Brigham Young University
BYU ScholarsArchive

Theses and Dissertations

2021-04-01

How Many Ways Can You Vocalize Emotion? Introducing an Audio Corpus of Acted Emotion

Logan Ricks Kowallis
Brigham Young University

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

Part of the Family, Life Course, and Society Commons

BYU ScholarsArchive Citation
Kowallis, Logan Ricks, "How Many Ways Can You Vocalize Emotion? Introducing an Audio Corpus of Acted Emotion" (2021). *Theses and Dissertations*. 8921.
https://scholarsarchive.byu.edu/etd/8921

This Dissertation is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.

How Many Ways Can You Vocalize Emotion?

Introducing an Audio Corpus of

Acted Emotion

Logan Ricks Kowallis

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Bruce Brown, Chair
Dawson Hedges
Steven Luke
Deryle Lonsdale
David Wingate

Department of Psychology

Brigham Young University

Copyright © 2021 Logan Ricks Kowallis

All Rights Reserved

ABSTRACT

**How Many Ways Can You Vocalize Emotion?
Introducing an Audio Corpus of
Acted Emotion**

Logan Ricks Kowallis
Department of Psychology, Brigham Young University
Doctor of Philosophy

Emotion recognition from facial expressions has been thoroughly explored and explained through decades of research, but emotion recognition from vocal expressions has yet to be fully explained. This project builds on previous experimental approaches to create a large audio corpus of acted vocal emotion. With a large enough sample size, both in number of speakers and number of recordings per speaker, new hypotheses can be explored for differentiating emotions. Recordings from 131 subjects were collected and made available in an online corpus under a Creative Commons license. Thirteen acoustic features from 120 subjects were used as dependent variables in a MANOVA model to differentiate emotions. As a comparison, a simple neural network model was evaluated for its predictive power. Additional recordings intended to exhaust possible ways to express emotion are also explored. This new corpus matches some features found in previous studies for each of the four emotions included (anger, fear, happiness, and sadness).

Keywords: emotion, vocal emotion, acting, psychology, neural networks, artificial intelligence, voice

ACKNOWLEDGMENTS

I'd like to thank the members of my dissertation committee, who are all skilled data scientists engaged in original research at Brigham Young University. Thanks to Steven Luke and Dawson Hedges for offering me succinct wisdom as I've navigated this project. Deryle Lonsdale has been a mentor to me for several years, and I thank him for aiding me in my exploration of linguistic tools and instilling in me a deeper appreciation for computer programming and the study of language. David Wingate has inspired me to work harder to learn neural network technologies, and I thank him for his patience with me. Bruce Brown, the committee chair, has been a mentor of mine since 2009, and has done more to aid me in my career at this point than any other person. A special thanks to him.

The research assistants for this project exhibited an impressive level of resilience. Through multiple difficult technical issues and through many tedious responsibilities, they made the completion of this project within one year possible. Thanks to each research assistant: Hiroki Baba, Karlee Hemmert, Angela Larkin, Emma Lowther, Mark Rust, Alicia Settle and Jacob Tubbs.

I'd also like to acknowledge the help of the employees at the Humanities Learning Research Center, where recording sessions took place, especially their supervisor, Russell Hansen.

Table of Contents

Table of Contents

Table of Contents

Table of Contents

List of Tables

List of Figures

How Many Ways Can You Vocalize Emotion? Introducing an Audio Corpus of Acted Emotion

Imagine that a stranger calls you on the phone. You answer your phone and the person starts by saying "Guess what happened today?" Somehow, without knowing the speaker's identity, situation, what they look like, or their personality, you have an accurate guess of how the person is feeling. Relying solely on the sound of the voice coming through your phone, you accurately infer the person's emotional state. If the person were to speak slowly, quietly, and with slurred words, you might infer that the person is feeling sad. If it happened to be a video call and you could see the person's face, you would have more information, more clues, to use as you make inferences about the emotion. If you were acquainted with this person, your familiarity with their voice would also aid your ability to discern how they are feeling. If you knew that this person experienced a loss of a loved one that day, you would gain greater confidence in attributing sadness to that person. With every clue added to the person's voice, you gain a more accurate and confident inference about their emotional state. Yet people are often able to accurately attribute what emotion is being felt with no clues other than the sound the of a speaker's voice. How are people able to complete this task? There must be something that is understood by the listeners about the emotional meaning of particular characteristics in a speaker's voice, but the specifics of how these decisions are made are still not completely understood by researchers.

The area of research focused on solving this mystery refers to this ability as *vocal emotion recognition*. The adjective *vocal* is necessary because understanding emotion in voice can be considered separately from *facial emotion recognition*, a popular area of research that has advanced far beyond its vocal counterpart. The *recognition* part of the term refers to the ability to correctly perceive and judge a vocally expressed emotion. *Emotion* is a term that can have many meanings, but a general definition is moods or affective states that humans often express and are commonly attributed to themselves or others as part of daily life. A more precise definition of emotion for use in this study will be discussed later in this document. Before explaining the importance and the necessity

of the current study within the vocal emotion recognition literature, it is helpful to understand how vocal emotion research is used in practical applications.

Applications for Vocal Emotion Research

Several research fields in particular can benefit from advances in vocal emotion research, including applied linguistics, artificial intelligence, and psychotherapy.

In applied linguistics, speech recognition research has reached an impressive level in the state of the art, yet there are still lingering issues to be resolved when it comes to semantics. For example, digital assistants, such as Apple's Siri, Microsoft's Cortana, and Amazon's Alexa, can often accurately transcribe your spoken words, but they struggle to accurately infer intentions of the speaker unless the speaker speaks very direct commands. Programmers might try to manually program semantic information from common idioms or common expressions when a person is emotional to help digital assistants improve their assistive functions, but it may be far easier to improve artificial intelligence (AI) to handle conversations the way a human would. Humans infer meaning from many sources of evidence, which I referred to as clues earlier. Building an AI that is able to infer how features in a person's voice affect their intended meaning could lead to much more accurate understanding of a person's intentions and directions. For my idiom example, a person may express something positive in their chosen words, but the tone of the person's voice could make it obvious to the listener that the speaker is feeling quite the opposite, maybe frustrated or dejected. If digital assistants had this skill of discerning emotional states, they could give more timely advice and information. They could also disambiguate statements whose meaning depend on what emotion is being expressed at that moment, such as "I'm fine" signals different meaning depending on the tone of voice, and could indicate a person is feeling happy, frustrated, afraid, or many other emotions. This applied linguistics example intersects the field of computer science, but this next example intersects computer science as it relates to psychology, in that it is desirable for artificial intelligence systems to have the feel of being more human.

One application of artificial intelligence is *conversation agents*, which are computerized systems design to have conversations with human users. In an attempt to

create more genuine interactions with users, conversation agents attempt to mimic natural human traits, but their attempts at vocal emotion are still not convincing. While text-based conversation agents have improved markedly in recent years (for example, see Zhou et al., 2018), speech-based agents have lagged behind. This is not surprising, given that the vocal cues of many emotions have yet to be disentangled by researchers. If we understood the acoustic components of each emotion thoroughly, we would be able to apply those components in speech production engines to create believable emotional speech by speech agents. Additionally, AI could both identify distress or anger and help de-escalate a person's heightened emotions to help them avoid harming themselves or others and employ sympathetic-sounding vocal patterns to assist in this process.

A third application of emotion recognition research is in the area of psychotherapy for those who struggle with emotion expression or recognition. Many disorders found in DSM-5 include an emotional component. For example, "deficits in social-emotional reciprocity" is a required diagnostic criterion for diagnosing autism spectrum disorder (American Psychiatric Association, 2013). Currently, a common behavioral approach is to model appropriate expression repeatedly for the client to mimic or recognize. Since researchers have yet to identify many vocal features that differentiate emotions, how can we be certain that the clinician is demonstrating the correct features for a client to learn? If we knew precisely what it was in the voice that identified it as angry as opposed to fearful, for example, we could clearly communicate those features to clients and create training materials to precisely target training for each feature.

Previous Research in Vocal Emotion

Having established the value of understanding vocal emotion, this section focuses on what issues are preventing researchers from reaching this understanding. Several studies about vocal emotion recognition will illustrate these issues and the need for this dissertation study.

Over three decades ago, Scherer (1986) reviewed vocal emotion research and identified an important trend: Although listeners can recognize the correct emotion in a voice at much better than chance rates, no analysis of the speakers' voices had yet to fully

account for those recognition rates with features identified in the audio recordings of the voice. These features are often called *acoustic features* because they convey specific information about the physical properties of the sounds recorded from a person's voice.

In the decades that followed, other attempts were made to account for this ability to recognize emotion in voice. One of the more elaborate attempts to understand emotion in voice came from Scherer himself, along with one colleague (Banse & Scherer, 1996). In this study, 14 emotions were analyzed acoustically and those acoustic parameters were used to predict listener recognition rates, with limited levels of success. Understanding of vocal emotion recognition was still incomplete by the end of the 20th century.

Spackman, Brown, and Otto (2009) continued in this line of research, focusing on four emotions rather than 14, and found similar results: Although tests of main effects show that emotions are significantly different from each other in expression, a single pattern could not be identified that predicts all expressions of one emotion.

A graduate dissertation by Brown's student, Lauritzen (2009), expanded on the previous results with potential acoustic profiles for each acoustic domain, but came to the same conclusion as the Spackman, Brown, and Otto study (2009): Each speaker seemed to have a unique acoustic profile which differed from that of the other speakers. This unique profile problem can be found in any study of vocal emotion by separating the within-speaker variance and comparing it to the between-subject variance. The within-subject variance is larger. In other words, the acoustic features vary a lot from speaker to speaker, which seems incompatible with the finding that listeners are often correct in identifying the emotion being expressed.

In 2018, I completed a project that explored the effect of speaker on predicting emotions in voice (Kowallis, 2018). Using the same type of lens model design as Lauritzen (2009) and using his original 8 subjects' data in addition to data from an additional 9 subjects that went unused from that study's sample, I systematically varied the lens model design to use all of the speaker and listener characteristics reported to see which changes affected acoustic variables' ability to mediate listeners' recognition of speakers' acted emotions. The current study was designed with that project's results

considered. Subsetting the grouped data into all possible *listener* groupings does not affect results, but subsetting the data by *speakers* drastically affects results. When only one speaker is considered at a time, the acoustic features of that person's voice were very consistent, leading to consistently high mediation indices for acoustics predicting the recognition relationship. However, when any speakers were grouped in an analysis, even in groups of only four speakers, the mediation indices drastically dropped. When all 16 speakers were included in the analysis, in a simple design that is sometimes found in emotion studies, the mediation index for ratio of mediation effect to total effect fell to between .05 and .20, meaning that only 5-20% of the pattern of recognition by listeners could be accounted for by the total pattern of acoustic variables. Going from nearly 100% average mediation for a single speaker, to an average of 40% of the effect mediated for four speakers, to an average of 10% mediated in 16 speakers demonstrates that the variation among speakers is critically important in understanding how vocal emotion recognition works. Somehow people recognize emotion accurately despite speakers having widely-varying acoustic properties expressed in their voices.

Dissertation Study

Having established the issues in vocal emotion research, I can now introduce how my study goes beyond previous research in an attempt to understand vocal emotion.

