93	2.25	-2.64	**5.88**
$F0_{jitt,\mathrm{M}}$	**4.81**	-2.23	-4.75	-0.22	0.98	1.73

6 Summary and Future Work

This thesis has attempted to detail the theory, methods, and composition of a corpus of emotional speech. Theories on the definitions for emotion and how specific emotions are differentiated were discussed in Chapter 2. Successive exploration of the voice production system established a link between the pressure waves, the body, and digital signals. Scherer's Component Process Theory established a reference for the expectations in how emotion might be expressed acoustically and a set of predictions made for anger, happiness and sadness [6].

For the classical category of prosodic cues such as *duration*, *intensity*, and *pitch*, Scherer predicted increasing or decreasing trends from a neutral or non-expressive state. *Duration* cues like speech rate (F_{syll}) were expected to increase along the arousal dimension, which suggests that anger and happiness would cause an increase in F_{syll}, while sadness would be marked by a decrease in F_{syll}. These predictions agree with the broad consensus in the literature [3] [22]. Cues belonging to the *intensity* category were also predicted to rise in level and variability for emotions of higher arousal, thus reflecting the same trends as was described for *duration*. General consensus in the literature reflected these predicted trends as well. Predictions for *pitch* (*F0*) cues were slightly less unidirectional than the previous categories were. Scherer (1986) expected elation (happiness) to cause increases in $F0_{\mathrm{M}}$, $F0_{\mathrm{R}}$, $F0_{\mathrm{SD}}$, and the short-term variability of pitch $F0_{jit}$. Of this list of *F0* predictions, the mean, range, and standard deviations were found to be consistent with the literature, however results from the few studies that included $F0_{jit}$ showed decreases for sadness against predicted trends. The *voice quality* category of cues offered a way to differentiate emotions along the valence dimension, although fewer of Scherer's predictions were consistent with the consensus determined by Juslin & Laukka (2003).

The development of recording procedures and acoustical optimization of the speech corpus was discussed in Chapter 3. Many factors were considered in maximizing the quality of the corpus including acoustical treatments to the recording room, communication between participating parties, and digital processing of the recorded stimuli.

In Chapter 4, the emotional speech corpus was processed and analyzed for markers of emotion. The primary goal of this chapter was to assess whether the acoustic cues known to correlate with emotion did so for this corpus. The size of the dataset and the number of possible salient cues increased the need for clearly stated cue definitions and a well-organized file structure. The results of these analyses were compared to theory and the most general consensus found in the literature. The lack of a consistent baseline and differences in cue definitions complicated both the comparison and conclusions that could be drawn from the analyses. Overall, the differences in low-level descriptor (LLD) values between a non-expressive and the emotional portrayals matched consensus of previous studies 58% of the time for the LLDs that could be compared. The effects of language content were also assessed. Although it was found that both script and utterance significantly affected the acoustics of the emotional portrayals, the magnitude of their effect was lower by a factor of ten.

Objectively speaking, the corpus of emotional speech had been found to comprise significant acoustical content, but perception and measurement can be vastly different from one another. To fully assess the accuracy and quality of these portrayals of emotion, a listening study was conducted. The main goals of this study were to determine if listeners could decode the intended emotion of the speaker, and to investigate which acoustic cues correlate with the perception of emotion as a category and its intensity. Recognition rates across all emotions were 92% on average, which was 3.7 times higher than chance. This result indicates unequivocally that the decoders were able to discriminate the intended emotion of the encoder as presented acoustically. Results from the analysis of variance indicate that semantic content did not affect the recognition rates or the response times in the recognition task.

The intensity ratings told a slightly different story. As encoded, it was assumed that the intended intensity of the emotion as portrayed was as high as possible or 100% of their perception. The average intensity ratings of the decoders was on average only 65% for the three emotions of anger, happiness, and sadness. This result was significantly lower than the assumed 100% intensity of the encoder. Treated as a dependent variable, rated emotional intensity was significantly affected by script, the interaction of emotion X script, as well as the interaction of rated emotion X script X utterance. Overall, the results from the intensity analysis indicate that semantic content significantly affected the perception of emotional intensity. In addition to classification performance and perception of emotional intensity, contextual effects on the perception of emotion was examined. On average, participants that listened to each utterance as they would occur in conversational speech gave lower clarity ratings than the average recognition rates from the previous tasks.

Comparisons to acoustical trends were also made. Of the LLDs that exhibited significant differences from the acoustical assessment in Chapter 3, only a handful significantly contributed to the classification of emotion. $F0_{\mathrm{SD}}$ strongly correlated with recognition of happiness, and lower values for HNR_{M} were found to significantly correlate with the recognition of anger. Both of these results match the acoustically determined trends. The acoustic trends matched the recognition of sadness, which correlated with a decrease in syllabic rate, and an increase in HNR_{M}. Sadness was also correlated with a decrease in $F0_{\mathrm{SD}}$, which is opposite the direction of the acoustically determined trend.

The wealth of information and the depths to which future analyses can go cannot be overstated. Just a fraction of the speech from only three out of the forty-two encoders have been assessed in this thesis. As a whole, the corpus here described is one of most extensive of its kind with over 3.8 hours of fully transcribed prompted emotional speech recorded in a laboratory setting. While many of the independent variables investigated were sound to have significant main effects on the dependent variable, it can hardly be said that these effects are representative of the entire corpus with such a small sample size. This work focused largely on global trends that were averaged across utterances or scripts. Future work would greatly benefit from an utterance level analysis of variance on the existing data set. Easy targets for future work includes analysis of the formants that are specific to vowels in the transcripts of the scripts. It is quite possible that much more variance could be explained by shifts in the formant structure that are unique to each vowel.

Lastly, the direction of more recent works is tending towards the use of higher level computational statistics as the world of machine learning continues to advance [34] [58]. This seems like a very natural tool for testing the theories proposed by Scherer's (1986) Component Process Model (see Figure 4.10). Although only binary decisions are represented here, a range of underlying activation potentials that are characteristic of neural networks remains to be explored. Such an analysis would synthesize one of the more profound advancements in modern technology within a framework constructed from psychological theory rather than a black box of statistical machines.

References

[1] T. Jay, The Psychology of Language, Upper Saddle River, New Jersey: Prentice Hall, 2003.

[2] K. Scherer, "Psychological Models of Emotion," in *The Neuropsychologo of Emotion*, New York, Oxford University Press, 2000, pp. 137-162.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine,* vol. 18, no. 1, pp. 32-80, 2001.

[4] C. Darwin, The Expression of the Emotions in Man and Animals, London: John Murray, 1872, p. 374.

[5] P. Ekman, "An Argument for Basic Emotions," *Cognition and Emotion,* vol. 6, no. (3/4), pp. 169-200, 1992.

[6] K. Scherer, "Vocal Affect Expression: A Review and a Model for Future Research," *Psychological Bulletin,* vol. 99, no. 2, pp. 143-165, 1986.

[7] C. E. Izzard, Human Emotion, New York: Plenum, 1977.

[8] H. Wagner and A. Manstead, Eds., "Vocal correlates of emotional arousal and affective disturbance," in *Handbook of social psychophysiology*, New York, Wiley, 1989, pp. 165-197.

[9] J. R. Krebs and R. Dawkins, "Animal signals: Mind-reading and manipulation," in *Behavioral ecology: An evolutionary approach*, Oxford, Blackwell, 1984, pp. 380-402.

[10] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication,* vol. 40, pp. 227-256, 2003.

[11] L. R. Rabiner and R. W. Schafer, Theory and Applications of Digital Speech Processing, 1 ed., A. Gilfillan, Ed., Upper Saddle River, NJ: Pearson Higher Education Inc., 2011.

[12] J. L. Flanagan, C. H. Coker, L. R. Rabiner, R. W. Schafer and N. Umeda, "Synthetic voices for computers," *IEEE Spectrum,* vol. 7, no. 10, pp. 22-45, October 1970.

[13] E. Breatnach, G. Abbott and R. Fraser, "Dimensions of the normal human trachea," *American Journal of Roentgenology,* vol. 142, no. 5, pp. 903-906, 1984.

[14] J. Elcner, J. Jedelsky, F. Liza and M. Jicha, "Velocity profiles in idealized

model of human respiratory tract," in *EPJ Web of Conferences*, 2013.

[15] A. Stevenson, Ed., Oxford Dictionary of English, Oxford University Press, 2010.

[16] P. Ekman, Darwin and Facial Expressions, New York: Academic Press, 1973.

[17] M. Davis and H. and College, Recognition of Facial Expressions, New York: Arno Press, 1975.

[18] K. Scherer and P. Ekman, Approaches to Emotion, Mahwah, New Jersey: Lawrence Erlbaum Associates, 1984.

[19] P. Ekman and W. Friesen, The Facial Action Coding system, San Francisco: Consulting Psychologists Press, 1978.

[20] V. Tartter, "Happy talk: Perceptual and acoustic effects of smiling on speech.," *Perception and Psychophysics,* vol. 27, no. 1, pp. 24-27, 1980.

[21] B. Schuller, A. Batliner, S. Steidl and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication,* vol. 53, no. 1, pp. 1062-1087, 2011.

[22] P. N. Juslin and P. Laukka, "Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code?," *Psychological Bulletin,* vol. 129, no. 5, pp. 770-814, 2003.

[23] A. M. Laukkanen, E. Vilkman, P. Alku and H. Oksanen, "Physical variations related to stress and emotional state: A preliminary study," *Journal of Phonetics,* vol. 24, pp. 313-335, 1996.

[24] M. Pell, S. Paulmann, C. Dara, A. Alasseri and S. Kotz, "Factors in the recognition of vocally expressed emotions: A comparision of four languages," *Journal of Phonetics,* vol. 37, no. 1, pp. 417-435, 2009.

[25] C. Busso, S. Lee and S. Narayanan, "Busso, C.; Lee, S.; Narayanan, S.," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 17, no. 17, pp. 582-596, 2009.

[26] D. R. Ladd, K. E. A. Silverman, F. Tolkmitt, G. Bergmann and K. R. Scherer, "Evidence of independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect," *Journal of the Acoustical Society of America,* vol. 78, no. 2, pp. 435-444, 1985.

[27] R. Banse and K. Scherer, "Acoustic Profiles in Vocal Emotion Expression," *Journal of Personality and Social Psychology,* vol. 70, no. 3, pp. 614-636, 1996.

[28] T. Bänziger and K. Scherer, "The role of intonation in emotional expressions," *Speech Communication,* vol. 46, no. 3-4, pp. 252-267, 2005.

[29] G. Borden, K. Harris and L. J. Raphael, Speech science primer: Physiology, acoustics and perception of speech, Baltimore: Williams & Williams, 1994.

[30] P. Boersma, "Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound," in *Proceeding of the*

Institute of Phonetic Sciences, 1993.

[31] S. Patel, K. R. Scherer, E. Björkner and J. Sundberg, "Mapping emotions into acoustic space: The role of voice production," *Biological Psychology,* vol. 87, no. 1, pp. 93-98, 2011.

[32] R. Van Bezooijen, Characteristics and Recognizability of Vocal Expressions of Emotions, Dordrecht: Foris, 1984.