Goals of the Study

The overall goal of this study is to help discover how people can express and recognize emotion in voice. Six specific goals were attempted in efforts to accomplish the overall goal:

1. Replicate the task conditions and emotions found in Spackman, Brown, and Otto (2009) and Lauritzen (2009)
2. Collect additional audio data according to new task conditions that are meant to further explore possible variations of emotional expressions
3. Collect a large audio corpus of emotion in a controlled recording environment
4. Host the audio corpus on the Internet for any researcher to use for their studies and projects

5. Create a tutorial for how to use the audio corpus for analyses about emotion expression
6. Compare two analytic approaches from two different fields: MANOVA modeling found in existing emotion studies in psychology and unsupervised deep neural networks taken from computer science

1. Replicate the task conditions and emotions found in Spackman, Brown, and Otto (2009) and Lauritzen (2009). Designing a study to resolve how emotion recognition works is difficult because all past attempts to resolve it have met with limited success. This has caused a lack of consensus for designing studies. The task of designing a study is further complicated because vocal expressions that are used as stimuli for recognition studies are themselves not well understood. Scherer (2013) details how there is no ideal way to study emotion recognition because real emotion is rarely ever expressed in consistent settings to allow for them to be recorded and used as stimuli in an experiment, but acted emotions can trade authenticity and variety for consistency. Because controlled experiments account for and control many confounds that a naturalistic study of vocal emotion cannot, the present study relies on an experimental design.

An alternative way to study emotion recognition is to remove the language component entirely. For example, one study used the Montreal Affective Voices (MAV) corpus to test recognition rates of subjects (Vasconcelos et al., 2017). The MAV corpus is large and impressive, but studies that rely on it are limited in generalizability. Because the sounds being vocalized in the corpus are all without words, being able to predict emotion from the corpus is also limited to situations where people create sounds with their vocal tracts but do not use words. An example of a situation when a person relies only on a nonverbal utterance to know what emotion a person is feeling would be interpreting a sigh as sadness, but there are many more situations not accounted for, like my previous examples of talking to a stranger in a phone call or having a digital assistant interpret your intentions accurately from a stated request. I choose to use spoken emotional portrayals so that my results can be generalized to situations involving speech.

While using words, though, there are many possible speech examples to use. A single script used for all acted emotional portrayals is a good way to control for linguistic variations in speech. Studies that use a single script utilize an ambiguous situation described in the first person voice so that the speakers can convincingly speak for all of the emotions used without the word choices and situation affecting listeners' judgments of what emotions were expressed. For example, Feldstein (1964) uses a script that describes a typical situation: boarding a public bus and noticing the people around you there. Spackman, Brown and Otto (2009) opted for a scenario involving a university student entering a classroom on the first day of classes, which may have better suited their university-student speakers and listeners than the bus scenario. Both of these studies used a first-person perspective as the speaker described events in the past tense. Many other studies use shorter scripts, such as focusing on one word that is present in all recordings and ignoring all of the other words spoken when the recordings are heard and when the data are analyzed. I believe this removes too much ecological validity from the study because hearing only one word and making a judgment of emotion expressed by the speaker is not as common a real-world situation as having a sentence or more heard before making a judgment. In other words, we lose the ability to generalize to typical monologues and conversations if we limit recordings to single words, so this study uses full sentences instead.

One study found that authenticity does not appear to affect recognition in a systematic way, with some emotions being more believable when acted and others more believable when real, yet all of them are recognized with about the same accuracy (Jürgens et al., 2013). Another study challenged that view and found that acted emotion yields a different recognition pattern compared to naturalistic expressions of emotion (Juslin et al., 2018). Also, the acting skill of speakers has been demonstrated to play an insignificant role in recognition rates of listeners in other studies (e.g., Jürgens et al., 2015; Spackman et al., 2009). In other words, untrained actors can express emotions just as well as trained and professional actors for the sake of a recognition task, but there are differences between real and acted emotion. This study uses acted emotion with untrained actors, relying on

the aforementioned research on their efforts being as good as professional actors, although that is a disputed effect. Holding the experiment in a controlled environment helps control for confounds that come from varying recording conditions.

The present study uses this controlled experiment approach with acted emotion, along with some important changes to design that distinguishes this study from previous studies. Several changes to design allow for the exploration of this theory: that each discrete emotion is expressed through multiple specific features from a possible set of distinguishable features. The aforementioned studies that use discrete emotions have a non-varying construct for each emotion, i.e. anger is either being portrayed or it is not. There is no room for understanding variations in expression unless the study uses other categories as discrete alternatives. For example, Banse and Scherer (1996) instructed speakers to express elation and happiness separately, thus measuring two related emotions that are similar to each other but distinct from the other emotions in the study.

Trying to study emotion recognition with continuously defined emotions rather than discrete emotions is not a desirable solution to the rigidity of past studies. Although continuous emotion spectra have been a popular topic among researchers in psychology, a lay-person's common experience with attributing a categorical emotion to another person is far detached from these theories. People use words to describe complex feelings, attributing a lot of meaning to these discrete terms, such as happy, angry, frustrated, depressed, sad, etc. These words are not expressed as a composite of several numeric values interpreted on continuous spectra. While the underlying mechanisms of feeling, expressing, and recognizing emotions might be best explained by continuous emotion spectra, there are endless ways of conceptualizing emotion as continuous and no clear way to map those theories onto humanity's discrete judgments of voices in a perceptual study. This study uses discrete emotions because it is a simple way to understand subjects' expressions and judgments based on the discrete words that are commonly used for attributing emotion.

Paul Ekman and his colleagues came up with basic emotions that are a good place to start for exploring common profiles of emotions (Ekman & Friesen, 1971). They found

evidence that facial expressions cross-culturally indicate consistently similar expressions of each individual emotion. This finding was based on subjects judging previously created stimuli, pictures of people's faces, as indicating some feeling. This experimental paradigm focused on expressions as stimuli and perceivers as subjects is emulated in many of the vocal emotion studies reviewed in this paper, though few vocal emotion studies focus on cross-cultural differences. I will not include cross-cultural hypotheses in my study because it is beyond the scope of this dissertation study and its mostly homogenous sample, though future studies can use my experimental design to collect new recordings for comparison to the recordings in this study.

Another relevant finding from Ekman was his survey sent to emotion researchers (Ekman, 2016). Most emotion researchers surveyed believe in basic emotions, i.e. emotions that are universal across cultures and time, and also agree on the identity of several of these emotions as basic. My simplest modeling approach will be utilized in describing the differences between four of Ekman's basic emotions: anger, fear, happiness, and sadness. In addition to these four, a fifth category of recordings, termed *neutral*, task the subjects with speaking in their normal speaking voice without attempting to act particularly emotional. This neutral condition helps to act as a baseline for what each subject's voice sounds like naturally, and is valuable when building a model to detect presence or absence of emotion in a person's voice. These are the same four emotions used in Spackman, Brown, and Otto (2009) and Lauritzen (2009), so the results from this study can be compared to their results.

2. Collect additional audio data according to new task conditions that are meant to further explore possible variations of emotional expressions. To help ensure that enough variations in recordings are present in the dataset for each subject, another new design feature is included in this study: a new task that requires variety in expressions. Subjects are asked for their last acted emotion to continue on and "act as many different ways you can think of." Instructing subjects to act in as many ways as they can imagine will exhaust their skill and understanding of emotion and guarantee variety in expression, in contrast to the usual study design that leaves variety of

expression up to the subject. Without this specific direction for variety, some subjects can exhibit a rehearsal mindset, repeating as closely as they can each portrayal in an effort to sound the same each time. Other subjects can experiment and vary their expression of emotions, or even have an epiphany partway through a session after repeating an approach to speaking only to vary drastically from that approach for the rest of the session. This additional, varied set of recordings is a first of its kind and will aid in the creation of better theories of emotions by offering contrasting expressions within each emotion.

3. Collect a large audio corpus of emotion in a controlled recording environment. I hypothesize that there will be more than one distinct pattern of expression found for each emotion that is meaningful. To help ensure that enough exemplars of each feature are identified to establish a pattern, a third change to existing study designs must be made: A much larger sample of audio recordings. Increasing the number of portrayals per emotion per subject by an order of magnitude should be sufficient for commonly-used features. Instead of the two portrayals used by Lauritzen (2009), this study has, when recordings meet certain quality standards, 20-30 portrayals per subject per emotion. The number of subjects should also increase substantially to allow more possibility for variety of expressions. The original plan for this study includes an order of magnitude greater than a study like Lauritzen (2009): 192 subjects planned instead of that study's 16 subjects. Circumstances described below led to 128 being chosen as the revised target number of subjects, with 131 subjects' data collected. Some features were expected be more commonly expressed than others, and some features may only exist for one subject at all, but the groundwork of a large corpus of recordings is a start that can guide future research to explore features that were overlooked or underrepresented.

Although the larger sample helps with some aspects of vocal variety and adds statistical power, the sample is mostly homogenous in demographics. Most notably, this study uses a typical undergraduate psychology student sample, and therefore lacks variety in the ages of subjects. This limitation is not as problematic as it may first appear, though.

Methods for testing accuracy of vocal emotion recognition vary over the lifespan, but the trend of recognition accuracy is roughly linear over the lifespan. Small children struggle to recognize motivations and emotions in others in anything other than the simplest of choice paradigms (for example, Bhullar, 2007), but progress rapidly in those abilities through the elementary school ages of 5 to 10 (Sauter et al., 2013; Allgood & Heaton, 2015). Adolescents still make more mistakes in recognition than adults, but they gradually reach that level of recognition, on average, as they age and mature (Morningstar et al., 2018). Older adults are significantly worse than younger adults at recognizing emotion (Laukka & Juslin, 2007), but gradual hearing loss may account for this difference. Young adults are commonly used as subjects in acted emotion studies because they are the most common age group attending universities, and universities are a common source of convenience samples in psychology research. They should, on average, have a greater mastery of emotion expression than all ages of children, and approach the upper limit of human ability for recognition and expression of emotion. Strength of recognition continues until age-related health issues cause adults' recognition skills to plummet below those of young- and middle-aged adults. Therefore my sample that consists of mostly young adult university students is justified as they will be near their own lifespans' peak mastery of emotional expression. Expression has not been thoroughly studied in the lifespan the way that recognition has, but there is significant overlap in expression and recognition skills because they both rely on a cognitive representation of each emotion that is accurate to common beliefs about that emotion.

4. Host the audio corpus on the Internet for any researcher to use for their studies and projects. Given the size and nature of this study's corpus, it will offer a new and unique resource for researchers. Experimentally controlled recordings for vocal emotion are rarer than uncontrolled conditions, and even rarer are corpora that include many different emotions. The controlled lexicon by using a script for all recordings helps control for variance due to different words' different acoustic sounds. Although a handful of studies already mentioned above use a similar paradigm for collecting emotional portrayals, none of those datasets are widely available on the Internet for anyone to use.