[33] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg and L. Wedin, "Perceptual and acoustic correlates of voice qualities," *Acta Otolaryngologica,* vol. 90, pp. 441-451, 1980.

[34] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication,* vol. 48, no. 9-10 (SI), pp. 1162-1181, 2006.

[35] I. Engberg, A. Hansen, O. Andersen and P. Dalsgaard, "Design, recording and verification of a Danish emotional speech database," in *Proceedings of the Fifth European Conference on Speech Communication and Technology*, Rhodes, 1997.

[36] J. Hirschberg, J. Liscombe and J. Venditti, "Experiments in emotional speech," in *Proceedings of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, 2003.

[37] R. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. and Weiss, "A database of German emotional speech," in *Proceedings of Interspeech*, Lisbon, 2005.

[38] T. Bänziger, H. Pirker and K. Scherer, "Gemep – Geneva multimodal emotion portrayals: a corpus for the study of multimodal emotional expressions," in *Proceedings of LREC Workshop on Corpora for Research on Emotion and Affect*, Genova, 2006.

[39] P. Belin, S. Fillion-Bilodeau and F. and Gosselin, "The 'Montreal affective voices': A validated set of nonverbal affect bursts for research on auditory affective processing," *Behavior Research Methods,* vol. 40, no. 2, pp. 531-539, 2008.

[40] A. Mehrabian and M. Wiener, "Decoding of Inconsistent Communications," *Journal of Personality and Social Psychology,* vol. 6, no. 1, pp. 109-114, 1967.

[41] J. Morton and S. Truhub, "Children's Understanding of Emotion in Speech," *Child Development,* vol. 72, no. 3, pp. 834-843, 2001.

[42] P. Lieberman and S. Michaels, "Some aspects of Fundamental Frequency and Envelope Amplitude as Related to the Emotional Content of Speech," *Journal of the Acoustical Society of America,* vol. 34, no. 7, pp. 922-927, 1962.

[43] S. Castro and C. Lima, "Recognizing emotions in spoken language: A validated set of Protuguese sentences and speudosentences for research on emotional prosody," *Behavior Research Methods,* vol. 42, no. 1, pp. 74-81, 2010.

[44] P. Laukka, D. Neiberg, M. Forsell, I. Karlsson and K. Elenius, "Expression of affect in spontaneous speech: Acoustic corelates and automatic detection of irritation and resignation," *Computer Speech and Language,* vol. 25, no. 1, pp. 84-104, 2011.

[45] S. Nowicki and M. P. Duke, "Individual Differences in the Nonverbal Communication of Affect: The Diagnostic Analysis of Nonverbal Accuracy Scale," *Journal of Nonverbal Behavior,* vol. 18, no. 1, pp. 9-35, 1994.

[46] K. Shifflett-Simpson and E. Cummings, "Mixed message resolution and children's response to interadult conflict," *Child Development,* vol. 67, no. 1, pp. 437-448, 1996.

[47] T. D. Rossing, Springer Handbook of Acoustics, New York, New York: Springer New York, 2014.

[48] Pixar Animation Studios, "Inside Out Behind The Scenes Footage - Amy Poehler, Bill Hader, Mindy Kaling, Lewis Black," Walt Disney Pictures, 19 June 2015. [Online]. Available: https://www.youtube.com/watch?v=6YiqKqgd_jU&t=1s. [Accessed 18 September 2017].

[49] Occupational Safety and Health Administration, "Occupational noise exposure," 1991. [Online]. Available: https://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=standards&p_id=9735. [Accessed 23 September 2017].

[50] R. Technologies, "MRI Video," 2017. [Online]. Available: http://www.mrivideo.com/visuastimdigital.php. [Accessed 28 10 2017].

[51] Adobe Systems, "Adobe Audition," San Jose, 2015.

[52] The Mathworks Inc., "MATLAB," Natick, 2017.

[53] E. Berger, "Hearing Protection Devices," in *The Noise Manual*, 5 ed., E. H. Berger, L. H. Royster, J. D. Royster, D. P. Driscoll and a. M. Layne, Eds., Fairfax, Virginia: American Industrial Hygiene Association, 2000, pp. 379-455.

[54] P. Cole, R. O. Gilmore, K. S. Scherf and K. & Perez-Edgar, "Databrary," June 2016. [Online]. Available: https://nyu.databrary.org/volume/248. [Accessed 19 September 2017].

[55] P. Juslin and P. Laukka, "Impact of Intended Emotion Intensity on Cue Utilization and Decoding Accuracy in Vocal Expression of Emotion," *Emotion,* vol. 1, no. 4, pp. 381-412, 2001.

[56] P. Boersma and D. Weenick. [Online]. Available: http://www.praat.org/. [Accessed 10 October 2017].

[57] S. Wiethoff, D. Wildgruber, B. Kreifelts, H. Becker, C. Herbert, W. Grodd and T. Ethofer, "Cerebral processing of emotional prosody- influence of acoustic parameters and arousal," *NeuroImage,* vol. 39, no. 2, pp. 885-893, 2008.

[58] R. Fernandez and R. Picard, "Reconizing affect from speech prosody using hierarchical graphical models," *Speech Communication,* vol. 53, pp. 1088-1103, 2011.

[59] C. Lima and S. Castro, "Speaking to the Trained Ear: Musical Expertise Enhances the Recognition of Emotions in Speech Prosody," *Emotion,* vol. 11, no. 5, pp. 1021-1031, 2011.

[60] J. Sundberg, S. Patel, E. Björkner and K. R. Scherer, "Interdependencies among Voice Source Parameters in Emotional Speech," *IEEE Transactions on Affective Computing,* vol. 2, no. 3, pp. 162-174, 2011.

[61] M. Goudbeek and K. Scherer, "Beyond arousal: Valence and potency/control cues in the vocal expression of emotion," *Journal of the Acoustical Society of America,* vol. 128, no. 3, pp. 1322-1336, 2010.

[62] S. Patel, K. R. Scherer, J. Sundberg and a. E. Björkner, "Acoustic Markers of Emotions Based on Voice Physiology," in *Proceedings of Speech Prosody*, Chicago, 2010.

[63] D. Bone, C.-C. Lee and S. Narayanan, "Robust Unsupervised Arousal Rating: A Rule-Based Framework with Knowledge-Inspired Vocal Features," *IEEE Transactions on Affective Computing,* vol. 5, no. 2, pp. 201-213, 2014.

[64] K. Scherer, J. Sundberg, L. Tammarit and G. Salomão, "Comparing the acoustic expression of emotion in the speakering and the singing voice," *Computer Speech and Language,* vol. 29, no. 1, pp. 218-235, 2015.

[65] G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels," *Journal of the Acoustical Society of America,* vol. 24, no. 2, pp. 175-184, 1952.

[66] E. Bozkurt, E. Erzin, C. d. E. Erdem and A. T. Erdem, "Formant position based weighted spectral features for emotion recognition," *Speech Communication,* vol. 53, pp. 1186-1197, 2011.

[67] C. Gobl and A. N. Chasaide, "The role of voice quality in comminicating emotion, mood, and attitude," *Speech Communication,* vol. 40, pp. 189-212, 2003.

[68] E. Denoel and J.-P. Solvay, "Linear prediction of speech with a least absolute error criterion," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 33, no. 6, pp. 1397-1403, 1985.

[69] D. D. Deliyski, "Acoustic Model and Evaluation of Pathological Voice Production," in *Conference on Speech Communication and Technology*, Berlin, 1993.

[70] SAS Institute Inc., "SAS Analytics and Software Solutions," 2017. [Online]. Available: https://www.sas.com/en_us/home.html. [Accessed 28 October 2017].

[71] R. O. Gilmore, K. Perez-Edgar and K. S. Scherf, "Chilren's Neural Processing of the Emotional Environment (PEEP-II)," 2017. [Online]. Available: https://nyu.databrary.org/volume/339. [Accessed 24 October 2017].

[72] The Pennsylvania State Univeristy Clinical and Translational Science Institute, "StudyFinder," 2017. [Online]. Available: https://studyfinder.psu.edu/. [Accessed 24 October 2017].

[73] S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin and E. Clubley, "The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/high-functioning autism, males and females, scientists and mathematicians," *Journal of Autism and Developmental Disorders,* vol. 31, no. 1, pp. 5-17, 2001.

[74] M-Audio, "Support Knowledge Base," 25 February 2013. [Online]. Available: http://www.m-audio.com/support/documents-search. [Accessed 18 September 2017].

[75] K. R. Scherer, "Vocal correlates of emotional arousal and affective disturbance," in *Handbook of social psychophysiology*, H. Wagner and A. Manstead, Eds., New York, Wiley, 1989, pp. 165-197.

[76] C. Lima, S. Castro and S. K. Scott, "When voices get emotional: A corpus of nonverbal vocalizations for research on emotion processing," *Behavior Research Methods,* vol. 45, pp. 1234-1245, 2013.

[77] F. Eyben, Real-time Speech and Music Classification by Large Audio Feature Space Extraction, Munich: SpringerNature, 2016.

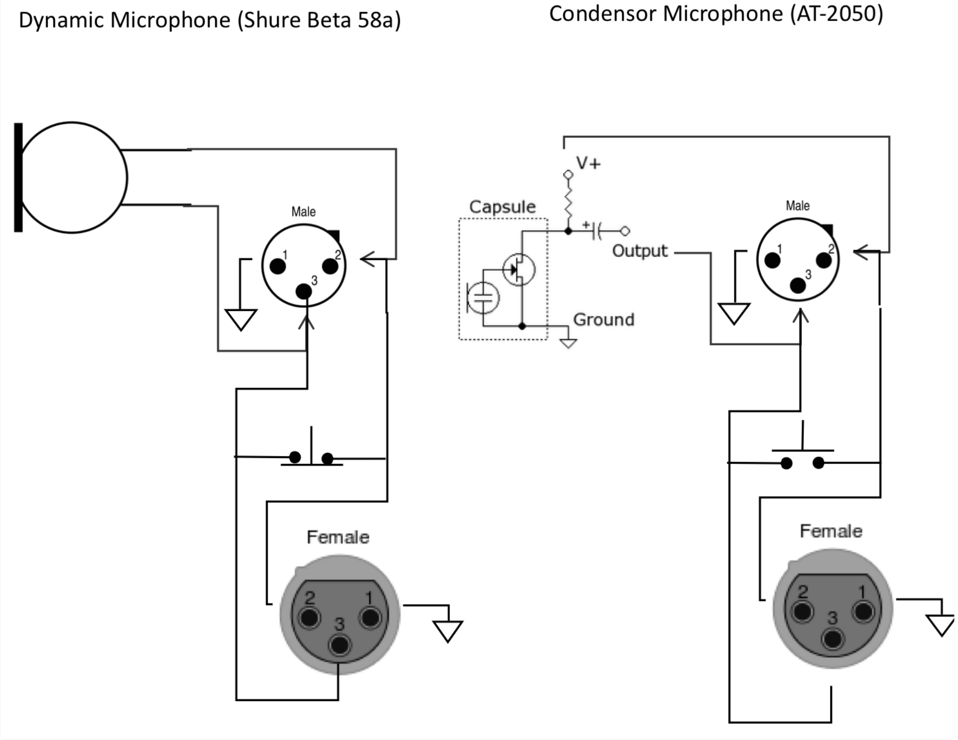
[78] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication,* vol. 40, no. 1, pp. 5-32, 2003.