Datasets have also been created from existing, heterogeneous recording sources of data that are then treated as emotion portrayals, but as I will discuss below, this approach causes more problems than it solves. Instead, a corpus of recordings deliberately created through acting and performed in a silent sound stage with professional equipment is the approach chosen for the present study. Having a large corpus of these high-quality recordings can aid several practical applications that are cross-disciplinary. Having my corpus freely available on the Internet is an attempt facilitate collaboration across disciplines as well as answer the questions posed by psychologists about vocal emotion recognition.

5. Create a tutorial for how to use the audio corpus for analyses about emotion expression. This study also aims to serve as a tutorial for how to use the corpus of recordings to conduct research. All analyses and data processing steps were systematically tracked in documents that are preserved as files hosted alongside the audio files of the corpus. These documents serve as a guide for anyone wanting to analyze the corpus for themselves.

6. Compare two analytic approaches from two different fields: MANOVA modeling found in existing emotion studies in psychology and unsupervised deep neural networks taken from computer science. In addition to data collaboration, this study aims highlight the strengths of various analytic techniques that could answer questions about vocal emotion recognition. Using the multivariate version of a common technique used in psychology studies, Analysis of Variance (ANOVA), I will highlight how to detect overall differences between emotions. Also, accounting for within-subjects variability could drastically improve the ability to predict emotion from acoustic properties when compared to simpler statistical models. As a comparison, an unsupervised neural network approach from the field of computer science illustrates that a data-driven approach can offer new avenues for finding patterns for emotions that ANOVA models have failed to find. Complex networks created from all of the available data instead of pre-selected features could offer new insights.

Hypotheses

Hypothesis 1: Emotions differ on acoustic measures. The null hypothesis for the MANOVA model is that all four emotions (anger, fear, happiness, and sadness) will have equal means for each emotion, and that the overall pattern across all variables is the same for each emotion. The alternative hypothesis is that emotions are not all equal in means and that the multivariate statistics show that emotions vary from each other overall.

Hypothesis 2: Using some audio data to build a model, you can accurately identify the emotion in newly presented audio data. The prediction for the neural networks model is that the prediction accuracy of the test data will be greater than the rate of accuracy at a chance level (20% when using all five emotions).

Hypothesis 3: There will be a clear mode for unique emotional expressions. This hypothesis is harder to define because no study has previously investigated unique ways to act emotion by individuals, but most people expressing around the same number of unique expressions seems reasonable if most people have the skill of expressing emotion effectively. I choose five as my expected statistical mode for number of unique expressions per emotion.

Method

Figure 1 depicts the data collection process. The sample includes BYU students who volunteer through BYU's SONA research system. Students chose among alternatives to participate in this study for course credit as compensation for their research participation. SONA study participation is largely limited to students who take psychology courses, and the majority of psychology course students major in psychology.

Procedure

Demographic questions given to subjects were chosen for their relevance to vocal properties: age, sex, home country, first language, and home state or territory if their home country is the United States of America. Although the current study does not explore the effects of these demographics on the subjects' recordings, they were collected so that future studies can benefit from using them in their emotion recognition models.

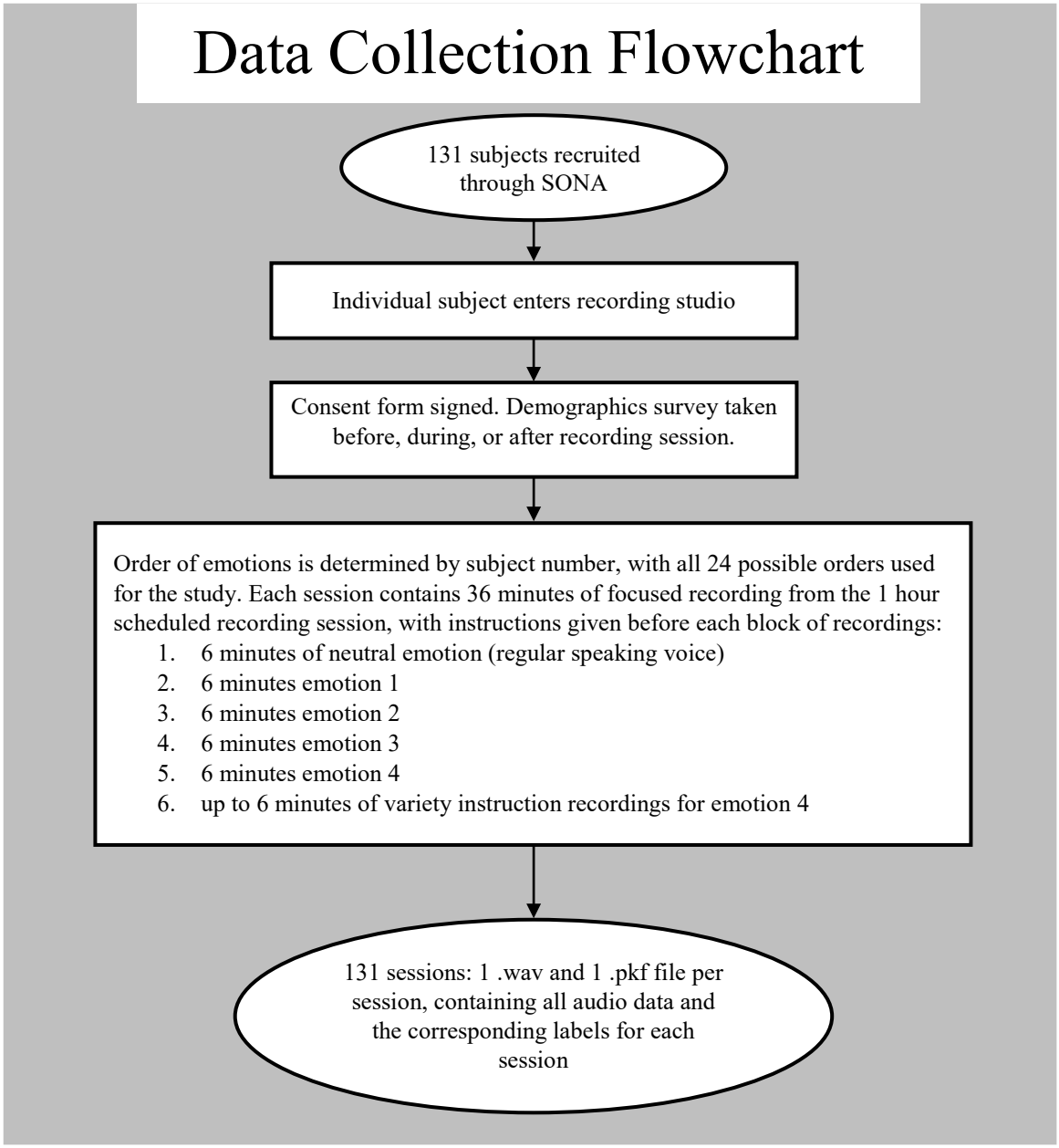

Figure 1. Data collection flowchart.

In total, data from 131 subjects are included in this study.

The recordings were collected in a professional recording-booth environment, with standards in place to ensure consistent conditions for all recordings. A research assistant observed during recording and ensured that the subjects remained within two inches of the pop filter, which was itself as close to the microphone without touching it.

Research assistants received correlation training in the months leading up to the data collection period to help ensure that they followed the written protocol for recording procedures precisely. Subjects were given six minutes to read the script out loud as many times as possible. They were told to prioritize accuracy in reading all of the words in order over number of recordings completed. Recordings where the subject deviated from the script were identified by research assistants and later placed in separate folders for the corpus. All recordings that were unfinished or that failed to maintain an acted voice throughout were deleted and not included anywhere in the corpus. All statements that were not portrayals of the script, such as conversations between subjects and the research assistants, were also deleted. The microphone used for recording was a Shure KSM32 connected to a PreSonus Studio 1824c audio mixer, which was then connected to a Mac Mini. Recordings were encoded as PCM 16-bit 44.1 kHz mono .wav audio using the recording software Adobe Audition (Adobe Audition, version 12).

My choice of script seeks to describe an ambiguous situation that can be more general than previous studies that used as contexts public transit (Feldstein, 1964) or attending a university (Spackman et al., 2009; Lauritzen, 2009). A single script needs to be used for all emotions so that linguistic variation is accounted for, but the script itself must be free of emotionally charged words that favor one emotion over others. Furthermore, the situation described must include enough ambiguity to not make one or a few emotions more likely as a response to the situation than the other emotions. There are limitations to any situation described because not everyone will interpret a situation in the same way, but it is understood that this one corpus of recordings will not generalize to all possible situations for which humans recognize emotions. I wrote my own script to describe a fairly general situation that many people have experienced. Here is the script read by subjects:

> I went for a walk. It seems that I have new neighbors moving in. You wouldn't believe what happened, it was ... interesting. If you were there, you'd have the same reaction. Why don't you come with me next time I see them?

Subjects were prompted to act as best they could for each emotion, and were

reminded that even if they feel they are not skillful at acting their portrayals would be useful. Subjects were also told to focus first on accuracy, then second on acting genuinely, because the analysis required them to speak every word of the script. Each subject began with neutral trials, being told to read the script without any particular emotion, in their normal speaking voice. This block of trials was timed by the research assistant to ensure that it lasted for six minutes. All subsequent blocks of trials were also six minutes each, with timing also beginning after instructions were given, at the onset of the first attempted portrayal for that block.

After the neutral block, each subject was given instructions for his or her first predetermined emotion block. The four emotions were assigned by subject number ahead of time, with all 24 possible permutations of four emotions being used in this study, and the number of subjects for each order was balanced.

After proceeding to act for each of the four emotions, subjects were instructed that the task was changing for the last block: they would be required to continue their last completed block's emotion for portrayals, but this time varying each portrayal to always act it differently for all the subsequent portrayals. They were told to make sure that each portrayal was truly different for each trial, with no repeats. The research assistant was able to help them judge if portrayals were the same if they were unsure by consulting the waveform and using their best judgment for what sounded the same, but were otherwise trained to not direct the subjects in their acting. Some subjects ended up needing to be interrupted by confronting them about their recordings sounding too similar after many trials. These situations were not carefully tracked, though they may be apparent by their high number of recordings compared to the average for the rich dataset. Subjects were told that they could reuse any of their approaches for expressing the emotion that took place in the previous block of recordings for that emotion. Once the subject indicated that they were out of ideas, they were pressed one time by the research assistant to keep attempting more after some time thinking, but ultimately is the subject who decided if their recordings started sounding too similar and when this portion of the session was finished. Six minutes was allotted to this section, but no subjects ever reached the time

limit and their recordings can be interpreted as their limit on ways they could act out that emotion.

Irregularities During Data Collection

Recording began in September 2019. Partway into the study, however, issues arose. The worst of these issues was that a Mac OS X update broke driver compatibility with the audio mixer being used by the recording computer. Unfortunately, this issue meant that audio monitoring could not occur simultaneously with audio recording, so the research assistant present at each recording session could not give feedback to participants while recording. Because of this issue, all recordings were checked after the study completed so that recordings not meeting quality standards could be identified and removed. In a few cases, subjects did not meet the strict criteria for inclusion in the dissertation analysis because of issues that were not corrected because there was no audio monitoring during recording. These cases are explained in the data processing section below.