[79] F. Sai, "The role of the mother's voice in developing mother's face preference: Evidence for intermodal perception at birth," *Infant and Child Development,* vol. 14, pp. 29-50, 2005.

[80] B. M. Repacholi and A. N. Meltzoff, "Emotional eavesdropping: Infants selectively respond to indirect emotional signals," *Child Development,* vol. 78, pp. 508-521, 2005.

Appendix A Electrical Circuit Diagram of the Push-to-Talk Box

The wiring of the “Push-to-Talk” box is given in the schematic below:

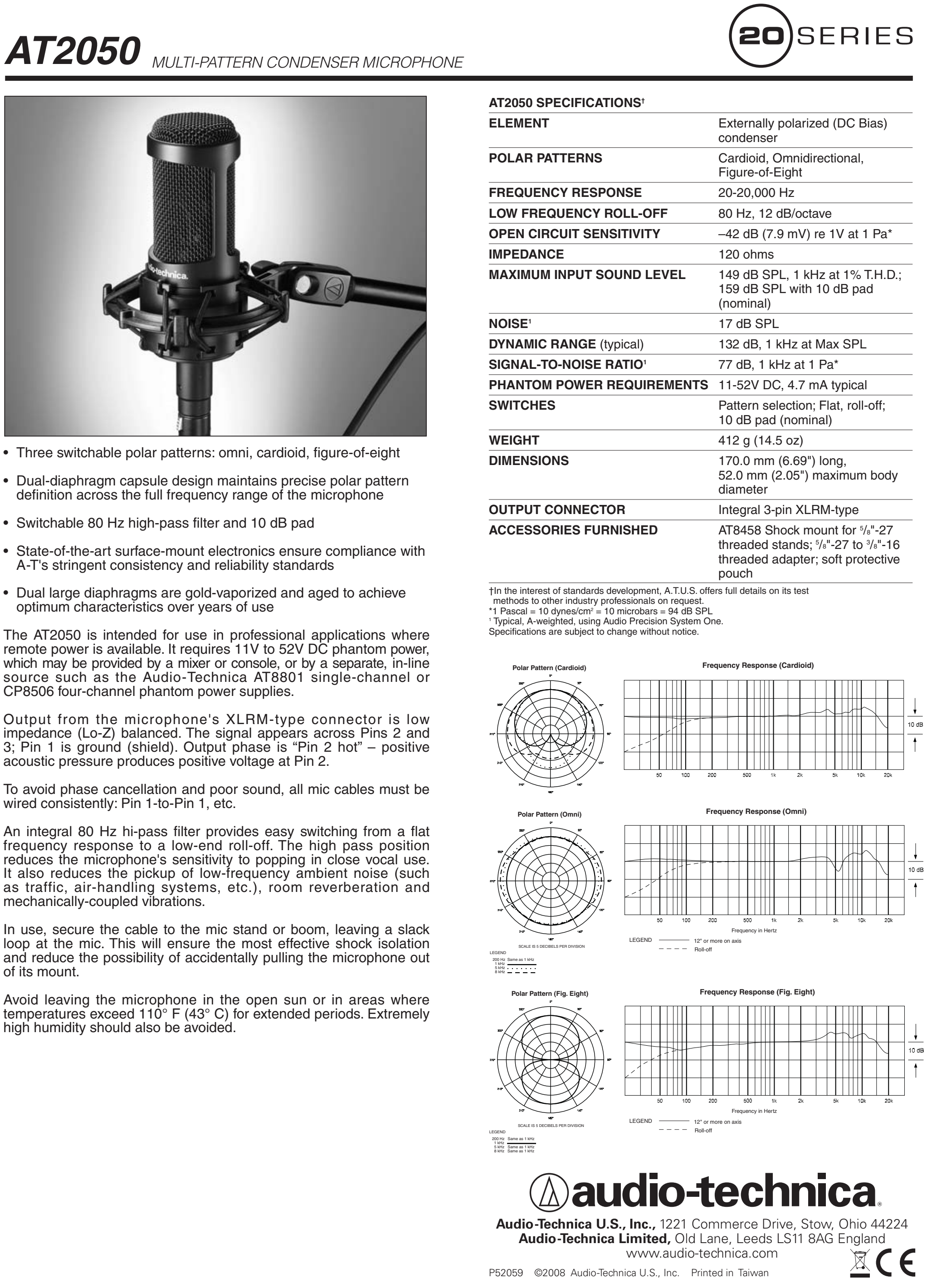


Dynamic Microphone (Shure Beta 58a)
Condensor Microphone (AT-2050)
Male
1
2
3
V+
Capsule
+
Output
Ground
Male
1
2
3
Female
2
1
3
Female
2
1
3

The “Push-to-Talk” box consists of the male XLR, button, and female XLR terminals for both the dynamic and condenser microphone channels.

Appendix B Measurement System Specifications

Audio-Technica AT-2050 technical specifications:



AT2050 *MULTI-PATTERN CONDENSER MICROPHONE*

20 SERIES

- Three switchable polar patterns: omni, cardioid, figure-of-eight
- Dual-diaphragm capsule design maintains precise polar pattern definition across the full frequency range of the microphone
- Switchable 80 Hz high-pass filter and 10 dB pad
- State-of-the-art surface-mount electronics ensure compliance with A-T's stringent consistency and reliability standards
- Dual large diaphragms are gold-vaporized and aged to achieve optimum characteristics over years of use

The AT2050 is intended for use in professional applications where remote power is available. It requires 11V to 52V DC phantom power, which may be provided by a mixer or console, or by a separate, in-line source such as the Audio-Technica AT8801 single-channel or CP8506 four-channel phantom power supplies.

Output from the microphone's XLRM-type connector is low impedance (Lo-Z) balanced. The signal appears across Pins 2 and 3; Pin 1 is ground (shield). Output phase is "Pin 2 hot" – positive acoustic pressure produces positive voltage at Pin 2.

To avoid phase cancellation and poor sound, all mic cables must be wired consistently: Pin 1-to-Pin 1, etc.

An integral 80 Hz hi-pass filter provides easy switching from a flat frequency response to a low-end roll-off. The high pass position reduces the microphone's sensitivity to popping in close vocal use. It also reduces the pickup of low-frequency ambient noise (such as traffic, air-handling systems, etc.), room reverberation and mechanically-coupled vibrations.

In use, secure the cable to the mic stand or boom, leaving a slack loop at the mic. This will ensure the most effective shock isolation and reduce the possibility of accidentally pulling the microphone out of its mount.

Avoid leaving the microphone in the open sun or in areas where temperatures exceed 110° F (43° C) for extended periods. Extremely high humidity should also be avoided.

AT2050 SPECIFICATIONS†	
ELEMENT	Externally polarized (DC Bias) condenser
POLAR PATTERNS	Cardioid, Omnidirectional, Figure-of-Eight
FREQUENCY RESPONSE	20-20,000 Hz
LOW FREQUENCY ROLL-OFF	80 Hz, 12 dB/octave
OPEN CIRCUIT SENSITIVITY	–42 dB (7.9 mV) re 1V at 1 Pa*
IMPEDANCE	120 ohms
MAXIMUM INPUT SOUND LEVEL	149 dB SPL, 1 kHz at 1% T.H.D.; 159 dB SPL with 10 dB pad (nominal)
NOISE[1]	17 dB SPL
DYNAMIC RANGE (typical)	132 dB, 1 kHz at Max SPL
SIGNAL-TO-NOISE RATIO[1]	77 dB, 1 kHz at 1 Pa*
PHANTOM POWER REQUIREMENTS	11-52V DC, 4.7 mA typical
SWITCHES	Pattern selection; Flat, roll-off; 10 dB pad (nominal)
WEIGHT	412 g (14.5 oz)
DIMENSIONS	170.0 mm (6.69") long, 52.0 mm (2.05") maximum body diameter
OUTPUT CONNECTOR	Integral 3-pin XLRM-type
ACCESSORIES FURNISHED	AT8458 Shock mount for 5/8"-27 threaded stands; 5/8"-27 to 3/8"-16 threaded adapter; soft protective pouch

†In the interest of standards development, A.T.U.S. offers full details on its test methods to other industry professionals on request.

*1 Pascal = 10 dynes/cm^2 = 10 microbars = 94 dB SPL

[1] Typical, A-weighted, using Audio Precision System One.

Specifications are subject to change without notice.

Polar Pattern (Cardioid)

Frequency Response (Cardioid)

Polar Pattern (Omni)

Frequency Response (Omni)

SCALE IS 5 DECIBELS PER DIVISION

LEGEND

200 Hz Same as 1 kHz
1 kHz
5 kHz
8 kHz

Polar Pattern (Fig. Eight)

Frequency Response (Fig. Eight)