Project delays made it impossible to collect data from 192 subjects as originally planned, so a new target of 128 subjects was created. This value allows for counterbalancing of emotion order to still use all possible orders with an equal number of subjects per order. Some data had already been identified as unusable or sparse, so the final subjects scheduled were given subject numbers to complete the counterbalancing of emotion order as well as possible. Numbering found in labels of files in the corpus is according to these subject numbers, so some numbers were skipped when assigning. The subject numbers range from 001 to 142, but there are only 131 subjects included in the corpus.

One subject did not complete all of the questions on the demographics survey. This subject's data are still included in the corpus.

Data Processing

Figure 2 details the data processing steps and the exclusion criteria for data in this study's analysis. Recordings from 131 subjects were split into three directories: corpus

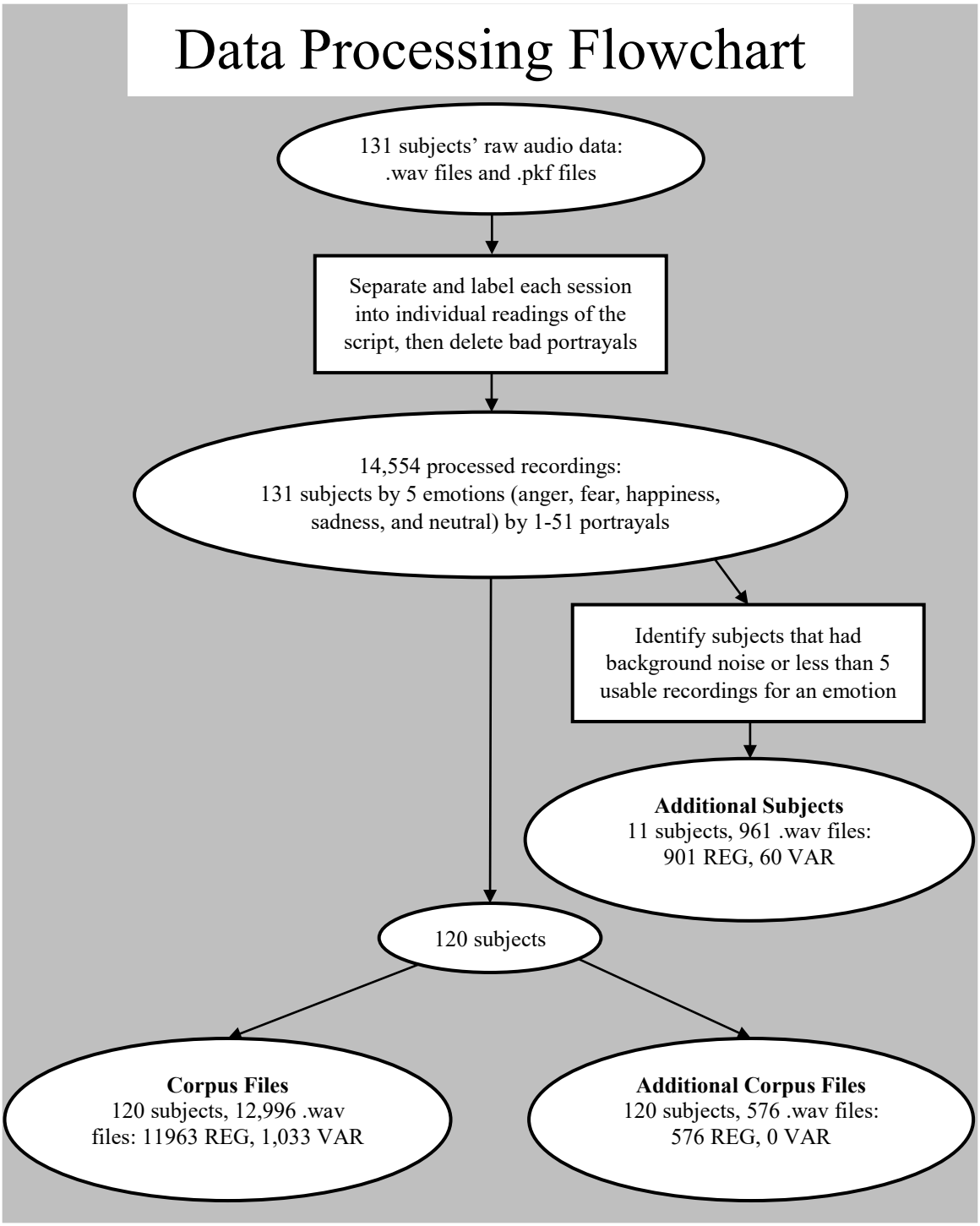

Figure 2. Data processing flowchart.
Note. REG: Regular Instructions. VAR: Variety Instructions.

recordings, additional corpus recordings, and additional subjects. *Corpus recordings* contains data from 120 subjects that have good recording data for each of the five emotions and were collected according to correct recording protocol. *Additional corpus recordings* contains examples of useful but not perfect recordings from the 120 subjects found in the corpus recordings folder. Reasons for being included in additional corpus recordings: in character for the entire recording, reach the end of reading the script, can have noises that obscure portions of the spoken script but not substantial portions (e.g., bumped microphone, coughed, or stuttered), and can have errors in reading the script but still sounds natural (e.g., neighbor instead of neighbors, would not instead of wouldn't, the next time instead of next time). Most common reasons for inclusion in the additional corpus files is for minor deviations from the script in otherwise standard recordings. Recordings were excluded from the forced alignment analysis because they were expected to create erroneous time markings for words and phones. They may be just as useful as the corpus files, but could not be included in this analysis, and so they are marked as additional in the corpus.

Another folder called *additional subjects* was created for recordings in which a global issue caused more than one or a few recordings to be faulty. Out of 131 total subjects, 11 had serious issues that prevented them from being used in whole, so they were excluded from analysis but may still be useful. Because the two planned analyses require examples from each emotion for each subject, eight subjects who lacked at least five usable recordings for any one of the five emotions were not analyzed, and their recordings were placed into the additional subjects folder. Two subjects were not properly recorded according to the protocol and were instructed to only complete one reading of the script in each emotion, they were also excluded. One subject's recording session was marred by intrusive background noise that resulted from the recording protocol not being followed, and was excluded from analysis as well.

Designs of Analyses

Three separate analyses from different research traditions are described below:

1.) A linguistic analysis, which uses theory-based extraction of acoustic features and

null-hypothesis testing

2.) A neural network analysis, which uses few assumptions to infer a model from raw audio data, then test that model using data that was not used in creating the model
3.) An exploratory analysis, relying on visual and descriptive representations of data to help create new theories of emotion

Linguistic Analysis Design

Figure 3 details the process of processing and analyzing the audio recordings for this analysis. The process can be subdivided into three main analysis steps:

1.) Extracting timestamps for each word and sound via *forced alignment*
2.) Creating 13 acoustic feature variables with one value per recording
3.) Using a MANOVA model to test the null hypothesis that all emotions have equal means for the thirteen acoustic feature variables

1. Extracting timestamps for each word and sound via forced alignment. Forced alignment is a technique designed to quickly identify and mark the timing speech sounds, with the strong assumptions of a small set of words spoken in a specific order. These assumptions help improve accuracy of identification by limiting the possible speech sounds to a small set. Software called Montreal Forced Aligner (McAuliffe et al., 2017) was used to complete the forced alignment. After alignment, 132 of the recordings were identified by the software as failing to be aligned properly. The planned approach to fix these alignments was to use altered settings during forced alignment, but this was not possible due to software bug in Montreal Forced Aligner that prevents settings from being applied, resulting in default settings always being applied for any forced alignment. It was decided that these 132 recordings would be removed from this and all other analyses to ensure that the same recordings were being compared across analyses. The completed text files marking alignment of words and phones are called TextGrid files.

2. Creating 13 acoustic feature variables with one value per recording. After alignment, 13 acoustic variables were created based on theoretically-relevant phenomena. These phenomena can be viewed as a small subset of possible acoustic features that can demonstrate how this corpus can be useful. Two features summarize entire recordings.

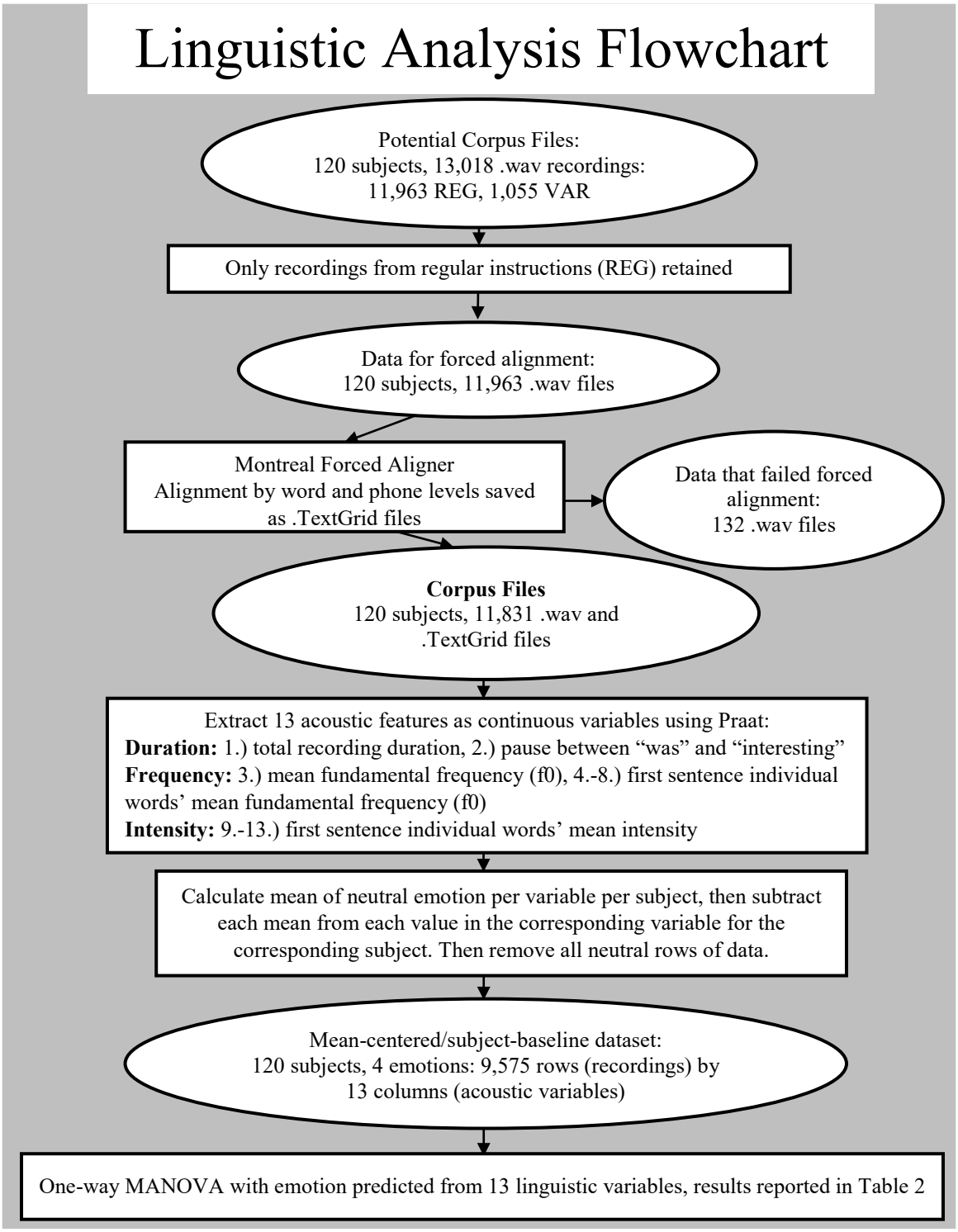

Figure 3. Linguistic analysis flowchart.
Note. REG: Regular Instructions. VAR: Variety Instructions.

These overall features, sometimes called *global*, include: total duration of the recording in seconds and mean fundamental frequency (f0) of the total recording in Hz. The other 11 features are for smaller features than global features, and include: duration of the ellipsis pause in seconds, mean fundamental frequency for five individual words (I, went, for, a, walk) in Hz, and mean intensity for five individual words (I, went, for, a, walk) in dB. Acoustic features were calculated using the acoustic analysis software Praat (Boersma & Weenink, 2019). For all feature extraction processes, the TextGrid files created by the Montreal Forced Aligner were assumed to be accurate and were used to create the variables.

Next, each variable is averaged within each emotion (neutral, anger, fear, happiness, and sadness) and then neutral's values are subtracted from each individual value found in the other emotions to create difference scores. This process allows for each subject to serve as their own baseline, and then each emotion's averaged values can be interpreted as amount of change subjects made in their voice away from their normal speaking voice. The reason for this is to prevent variation in vocal ranges due to factors other than acted emotion, such as physical differences in vocal tract size, to influence the results.

3. Using a MANOVA model to test the null hypothesis that all emotions have equal means for the thirteen acoustic feature variables. The MANOVA model is simple: one independent variable with four levels (emotion: anger, fear, happiness, sadness) predicting the thirteen difference-score dependent variables (total duration, mean f0, ellipsis pause duration, f0 for "I", f0 for "went", f0 for "for", f0 for "a", f0 for "walk", intensity for "I", intensity for "went", intensity for "for", intensity for "a", intensity for "walk"), with the error term specified as the repeated variable for subject (sample of 120 subjects, with varying number of recordings per subject per emotion). The statistical software Stata (StataCorp, version 16) was used for calculating this MANOVA model and the accompanying descriptive statistics.

Neural Network Design

In addition to the previously described statistical methods, a bottom-up approach

may help researchers differentiate emotions without relying on formulated theories of emotion. Some forms of neural networks can use raw data as initial input to extrapolate useful features that yield accurate predictions of a categorical variable. In this way, the neural network model will help offer insights into relevant features for discriminating emotions as well as a final model that can be tested for accuracy of emotion recognition.

Figure 4 details the processing steps and design of the neural network model. The neural network model uses a convolutional neural network (CNN) to quantify higher-order features into a smaller representation of the raw audio spectrograms and to introduce some noise into the data. The output of the CNN portion of the model is then run through a standard dropout procedure, zeroing out 20% of the values present. It is then sent through a simple feed-forward network comprising linearly connected layers in progressively smaller sizes. The model is created using the Python-based neural network framework, PyTorch (Paszke et al., 2019, version 1.5.1).

In pre-processing, the audio data are made commensurate in length by padding silence at the end of each recording using the command-line tool ffmpeg (ffmpeg Developers, 2020, version git-2020-04-26-1128aa8). Then a high-resolution spectrogram is created for each recording using PyTorch. This step allows for the data to be transformed from two-dimensional (time by amplitude) to three-dimensional (time by frequency band by power). These spectrograms are then randomly assigned according to an 80-20 training and test design, with 96 subjects in the training set and 24 subjects in the test set. The training spectrograms are used as input for creating the model. When the model is finished training, it is evaluated using the test set of spectrograms, yielding an accuracy statistic, which is the percentage of the recordings that have their emotion correctly identified by the model out of the total number of recordings in the test set.

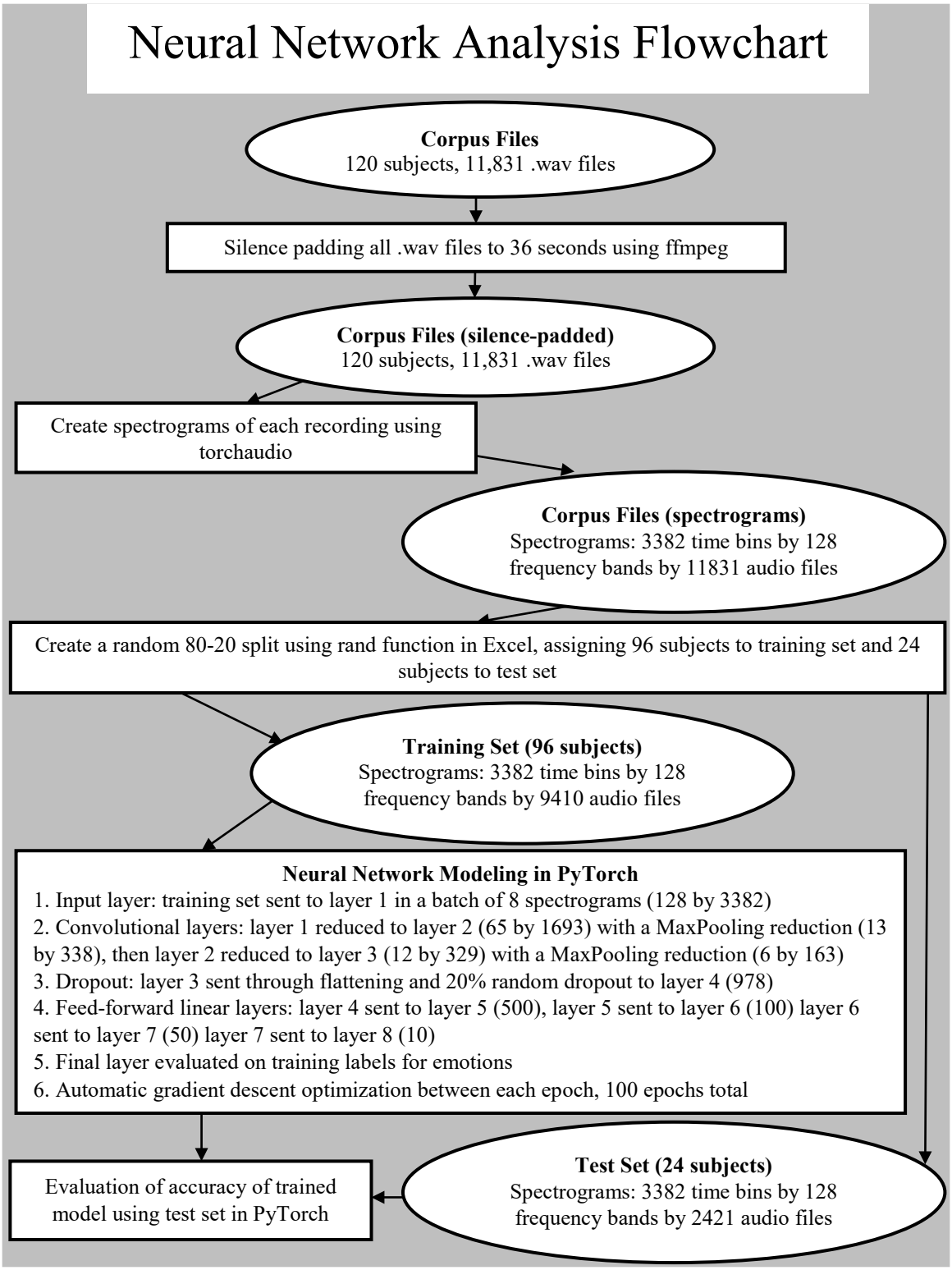

Figure 4. Neural network processing, modeling, and evaluation for prediction of emotion from audio recordings.

Exploratory Analysis Design

The exploration of the new instruction set for acted emotions is designed to include basic descriptive statistics as a starting point for developing a hypothesis-based analysis of the possible ways to express each emotion.

The modal number of recordings subjects complete before indicating that they can do no more unique portrayals in the variety block could be a good indicator of the number of unique acoustic profiles for each emotion, so it they are reported along with the range, mean, standard deviation, and median.

Results

Table 1 contains the demographic information for the sample. 95% of the student population at BYU as a whole are from the US, with roughly half of them from the western US, and 65% speak more than one language, but the official demographics do not include the percentage of students whose first language is something other than English (Brigham Young University, 2020). The sample in this study follows a similar pattern in demographics as these overall university statistics. Although the majority of subjects are from the US and have English as their first language, the sample does contain several subjects whose home country is outside the US or whose first language is not English. One subject failed to complete several items on the survey and is marked as unknown for those items.

Table 1

Demographics of Subjects Included in the Emotion Corpus

	Corpus Files (n=120)		Additional Subjects (n=11)	
mean age (standard deviation of age)	20.27 years (2.28 years)		20.73 years (3.22 years)	
gender	female	77	female	6
	male	43	male	5
first language	English	118	English	10
	Spanish	1	Spanish	1
	unknown	1		
home country	USA	116	USA	9
	Mexico	1	Mexico	1
	Ghana	1	Australia	1
	South Korea	1		
	unknown	1		
home state/territory (if home country is USA)	Arizona	10	California	2
	Arkansas	1	Massachusetts	1
	California	10	Michigan	1
	Colorado	7	Oregon	1
	Idaho	4	Utah	2
	Iowa	1	Washington	1
	Kansas	1		
	Maryland	1		
	Michigan	1		
	Minnesota	1		
	Montana	1		
	Nevada	2		
	New Hampshire	1		
	New Mexico	2		
	New York	2		
	Ohio	3		
	Oregon	3		
	Tennessee	1		
	Texas	13		
	Utah	38		
	Virginia	3		
	Washington	8		
	Wisconsin	1		
	Wyoming	1		
	unknown	1		

Figure 5 contains box-and-whisker plots of all values used in the linguistic analysis for the first three acoustic feature variables. These variables are total duration, total mean fundamental frequency, and the ellipsis pause, which can be identified by the nearby words in the script as the pause between the words “was” and “interesting”. Interesting patterns can be noted here in the original data before neutral-baselining is performed. The distribution of values for total duration of happiness and neutral recordings are very similar, and are overall shorter and narrower than the other emotions. Total duration contains many outliers on the longer side, with fear and sadness in particular having recordings that stretch on past the 30 second mark. In contrast, total mean fundamental frequency has fewer outliers in each emotion and all five emotions overlap in distribution, with similar ranges for their distributions. The pause duration between “was” and “interesting” follows an almost identical pattern to total duration, which may mean that modifying pause duration between words was a primary method employed by subjects to convey emotion as it relates to duration, though no evidence is presented in this study for modulation of duration of words spoken.