SCALE IS 5 DECIBELS PER DIVISION

LEGEND

200 Hz Same as 1 kHz
1 kHz
5 kHz Same as 1 kHz
8 kHz Same as 1 kHz

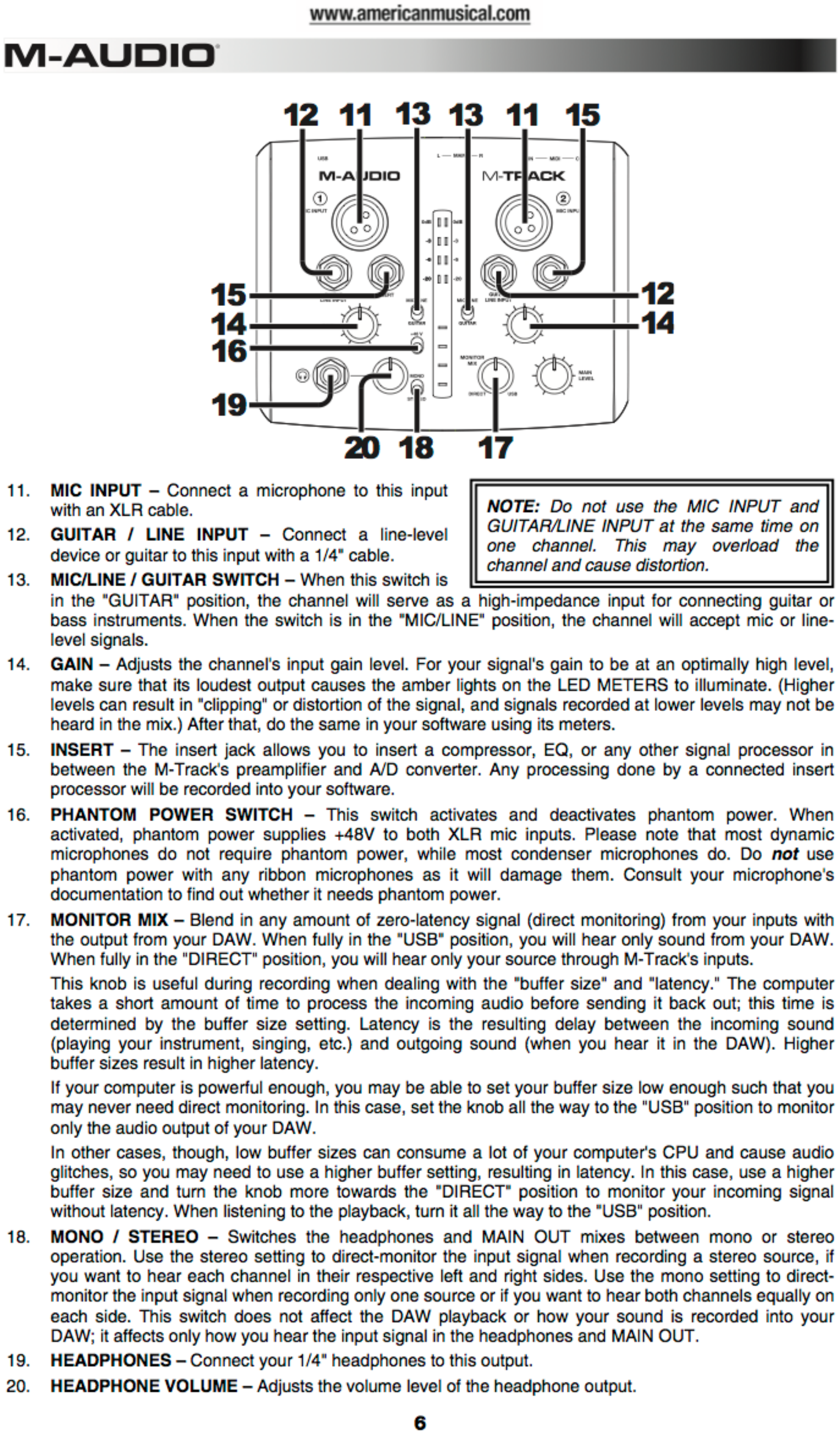
audio-technica®

Audio-Technica U.S., Inc., 1221 Commerce Drive, Stow, Ohio 44224
Audio-Technica Limited, Old Lane, Leeds LS11 8AG England
www.audio-technica.com

P52059 ©2008 Audio-Technica U.S., Inc. Printed in Taiwan

CE

M-Audio M-Track technical specifications [74]:



www.americanmusical.com

M-AUDIO

12
11
13
13
11
15
M-AUDIO
M-TRACK
15
14
16
19
12
14
20
18
17

11. **MIC INPUT –** Connect a microphone to this input with an XLR cable.
12. **GUITAR / LINE INPUT –** Connect a line-level device or guitar to this input with a 1/4" cable.
13. **MIC/LINE / GUITAR SWITCH –** When this switch is in the "GUITAR" position, the channel will serve as a high-impedance input for connecting guitar or bass instruments. When the switch is in the "MIC/LINE" position, the channel will accept mic or line-level signals.

***NOTE:** Do not use the MIC INPUT and GUITAR/LINE INPUT at the same time on one channel. This may overload the channel and cause distortion.*

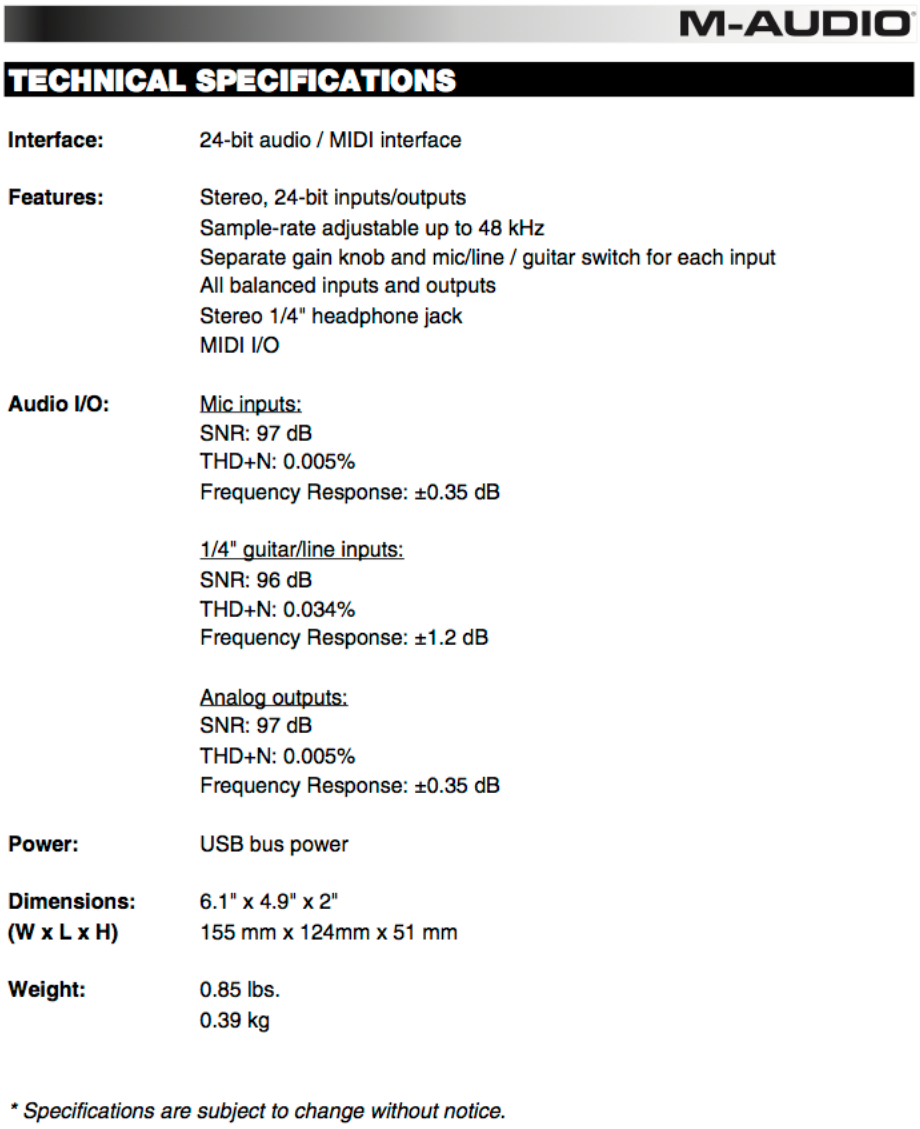
14. **GAIN –** Adjusts the channel's input gain level. For your signal's gain to be at an optimally high level, make sure that its loudest output causes the amber lights on the LED METERS to illuminate. (Higher levels can result in "clipping" or distortion of the signal, and signals recorded at lower levels may not be heard in the mix.) After that, do the same in your software using its meters.
15. **INSERT –** The insert jack allows you to insert a compressor, EQ, or any other signal processor in between the M-Track's preamplifier and A/D converter. Any processing done by a connected insert processor will be recorded into your software.
16. **PHANTOM POWER SWITCH –** This switch activates and deactivates phantom power. When activated, phantom power supplies +48V to both XLR mic inputs. Please note that most dynamic microphones do not require phantom power, while most condenser microphones do. Do ***not*** use phantom power with any ribbon microphones as it will damage them. Consult your microphone's documentation to find out whether it needs phantom power.
17. **MONITOR MIX –** Blend in any amount of zero-latency signal (direct monitoring) from your inputs with the output from your DAW. When fully in the "USB" position, you will hear only sound from your DAW. When fully in the "DIRECT" position, you will hear only your source through M-Track's inputs.

    This knob is useful during recording when dealing with the "buffer size" and "latency." The computer takes a short amount of time to process the incoming audio before sending it back out; this time is determined by the buffer size setting. Latency is the resulting delay between the incoming sound (playing your instrument, singing, etc.) and outgoing sound (when you hear it in the DAW). Higher buffer sizes result in higher latency.

    If your computer is powerful enough, you may be able to set your buffer size low enough such that you may never need direct monitoring. In this case, set the knob all the way to the "USB" position to monitor only the audio output of your DAW.

    In other cases, though, low buffer sizes can consume a lot of your computer's CPU and cause audio glitches, so you may need to use a higher buffer setting, resulting in latency. In this case, use a higher buffer size and turn the knob more towards the "DIRECT" position to monitor your incoming signal without latency. When listening to the playback, turn it all the way to the "USB" position.
18. **MONO / STEREO –** Switches the headphones and MAIN OUT mixes between mono or stereo operation. Use the stereo setting to direct-monitor the input signal when recording a stereo source, if you want to hear each channel in their respective left and right sides. Use the mono setting to direct-monitor the input signal when recording only one source or if you want to hear both channels equally on each side. This switch does not affect the DAW playback or how your sound is recorded into your DAW; it affects only how you hear the input signal in the headphones and MAIN OUT.
19. **HEADPHONES –** Connect your 1/4" headphones to this output.
20. **HEADPHONE VOLUME –** Adjusts the volume level of the headphone output.

6



M-AUDIO

TECHNICAL SPECIFICATIONS

Interface: 24-bit audio / MIDI interface

Features:
Stereo, 24-bit inputs/outputs
Sample-rate adjustable up to 48 kHz
Separate gain knob and mic/line / guitar switch for each input
All balanced inputs and outputs
Stereo 1/4" headphone jack
MIDI I/O

Audio I/O:

Mic inputs:
SNR: 97 dB
THD+N: 0.005%
Frequency Response: ±0.35 dB

1/4" guitar/line inputs:
SNR: 96 dB
THD+N: 0.034%
Frequency Response: ±1.2 dB

Analog outputs:
SNR: 97 dB
THD+N: 0.005%
Frequency Response: ±0.35 dB

Power: USB bus power

Dimensions:
(W x L x H)
6.1" x 4.9" x 2"
155 mm x 124mm x 51 mm

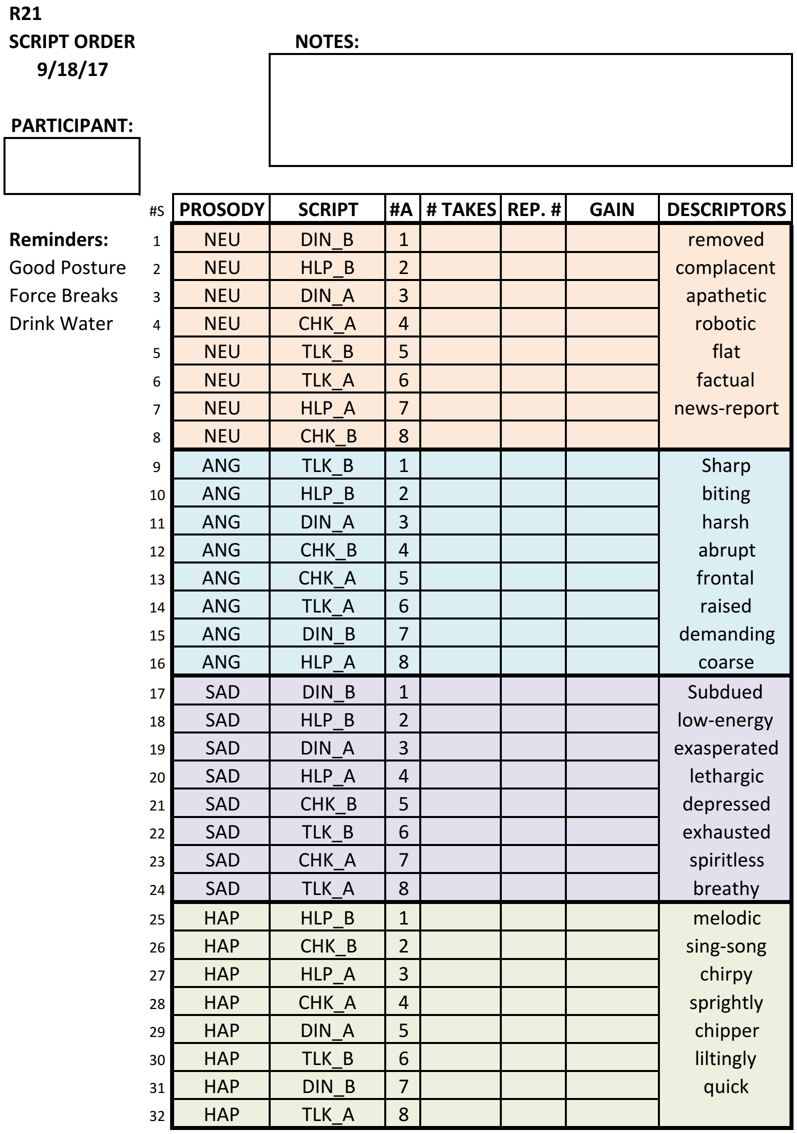
Weight:
0.85 lbs.
0.39 kg

** Specifications are subject to change without notice.*

Comprehensive Equipment list for the measurement system:

ID#	Function	Brand	Model
1	Condenser Microphone	Audio-Technica	AT-2050
2	Sound Card	M-Audio	M-Track
3	Laptop Computer	Lenovo	Thinkpad
4	Headphones	Audio-Technica	ATH-M40fs
5	Powered Monitors	Quickshot	Sound Force 680
6	Digital Audio Workstation	Adobe	Audition
7	Headphone Amplifier	Behringer	MicroAmp HA400

Appendix C Data Collection Instruments



R21
SCRIPT ORDER
9/18/17

PARTICIPANT:

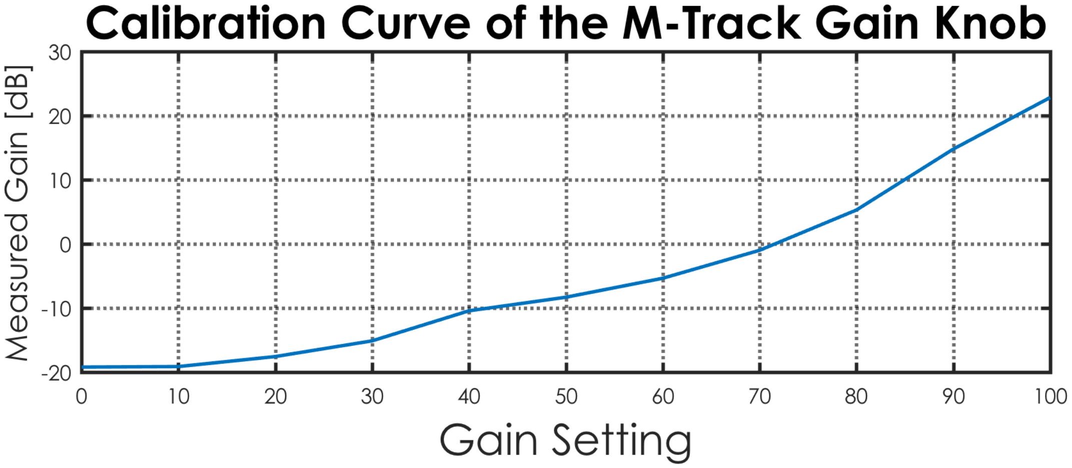
NOTES:

Reminders:
Good Posture
Force Breaks
Drink Water

#S	PROSODY	SCRIPT	#A	# TAKES	REP. #	GAIN	DESCRIPTORS
1	NEU	DIN_B	1				removed
2	NEU	HLP_B	2				complacent
3	NEU	DIN_A	3				apathetic
4	NEU	CHK_A	4				robotic
5	NEU	TLK_B	5				flat
6	NEU	TLK_A	6				factual
7	NEU	HLP_A	7				news-report
8	NEU	CHK_B	8				
9	ANG	TLK_B	1				Sharp
10	ANG	HLP_B	2				biting
11	ANG	DIN_A	3				harsh
12	ANG	CHK_B	4				abrupt
13	ANG	CHK_A	5				frontal
14	ANG	TLK_A	6				raised
15	ANG	DIN_B	7				demanding
16	ANG	HLP_A	8				coarse
17	SAD	DIN_B	1				Subdued
18	SAD	HLP_B	2				low-energy
19	SAD	DIN_A	3				exasperated
20	SAD	HLP_A	4				lethargic
21	SAD	CHK_B	5				depressed
22	SAD	TLK_B	6				exhausted
23	SAD	CHK_A	7				spiritless
24	SAD	TLK_A	8				breathy
25	HAP	HLP_B	1				melodic
26	HAP	CHK_B	2				sing-song
27	HAP	HLP_A	3				chirpy
28	HAP	CHK_A	4				sprightly
29	HAP	DIN_A	5				chipper
30	HAP	TLK_B	6				liltingly
31	HAP	DIN_B	7				quick
32	HAP	TLK_A	8				

This is a scaled version of a script order sheet that the test administrators would use to keep track of the number of takes and gain setting for each stimulus.

Appendix D Calibration Curve for the Sound Card



Appendix E Signal Processing Pipeline

The main objectives for these analyses are to determine the value of each LLD for the entire corpus of speech and to statistically compare these values to those found in the literature. From the input stimulus file to the output LLD value the handling and storage of information was organized in a structured and repeatable manner.

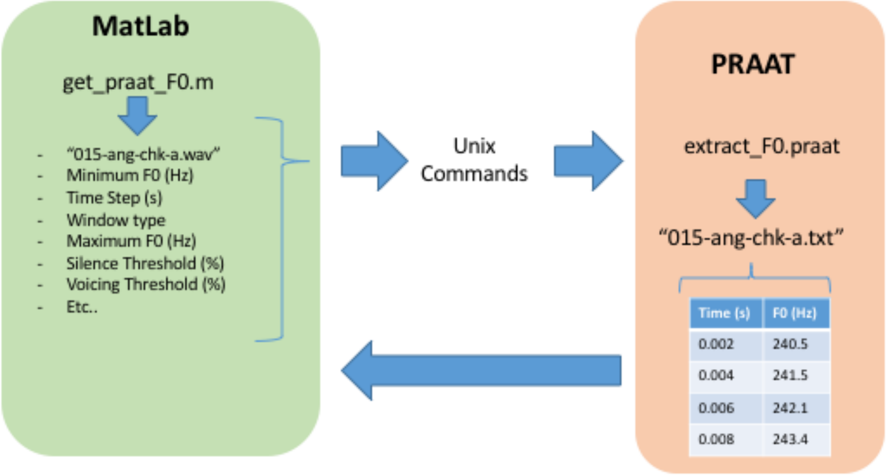
The analysis will focus on the stimuli that has been RMS normalized because the acoustical characteristics such as the RMS amplitudes are closer to the values during their presentation in PEEP. Having normalized the amplitudes of these stimuli, analysis of signal intensities will be limited to utterance-level variation. As previous work has indicated that loudness (intensity) is highly correlated with arousal, relative level differences between the unedited stimuli will be covered briefly.

The signal processing architecture was developed as a coordination between the computational software program MatLab [52] and the speech processing program PRAAT [56]. Although it was initially desired to unify the handling of data in a single software environment, the time saved by allocating analyses far outweighed the benefit of reinventing the wheel in code. Calculations of *F0*, HNR, formants, jitter, and shimmer were performed by PRAAT. The remaining LLDs were calculated with custom MatLab scripts and functions.

PRAAT's design as a scripted language accelerated what would have otherwise been a laborious and time-consuming manual procedure. PRAAT is both a speech processing software environment and a program that runs on scripts. In the PRAAT environment, stimuli can be manually imported and then processed depending on the LLD of interest. Alternatively, scripts written in the PRAAT language can be run that automatically imports and processes a stimulus to find one or several LLDs. PRAAT can also be opened and called to run a script with any number of unique arguments directly from a Unix or Linux system's command line. It was concluded that this last option offered the greatest time efficiency with the smallest probability of human error in data manipulation.

For each of the LLDs involving PRAAT, a custom script was written and saved with the following naming convention: "extract_LLD.praat", where "LLD" is replaced with *F0*, HNR, formants, and jitter. Shimmer was included in the PRAAT script for calculating jitter because both calculations apply to the same sets of *F0* values. Although it is very possible for each of these PRAAT scripts to be unified into one script that requires a larger number of arguments, separating each operation into separate files made the process of debugging less complicated. A final copy of each of these PRAAT scripts is provided in Appendix #. Although the input arguments to each of these scripts vary slightly, they all request a stimuli directory, filename, and either a time step or time interval parameter.

Because the rest of the LLDs were processed in MatLab, calls to the five PRAAT-based LLD scripts were also made in the MatLab environment. A graphical representation of the communication chain between both programs is shown in the figure below. Each of the PRAAT LLD scripts was complemented by a unique MatLab function controller. The naming convention for these MatLab-PRAAT controller functions followed similarly: “get_praat_LLD.m” where LLD “LLD” is replaced with *F0*, HNR, formants, and jitter. Each of these MatLab functions could be passed the arguments specific to each of the PRAAT algorithms, as well as speech file directories, file names, and temporal parameters as is depicted in the figure below.



Time (s)	F0 (Hz)
0.002	240.5
0.004	241.5
0.006	242.1
0.008	243.4

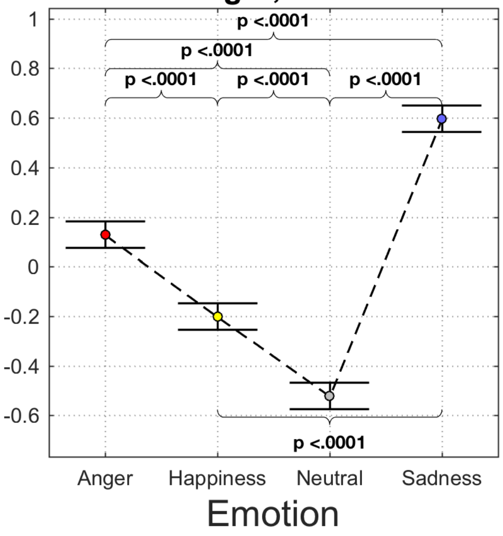
This is a visual representation of how a custom MatLab function was designed to call and pass arguments to a PRAAT script. Data from the text file generated by PRAAT was then read into MatLab for further analysis.

Appendix F Corpus Archival Structure

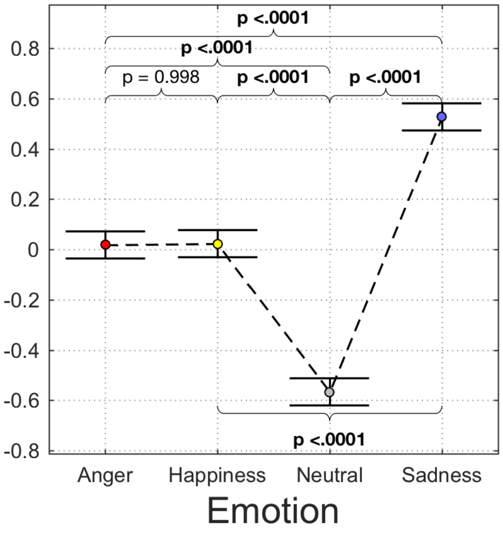
As originally saved, the speech corpus contains one folder for each encoder (55 total folders). Each of these folders is labeled by the three-digit identification number of the encoder as provided by PEEP (e.g. “001”, “002”). Within every encoder’s folder are two subfolders: one labeled “raw”, which contains the unedited stimulus files, and one labeled “norm_mono_unfiltered” which contains the single channel RMS normalized stimulus files.