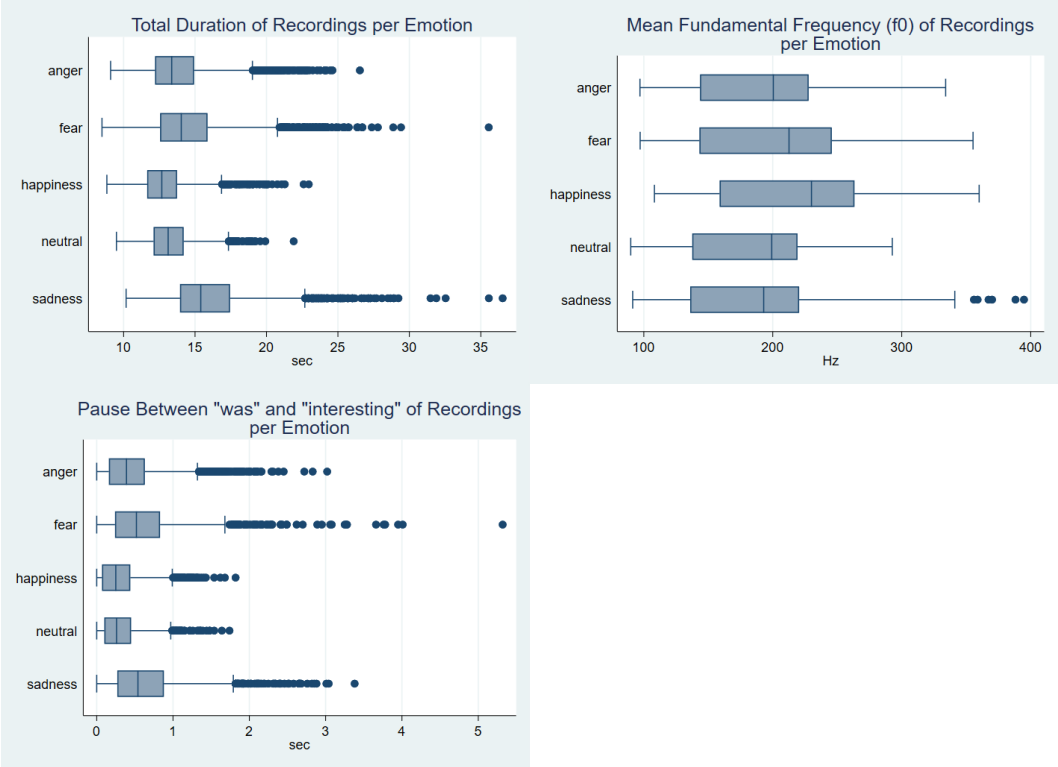

Figure 5. Box-and-whisker plots for total duration, mean fundamental frequency, and pause during the ellipsis between “was” and “interesting”, calculated at an individual recording level and with emotions as staggered plots.
Note. The shaded region is the interquartile range (IQR), whiskers are IQR multiplied by 1.5, the vertical line in the shaded region is the median, and points beyond the whiskers represent outliers.

Figure 6 contains box-and-whisker plots for all values for mean fundamental frequency in each word of the first sentence. The pattern of each of the five words closely resembles the total mean fundamental frequency plots seen in Figure 5. Although there are slight variations in distribution shape between each plot, no large variations are present.

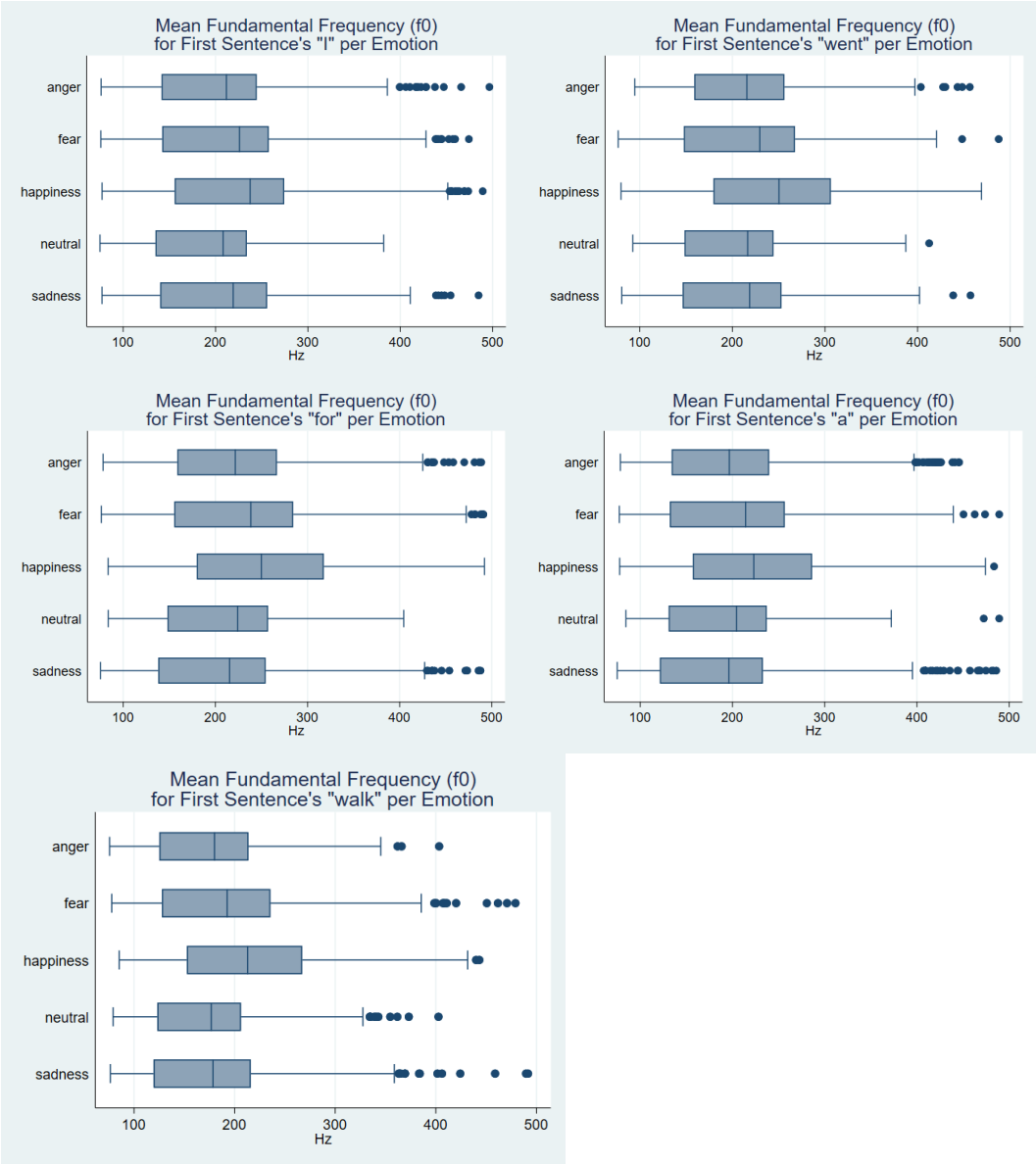

Figure 6. Box-and-whisker plots for mean fundamental frequency of each word in the first sentence calculated at an individual recording level and with emotions as staggered plots.

Note. The shaded region is the interquartile range (IQR), whiskers are IQR multiplied by 1.5, the vertical line in the shaded region is the median, and points beyond the whiskers represent outliers.

Figure 7 contains box-and-whisker plots for all values for mean intensity in each word of the first sentence. The pattern is very similar for the plots of all five words, with distributions of anger and happiness overlapping and being higher, and fear, neutral, and

sadness overlapping and being lower. It is interesting to note that for the words “I” “for” and “walk” the median intensity for all five emotions is below 40 dB, but for “went” and “a” the median intensity for all five emotions is above 40 dB, with the difference per emotion between medians of these sets of words equal to roughly 5-10 dB.

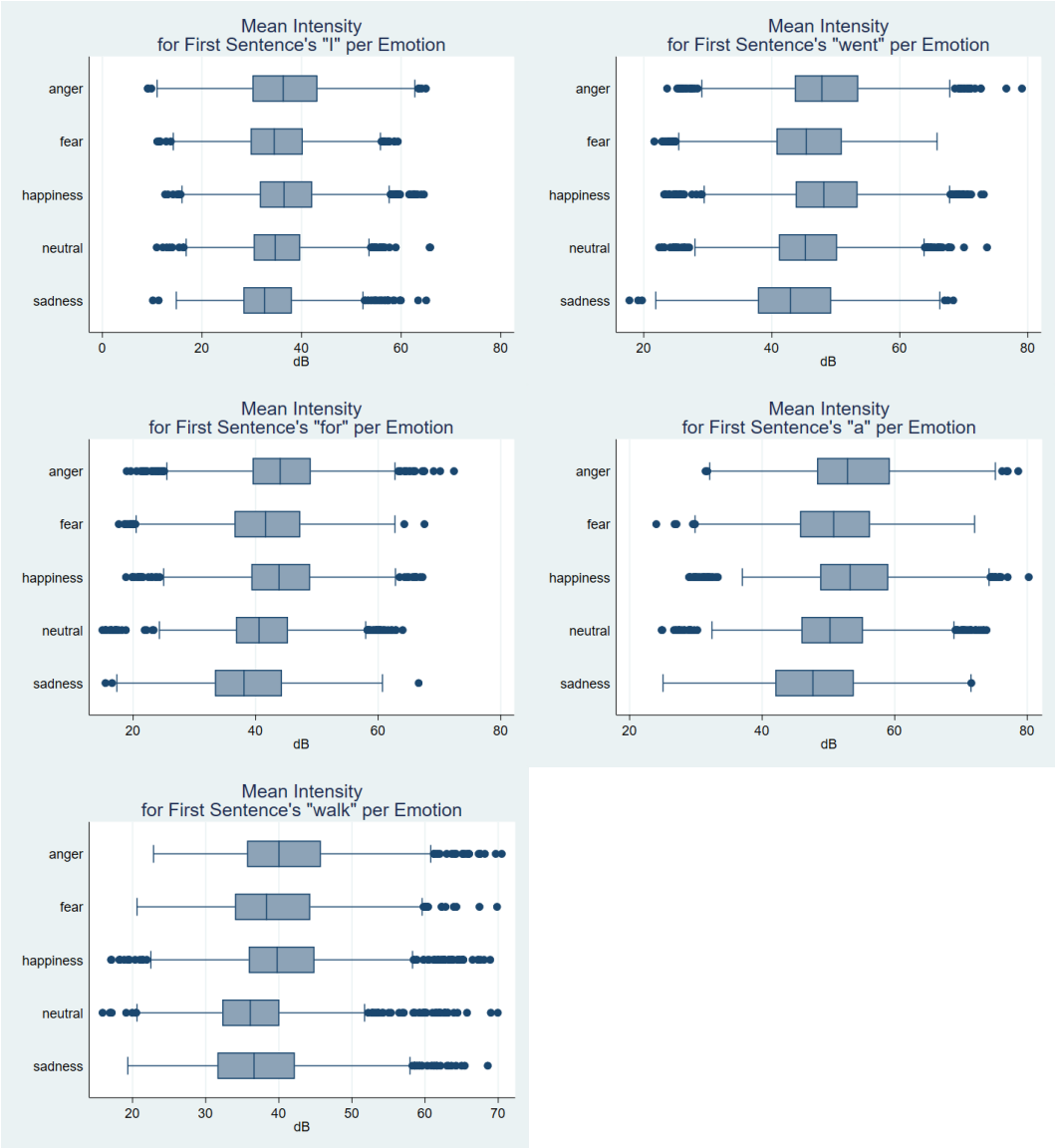

Figure 7. Box-and-whisker plots for mean intensity of each word in the first sentence calculated at an individual recording level and with emotions as staggered plots.
Note. The shaded region is the interquartile range (IQR), whiskers are IQR multiplied by 1.5, the vertical line in the shaded region is the median, and points beyond the whiskers represent outliers.

Table 2 contains the MANOVA results from the linguistic analysis. The MANOVA included all 13 linguistic features as dependent variables and emotion (with four categories) as the independent variable. The null hypothesis is rejected for each multivariate test for emotion.

Table 2
One-Way MANOVA Results with Emotion Predicted from Thirteen Subject-Baselined Linguistic Features

Multivariate Statistic	Predictor Variable	Estimate	*p*
Wilks' lambda	emotion	.423	<.0001
Pillai's trace	emotion	.710	<.0001
Lawley-Hotelling trace	emotion	1.067	<.0001

Table 3 contains the ANOVA results that accompany the MANOVA results for the linguistic analysis. For all 13 ANOVAs, a single linguistic feature is the dependent variable, emotion is the independent variable, and subject-within-emotions is the error term, and the null hypotheses are rejected with $p < .0001$. Although it was not planned as part of the study, subject-within-emotion as an independent variable with residual variance as the error term have their *F*-ratios reported in Table 3, with all these tests for the thirteen linguistic features also significantly different at the $p < .0001$ level.

Table 3

One-Way ANOVA Results with Emotion Predicted from Thirteen Subject-Baselined Linguistic Features

Predictor Variable	Source	*SS*	*df*	*MS*	*F*	*p*
total duration $R^2 = .8418$	emotion	14257.14	3	4752.38	55.69	<.0001
	subject-within-emotion	40617.27	476	85.33	75.96	<.0001
	residual	10216.95	9095	1.12		
total mean fundamental frequency $R^2 = .7107$	emotion	44708227.03	3	14902742.34	9.24	<.0001
	subject-within-emotion	767595561.42	476	1612595.72	44.07	<.0001
	residual	332809091.82	9095	36592.53		
pause duration between "was" and "interesting" $R^2 = .6072$	emotion	179.15	3	59.72	39.53	<.0001
	subject-within-emotion	719.18	476	1.51	23.97	<.0001
	residual	573.21	9095	0.06		
mean fundamental frequency of "I" $R^2 = .4705$	emotion	737599.59	3	245866.53	15.14	<.0001
	subject-within-emotion	7729198.10	476	16237.81	13.75	<.0001
	residual	9772041.90	8272	1181.34		
mean fundamental frequency of "went" $R^2 = .6562$	emotion	2178927.20	3	726309.08	34.04	<.0001
	subject-within-emotion	10157095.00	476	21338.43	29.20	<.0001
	residual	6640593.20	9088	730.70		
mean fundamental frequency of "for" $R^2 = .5852$	emotion	2497915.80	3	832638.60	28.70	<.0001
	subject-within-emotion	13810535.00	476	29013.73	22.26	<.0001
	residual	11745874.00	9013	1303.21		
mean fundamental frequency of "a" $R^2 = .6002$	emotion	2178428.90	3	726142.95	25.93	<.0001
	subject-within-emotion	13328824.00	476	28001.73	24.12	<.0001
	residual	10438956.00	8990	1161.17		
mean fundamental frequency of "walk" $R^2 = .5658$	emotion	2471480.40	3	823826.79	32.97	<.0001
	subject-within-emotion	11893760.00	476	24986.89	20.05	<.0001
	residual	11256986.00	9035	1245.93		

Table 3 - continued

Predictor Variable	Source	*SS*	*df*	*MS*	*F*	*p*
mean intensity of "I"	emotion	18672.26	3	6224.09	17.03	<.0001
	subject-within-emotion	173983.64	476	365.51	17.76	<.0001
R^2 = .5371	residual	170276.00	8272	20.58		
mean intensity of "went"	emotion	44271.85	3	14757.28	52.10	<.0001
	subject-within-emotion	134834.57	476	283.27	38.10	<.0001
R^2 = .7275	residual	67718.80	9088	7.45		
mean intensity of "for"	emotion	43824.35	3	14608.12	60.02	<.0001
	subject-within-emotion	115858.46	476	243.40	34.40	<.0001
R^2 = .7191	residual	63777.08	9013	7.08		
mean intensity of "a"	emotion	54263.37	3	18087.79	61.30	<.0001
	subject-within-emotion	140460.30	476	295.08	44.63	<.0001
R^2 = .7691	residual	59439.53	8990	6.61		
mean intensity of "walk"	emotion	21750.63	3	7250.21	26.09	<.0001
	subject-within-emotion	132263.77	476	277.87	33.52	<.0001
R^2 = .6750	residual	74885.67	9035	8.29		

The neural network analysis was evaluated using an 80-20 training and test split of the data. After 100 epochs of training, the model was evaluated using the test set, comprising 24 subjects' spectrograms. This model achieved an accuracy of 39% (945 spectrograms out of 2421 total spectrograms correctly identified). Note that the random chance of selecting the emotion correctly is 20% (1 out of 5 possible emotions), so this model offers a 19% improvement in accuracy from a chance model. Figure 8 details the accuracy for each training epoch, 1 through 100. The training accuracy did not approach 100%, leveling off closer to 60%. Possible reasons for this lack of accuracy during training will be explored below in the discussion section.

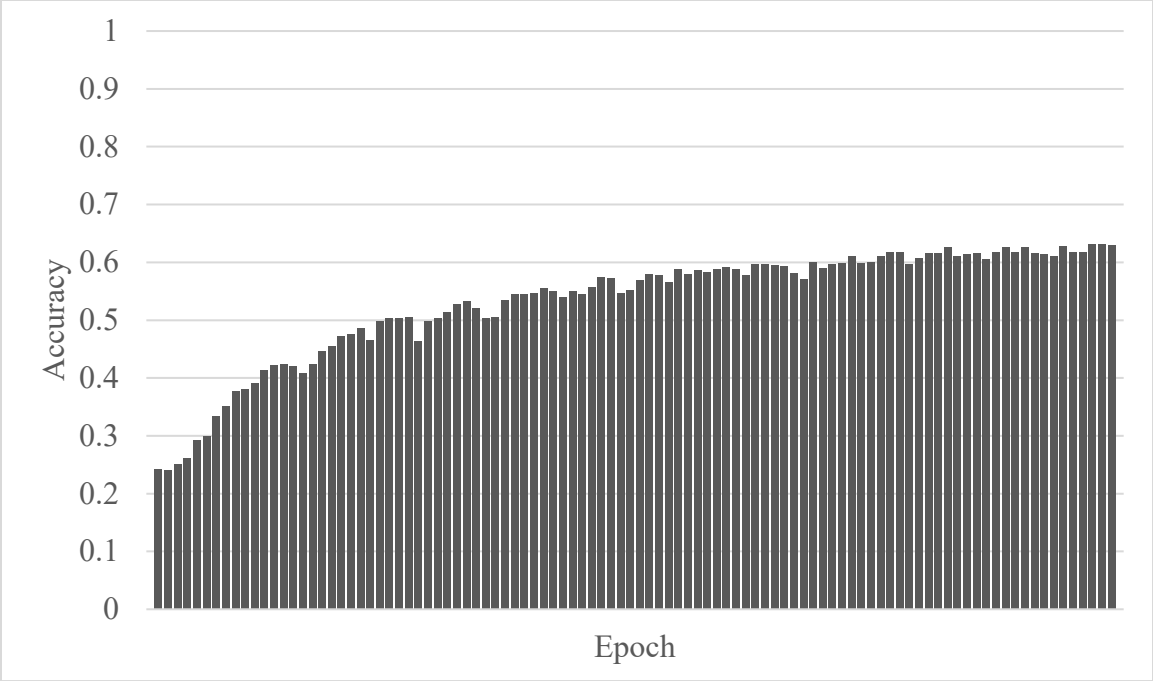

Figure 8. Neural network training set accuracy per epoch.

Table 4 contains descriptive statistics for the recordings in the variety instructions block of the study, along with the same statistics for the regular instructions data. An average of 20 recordings per emotion per subject were usable from each emotions' six-minute recording session in the regular instructions set. Because there were no recording sessions that reached the six-minute time limit for variety instructions, the number of unique expressions given by each subject can be considered exhaustive for each subject. Mean and median values representing the number of unique expressions recorded range from 7 to 9, which is higher than anticipated. The mode for each emotion is not 5, which was the hypothesized mode.

Table 4
Descriptive Statistics for Number of Recordings Per Subject Per Emotion in Each Instruction Set

instruction set	emotion	total	subject *n*	mean	standard deviation	mode	range	median
number of recordings for variety instructions	anger	240	30	8.000	4.763	3, 4, 6, 7	[2, 19]	7
	fear	289	30	9.633	6.178	6	[2, 21]	7.5
	happiness	244	29	8.414	4.851	10	[2, 23]	8
	sadness	226	31	7.290	4.656	4	[1, 23]	7
number of recordings for regular instructions	anger	2493	120	20.775	6.753	19, 21	[7, 51]	20.5
	fear	2307	120	19.225	6.852	17	[5, 49]	19
	happiness	2558	120	21.317	5.881	19, 25	[6, 36]	22
	neutral	2256	120	18.800	5.729	19	[3, 30]	19
	sadness	2217	120	18.475	5.734	17	[3, 37]	18

Note. The group subject size *n* for the variety instructions recordings represent a separate group of subjects for each emotion, but for regular instructions all 120 subjects are included for all five emotions.

Figure 9 contains the histogram with all subjects' number of different portrayals they managed to record before being unable to think of other distinct ways of expressing their assigned emotion. These distributions show some outliers on the higher end, with several subjects creating more than 20 unique expressions. Distributions are flatter than anticipated, with more variance possibly caused by differences in capabilities or confidence of subjects, among other possible explanations.