Appendix G Least-Squares Means Comparisons between LLDs

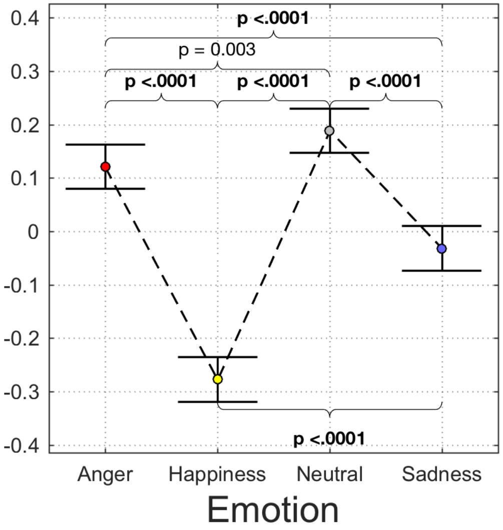
Utterance Length, n = 55 Encoders



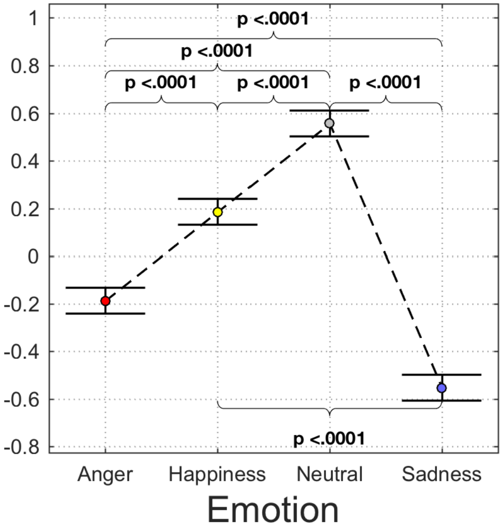
Voiced Duration, n = 55 Encoders

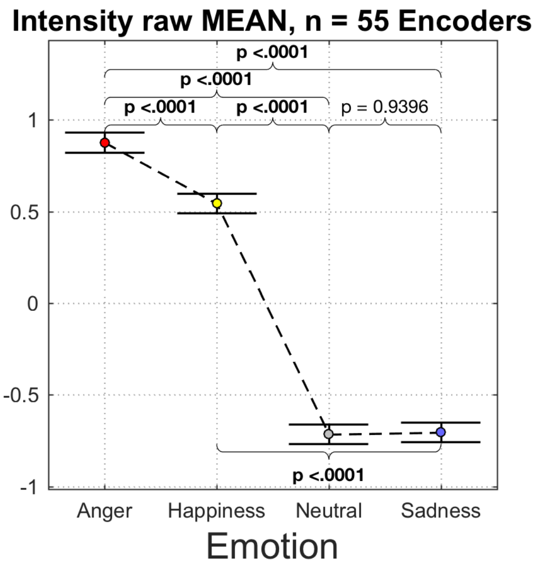


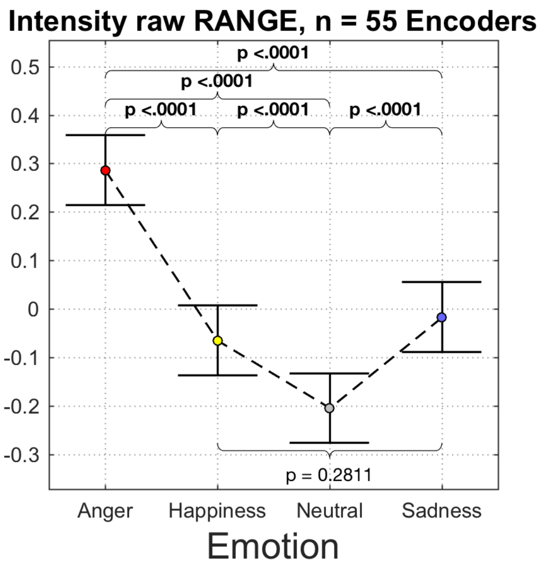
Proportion Pauses, n = 55 Encoders

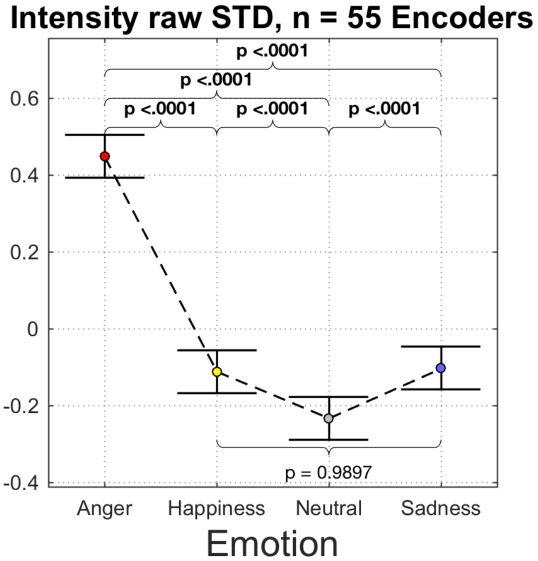


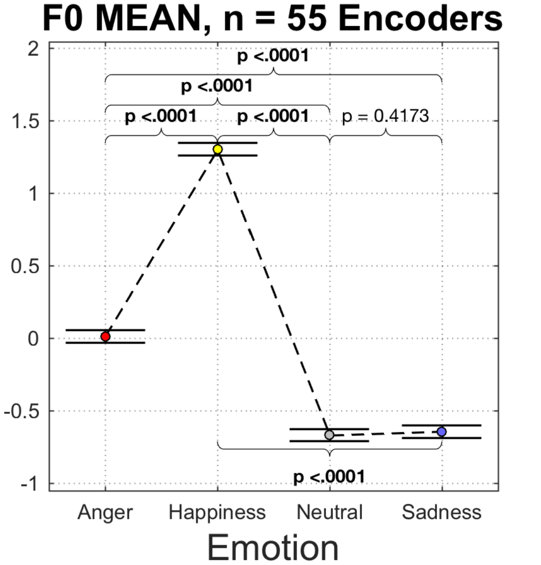
Syllabic Rate, n = 55 Encoders

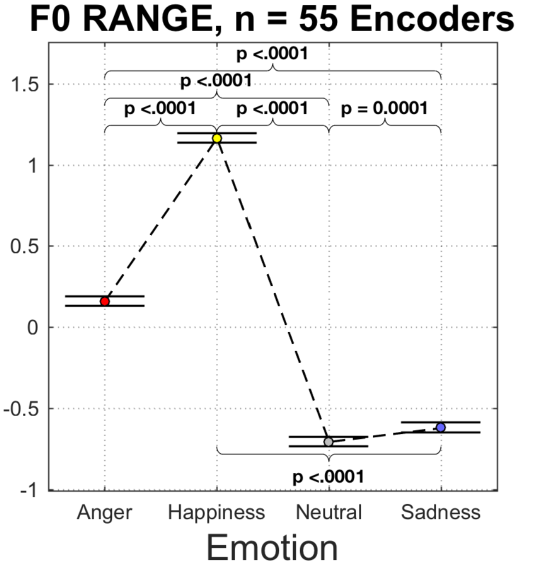


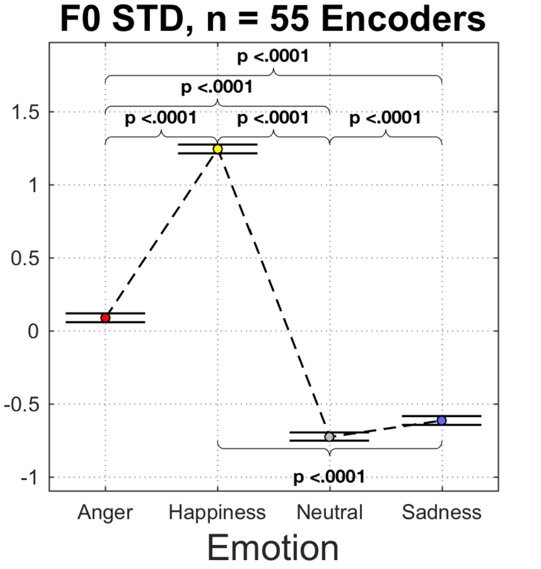

Intensity raw MEAN, n = 55 Encoders
p <.0001
p <.0001
p <.0001
p <.0001
p = 0.9396
p <.0001
1
0.5
0
-0.5
-1
Anger
Happiness
Neutral
Sadness
Emotion

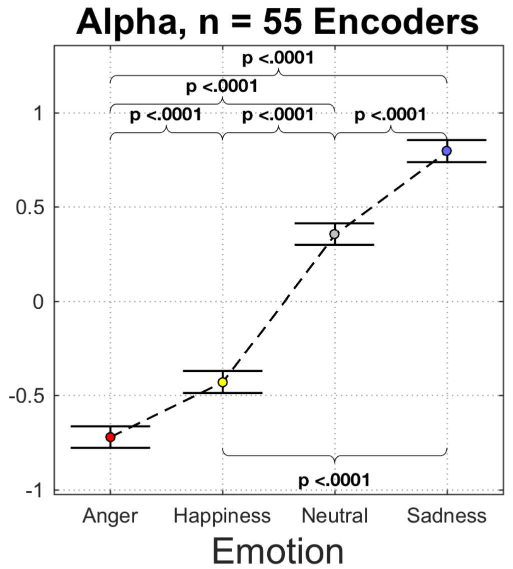

Intensity raw RANGE, n = 55 Encoders
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
p = 0.2811
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
Anger
Happiness
Neutral
Sadness
Emotion

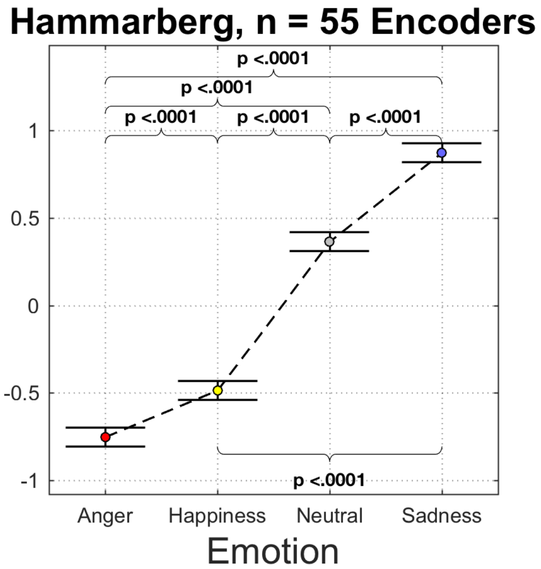

Intensity raw STD, n = 55 Encoders
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
p = 0.9897
0.6
0.4
0.2
0
-0.2
-0.4
Anger
Happiness
Neutral
Sadness
Emotion

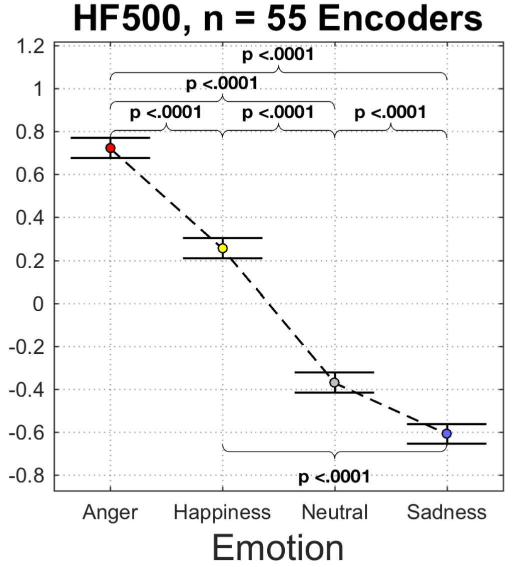

F0 MEAN, n = 55 Encoders
p <.0001
p <.0001
p <.0001
p <.0001
p = 0.4173
p <.0001
2
1.5
1
0.5
0
-0.5
-1
Anger
Happiness
Neutral
Sadness
Emotion

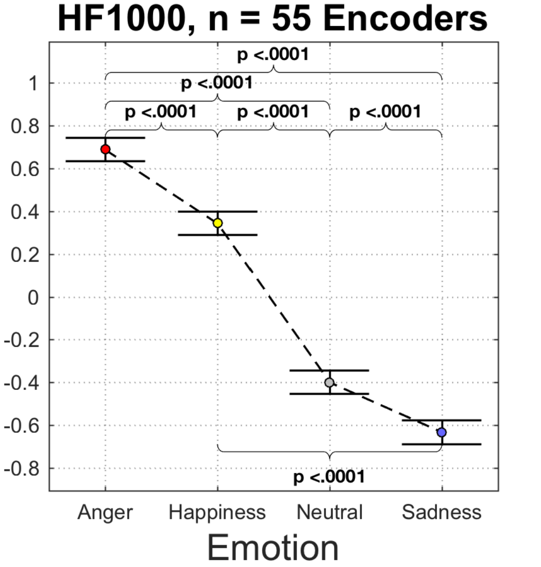

F0 RANGE, n = 55 Encoders
p <.0001
p <.0001
p <.0001
p <.0001
p = 0.0001
p <.0001
1.5
1
0.5
0
-0.5
-1
Anger
Happiness
Neutral
Sadness
Emotion

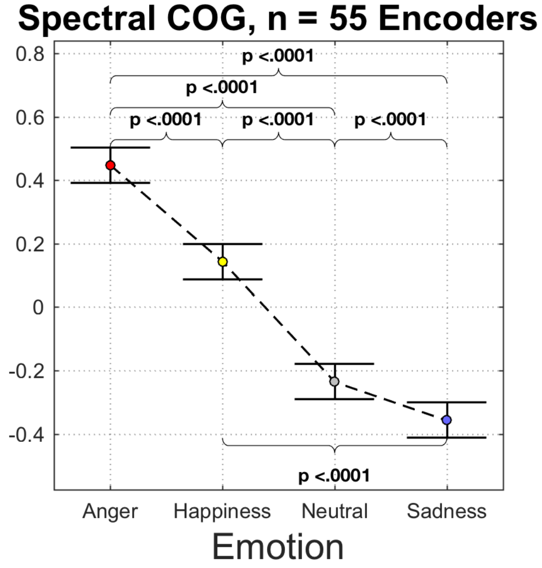

F0 STD, n = 55 Encoders
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
1.5
1
0.5
0
-0.5
-1
Anger
Happiness
Neutral
Sadness
Emotion

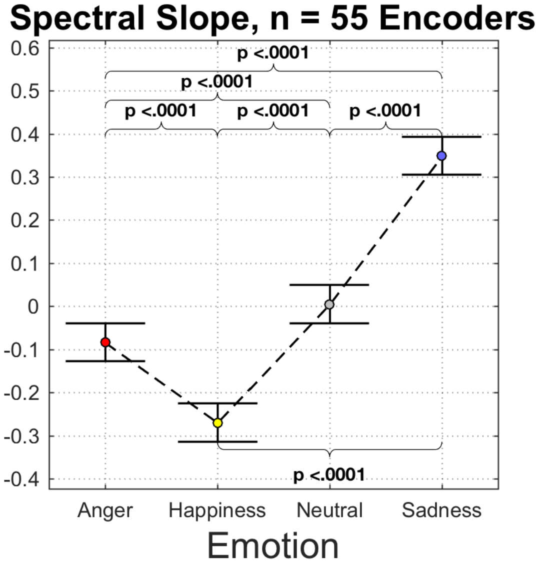

Alpha, n = 55 Encoders
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
1
0.5
0
-0.5
-1
Anger
Happiness
Neutral
Sadness
Emotion

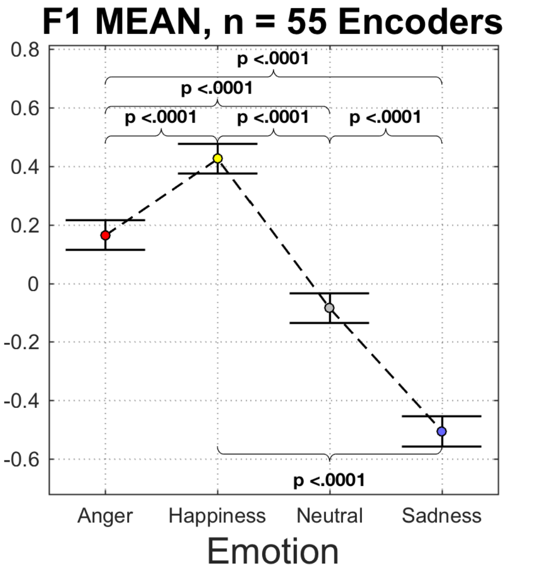

Hammarberg, n = 55 Encoders
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
1
0.5
0
-0.5
-1
Anger
Happiness
Neutral
Sadness
Emotion

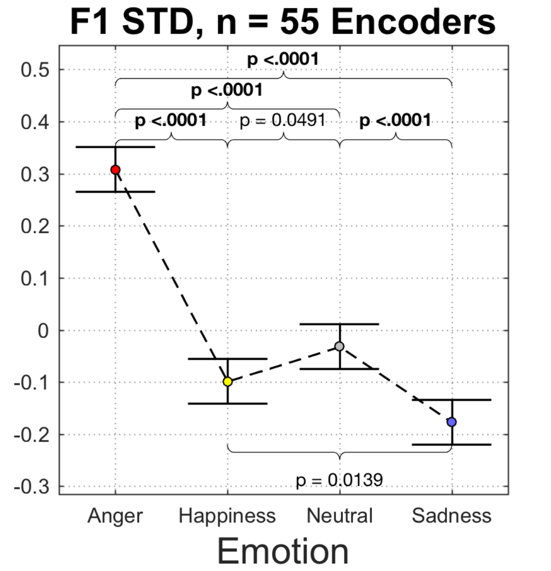

HF500, n = 55 Encoders
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
Anger
Happiness
Neutral
Sadness
Emotion

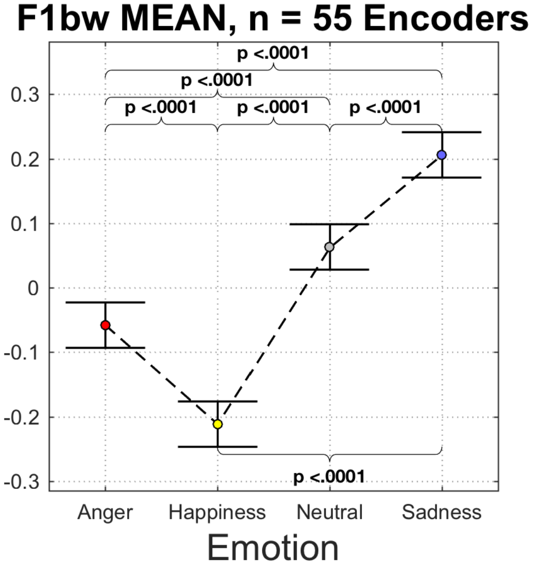

HF1000, n = 55 Encoders
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
Anger
Happiness
Neutral
Sadness
Emotion

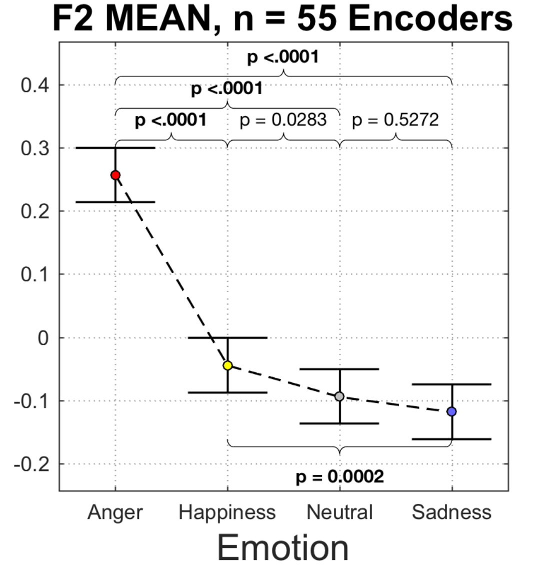

Spectral COG, n = 55 Encoders
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
0.8
0.6
0.4
0.2
0
-0.2
-0.4
Anger
Happiness
Neutral
Sadness
Emotion

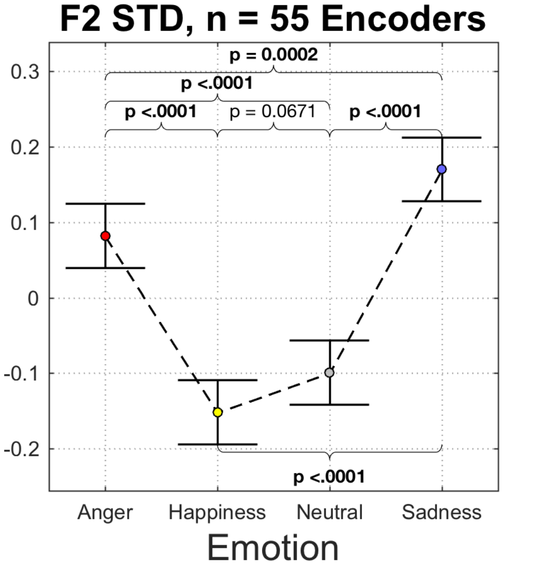

Spectral Slope, n = 55 Encoders
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
Anger
Happiness
Neutral
Sadness
Emotion

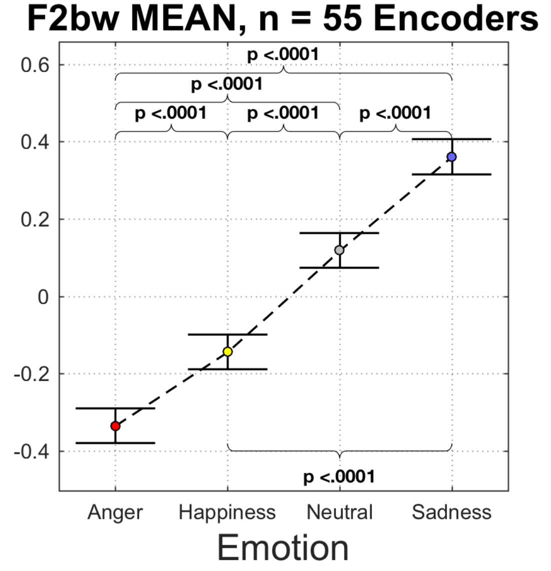

F1 MEAN, n = 55 Encoders
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
Anger
Happiness
Neutral
Sadness
Emotion


F1 STD, n = 55 Encoders
p <.0001
p <.0001
p <.0001
p = 0.0491
p <.0001
p = 0.0139
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
Anger
Happiness
Neutral
Sadness
Emotion

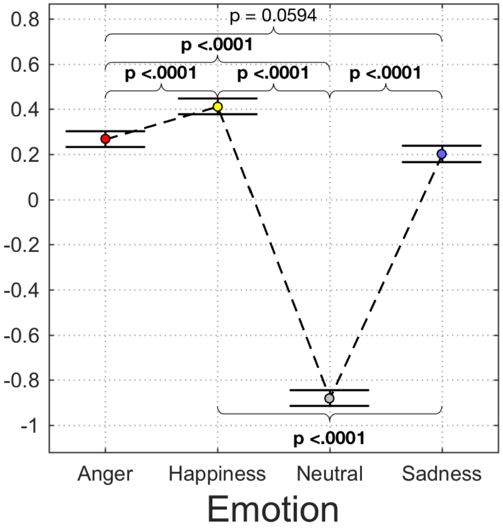

F1bw MEAN, n = 55 Encoders
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
Anger
Happiness
Neutral
Sadness
Emotion


F2 MEAN, n = 55 Encoders
p <.0001
p <.0001
p <.0001
p = 0.0283
p = 0.5272
p = 0.0002
0.4
0.3
0.2
0.1
0
-0.1
-0.2
Anger
Happiness
Neutral
Sadness
Emotion

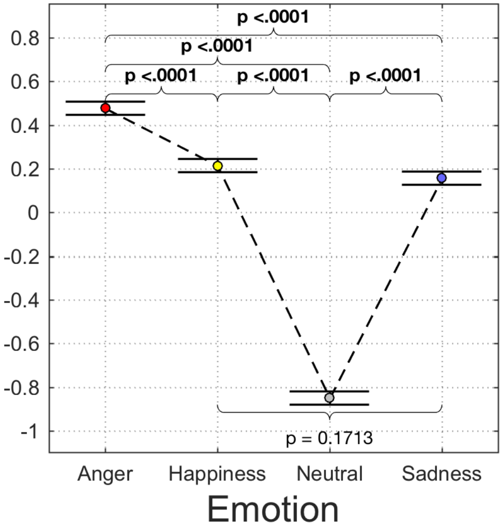

F2 STD, n = 55 Encoders
p = 0.0002
p <.0001
p <.0001
p = 0.0671
p <.0001
p <.0001
0.3
0.2
0.1
0
-0.1
-0.2
Anger
Happiness
Neutral
Sadness
Emotion


F2bw MEAN, n = 55 Encoders
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
p <.0001
0.6
0.4
0.2
0
-0.2
-0.4
Anger
Happiness
Neutral
Sadness
Emotion

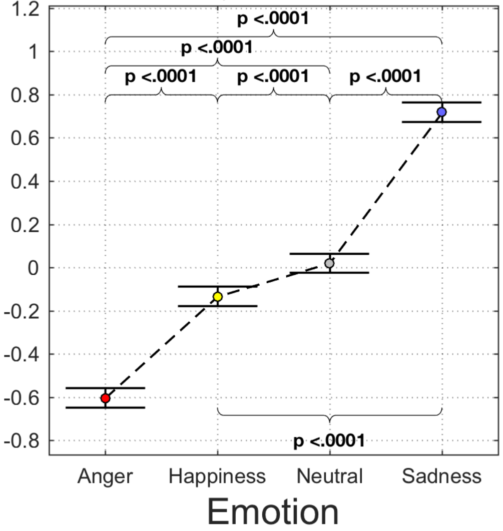
F1prec, n = 55 Encoders



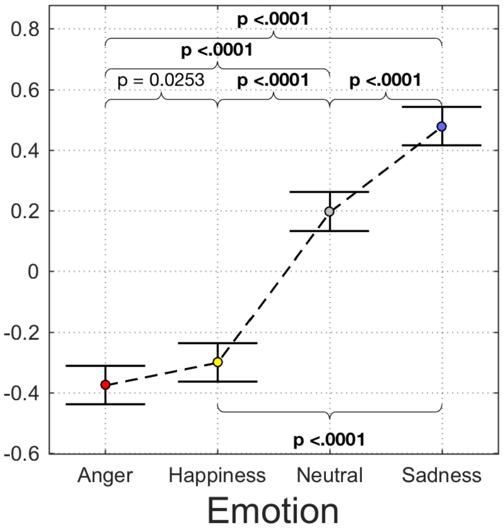
F2prec, n = 55 Encoders



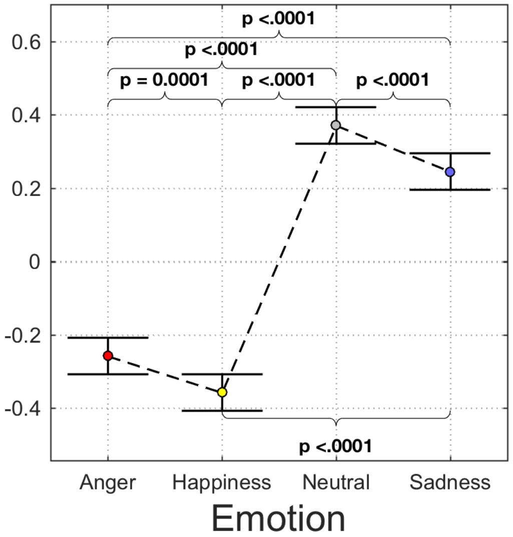
HNR MEAN, n = 55 Encoders



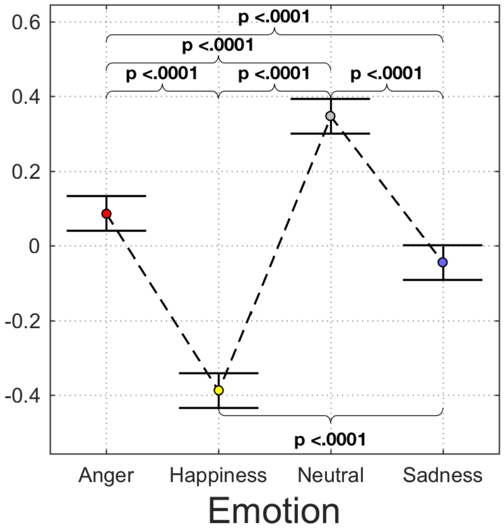
Autocorrelation, n = 55 Encoders



Jitter, n = 55 Encoders



Shimmer, n = 55 Encoders



Appendix H Participant Feedback Survey Results

Participant Feedback Survey Results

Survey Question	Range	x̄	σ
In general, how difficult was it to determine the emotion category for the stimuli?	(-2,2)	1.10	0.69
In general, how difficult was it to rate the emotional intensity of the recordings?	(-2,2)	-0.44	0.93
How authentic were the portrayals of each emotion? - Anger	(0-100)	78.47	20.60
How authentic were the portrayals of each emotion? - Happiness	(0-100)	66.17	26.25
How authentic were the portrayals of each emotion? - Sadness	(0-100)	60.07	29.59
For each emotion category, how frequently did the scripted words interfere with the task? - Anger	(0-100)	17.31	20.29
For each emotion category, how frequently did the scripted words interfere with the task? - Happiness	(0-100)	37.97	31.84
For each emotion category, how frequently did the scripted words interfere with the task? - Sadness	(0-100)	18.77	19.04
How would you rate the length of each of the following? - Overall test length	(-2,2)	-0.26	0.48
How would you rate the length of each of the following? - Question set length	(-2,2)	-0.29	0.50
How would you rate the length of each of the following? - Speech Recording length	(-2,2)	0.06	0.47
Were you fatigued or bored at any point during the test? - Fatigued	(0-100)	11.18	14.01
Were you fatigued or bored at any point during the test? - Bored	(0-100)	13.18	16.55
How would you rate the quality of the test for each of the following? - Audio quality	(-2,2)	1.53	0.66
How would you rate the quality of the test for each of the following? - User / Testing Interface	(-2,2)	1.74	0.39
How would you rate the quality of the test for each of the following? - Testing Tutorials	(-2,2)	1.81	0.62

Comments from the listening study.

Please provide any other comments or suggestions you might have.
The study was very well-designed. I believe in one of the early tutorials it may have mentioned to rank the speech on a scale from 1-10, but during the actual test the scale was from 1-100.
It would be beneficial to have male recordings as well as female recordings. I think it would be interesting to see how subjects can differentiate emotions in male voices as well as female voices.
The some of the "happy" voices sounded overly exaggerated and odd
Excellent scheduling process for the research study. Very clearly explained testing in tutorials.
Peter is super friendly and awesome! This was a very interesting study!
This was an interesting test
Very well conducted tests. UI was especially impressive.
It was fun to test my ability to match emotions to voices.
I thought that the speech could have been more authentic. For example, in the category of sadness, I did not hear a single person crying. It's very possible someone could have been crying, but I did not pick up on any sobs. I heard lots of sighing and soft-spoken speech, but to me, that just indicates distress. For anger, voices were raised, but again, it wasn't as intense as it could have been. If someone was truly angry with someone, they might be yelling at the top of their lungs to the other person, and I didn't necessarily hear that in this experiment. I definitely did hear some speech that sounded angry with raised pitch, but sometimes, instead of coming off as a person being angry, it merely comes off that the person in question was just annoyed, and has not really reached an emotional intensity that I would classify it as anger. With happiness, I think the opposite is present. I think that in certain speech, I heard over the top happiness, which I thought might have been inauthentic, especially for the text that they were saying, which could have definitely interfered (as you put above) with my perception of how happy those people really were when saying those lines of text. Other than that, the testing system worked very well with no glitches, and it seemed to do the job of recording my answers pretty well.
I thought the excerpts were somewhat short, but it was still relatively easy to distinguish the emotions in each.
Very professional.
I think that overall this was a very good experience and I don't have very many critiques. I mentioned above that I felt very slightly fatigued but quite honestly it was because I came in a bit tired.
first section is a little bit long, hard to focus the entire time
Sometimes the recording were too clearly trying to display an emotion (overselling it). I also was unsure whether I was trying to describe the emotional state that the speaker had or if I was trying to describe the emotion that the speaker was attempting to convey, i.e. the speaker is happy but trying to convey sadness. Which would I have said: happy or sad?
Easy test to follow
Excellent tutorials and opportunities for practice.