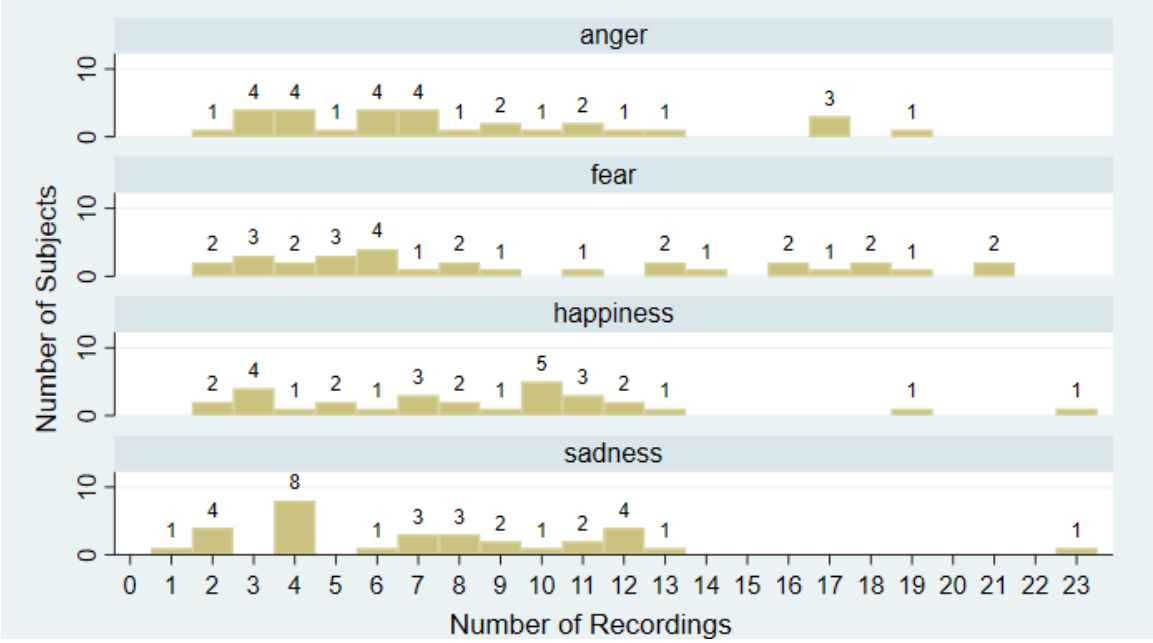

Figure 9. Frequency histogram for number of subjects per number of recordings completed during the variety instructions block, separated by which emotion the subject was instructed to act.

Discussion

The discussion section is organized into sections evaluating the goals, hypotheses, limitations, and possibilities for future research. Overall, the study was a success in creating a useful corpus and in making steps toward understanding emotion recognition.

Goals and Hypotheses

In this section, each goal and hypothesis is briefly considered before a general discussion about all of them is given.

Goals

The six goals of this study are evaluated individually below.

1. Replicate the task conditions and emotions found in Spackman, Brown, and Otto (2009) and Lauritzen (2009). All four emotions (anger, fear, happiness, and sadness) were replicated in this project. As for task conditions, the use of a professional microphone, a controlled recording environment, strict guidelines for the physical filter and seating arrangements, the counterbalanced ordering of emotions acted, and the ambiguous script read in all emotions were as close to identical as possible. However, the

intentional choice to not use the same exact script means that this is not a direct replication of their sample of actors. Also, they were concerned with sampling subjects in different majors to investigate acting experience affected recognition rates, but showing no sign of difference, that consideration was excluded from this study. Because of their results, it is still feasible to compare variation patterns in acoustic variables to compare patterns.

Figure 5 replicates an important finding: high arousal emotions (anger, fear, and happiness) overlap in duration measures, in contrast to low frequency emotions (sadness), which are distinguished by being much slower. Also, the smaller and simpler feature, pause duration between “was” and “interesting”, may offer a simple way to differentiate emotions. Instead of relying on duration of entire sentences or paragraphs, an online system could actively check pause duration between individual words as sentences are still being spoken, which could lead to faster identification of emotions being expressed. The large sample size means it may be possible to differentiate emotions despite the patterns overlapping. Using multiple features that vary in domain maximize patterns that vary between pairs of emotions, which in turn allows for discriminability between easily confusable emotions when all features are used for modeling emotion simultaneously. An alternative approach would be to cluster all recordings within each emotion and then identify subtypes per emotion, allowing for more finely tuned recognition when emotions have large intra-emotion variance.

Fear ended up being more difficult to categorize compared to the other emotions. For example, it is most similar to anger in total duration, to sadness in first sentence mean fundamental frequency, and partially to anger, sadness, and happiness in total mean fundamental frequency. The nuance to these differences make it more difficult to categorize fear as following the typical pattern of high arousal emotions, though duration being shorter does support that. It’s possible that sadness is responsible for the unexpected results in this study, since previous research supports it being lower frequency in many measures compared to all other emotions. These other studies may not have accounted for a subtype of sadness being distraught or inconsolable sadness,

meaning sadness where high energy of expression is still present. Sobbing and other speak styles that would disrupt speaking clearly were prevented in the present study, however, so follow-up studies will be necessary to determine if fear or sadness is abnormal compared to previous studies' results.

Other studies have similar structure from which we can draw some comparisons. Jiang and Cai (2004) studied prosodic features in Chinese using a sizable collection of recordings to identify different emotions. Having controlled conditions allowed for easier extraction of features, as was the case in the present study. Prosodic features and tones in that study relate most closely to frequency features used in the present study, such as fundamental frequency. It is interesting to consider that tones in English as well as Chinese can differentiate emotions despite Chinese being a tonal language, meaning that differentiating words relies on tonal changes. Having tone related to emotion could mean that there are clear biological conditions that affect tone even when a person is controlling tone carefully for the pronunciation of words, as is the case when mimicking emotion or acting.

2. Collect additional audio data according to new task conditions that are meant to further explore possible variations of emotional expressions. The variety block of instructions were successfully applied, with a large dataset of 120 subjects collected. These recordings exceeded researchers' expectations, although no subjects continued beyond the scheduled recording time. Figure 9 demonstrates an interesting pattern: most subjects were able to act differently several times, but a few subjects went beyond what was expected and expressed a double-digit quantity of unique expressions. Because each subject was the primary judge of what was considered unique, it would take another study with listener judges to determine if their recordings truly are different enough to be considered unique by people in general.

These data offer a new path in vocal emotion research, which will be discussed in the future research section below.

3. Collect a large audio corpus of emotion in a controlled recording environment. Statements about the corpus being large should be deeply qualified with

the assumptions implied by the type of corpus created: the corpus has more than one distinct emotion represented, it contains audio recordings widely available to researchers (e.g., digitally distributed on the Internet), and the control over the recording environment limits background noise and the recording equipment is consistent in position and settings. In absolute scale of size, 50 hours of audio recordings from 120 subjects presented in this corpus are considered small for a corpus, but when compared to similar corpora for vocal emotion research, it is large.

All studies that fit the limited scope and design of this study have already been introduced and discussed earlier, so connections to other research ideas are now considered. Other studies can be compared in size for audio data produced but represent a breadth of purposes beyond emotion recognition. One example comes from a study created a useful collection of audio recordings for the purpose of speech synthesis (Saratxaga et al., 2006). This study had a total of 760 Basque sentences recorded per subject. This would be useful for creating a voice that mimics that particular subject's voice. However, given the scope of that study, only two subjects' data were collected. This tangentially-related goal of speech synthesis yields better results from a wide variety of sounds in the target language.

The vocal emotion corpus presented here contains a variety of sounds, but not nearly all sounds possible in English and far from all possible pairs of sounds for producing all common combinations. Therefore, this study can only be somewhat useful when compared to other studies' application for speech synthesis.

4. Host the audio corpus on the Internet for any researcher to use for their studies and projects. Previous studies' data about vocal emotion are still useful, but they are often overlooked because they are not readily available on the internet for free. This problem is remedied in the current study: the vocal emotion corpus is available on the Internet without needing any account credentials or fees necessary for downloading it. The corpus is formally called Vocal Emotion Corpus Data and is hosted on ScholarsArchive (https://scholarsarchive.byu.edu/data/12). It is divided up by subsets of the audio recordings previously discussed, with *Regular Instructions* containing the

largest portion of data, which comprises the data that was included in the linguistic and neural network analyses presented in this study. Containing over 50 hours of audio files in total, it offers a useful tool for anyone wanting a project exploring artificial intelligence, voice synthesis, or voice recognition. It has some limits on usefulness, as further explained in the limitations section below, but offers a useful dataset for many types of research. For exploring the breadth of possible emotional expressions, the *Variety Instructions* recordings can also be useful. Recordings that are more challenging to analyze but may be useful for speech error detection or noise detection in emotion recognition tasks are found in *Additional Corpus Files* and *Additional Subjects* directories.

5. Create a tutorial for how to use the audio corpus for analyses about emotion expression. All scripts used for collecting, processing, and analyzing the recordings are included on the previously mentioned ScholarsArchive page called Vocal Emotion Corpus Data in the *Documents* directory. A data description file explains many of the researcher decisions that led to choosing each scripting language and in resolving unexpected difficulties during programming. It serves as a programming journal and walkthrough presenting each script in succession, so that someone could replicate the results of this dissertation precisely by following along. In addition to these resources, recording protocols, the exact script for reading, and the license governing the corpus are included. All resources can be used for noncommercial projects under the provided Creative Commons license (CC BY-NC-SA 4.0, https://creativecommons.org/licenses/by-nc-sa/4.0/). Software packages for analysis, however, are governed by their respective copyrights, so although the scripts are available, researchers interested in using them need to acquire access to these software packages themselves.

6. Compare two analytic approaches from two different fields: MANOVA modeling found in existing emotion studies in psychology and unsupervised deep neural networks taken from computer science. The MANOVA approach worked well enough as a general way of differentiating emotions. Table 3 shows that even when

reducing the degrees of freedom due to the repeated measures aspect of having multiple recordings per subject, the variance in each dependent variable were sufficient to explain a large proportion of variance due to the emotion categories, ranging from 47-84%. The decision to use difference scores as a way of making each subject's neutral voice their own baseline may have contributed to the strength of the model. The MANOVA model is limited in that means for the four emotions are all compared to each other simultaneously in each test, so that large differences between contrasting emotions lead to significant differences despite similar emotions having similar means and variances. The accompanying ANOVAs suffer from the same limitation. Using enough linguistic features to differentiate every pairing of emotions in at least one feature, however, compensates for this limitation, ensuring there is always a way to differentiate emotions in all possible cases of confusability. The large sample size from using features extracted from thousands of recordings was a factor in the large *F*-ratios reported in Table 3.

Each set of emotions is overall different, with a high amount of overlap due to variability in each acoustic feature. One of the best discriminators of emotions was the total duration feature, which, on average, tended to have low values for sadness and high values for the other three emotions. The features that represent smaller subsets of the available audio, such as the word "went", were smaller in effect size than comparable the global features, and tended to have more variation and more overlap between emotions, as can be seen in Figure 6.

The neural network model did not perform as well as expected. Humans can identify four or five emotions on average at an above 90% accuracy level, so an accuracy of 39% from this model is far below that. A simple, unsupervised model that was selected for its simplicity, leaving out many possible researcher decisions that more complex models require. Because the intention of this analysis was to contrast the top-down linguistic approach, fewer assumptions and decisions were desired when designing it. These choices for design decisions likely led to the unsatisfying results. A combined approach, using knowledge of typical similarities between some pairings of emotions to enhance an existing data-driven neural network could possibly improve accuracy.

Figure 8 shows the training set's accuracy per epoch. The strength of neural networks come from their ability to create abstractions beyond what a human could readily see from data, but they suffer from issues with overfitting and underfitting. Overfitting means that a model created from a training set can be too sensitive to idiosyncrasies of the training set that prevent it from generalizing to other data. Underfitting means that the model does not appropriately fit parameters that would yield high accuracy, often measured as low accuracy in evaluating the training set as well as the test set. Figure 8 shows this model potentially has issues with fit, never fully predicting the training set. Some post-hoc analyses using alternate network designs and training designs were used for comparison. The initial design and subsequent post-hoc designs were found to be inaccurate due to a programming bug, so they were discarded in favor of the first model that followed the planned design philosophy without the bug. This design is the one reported in Figure 4 and the one for which all results of the analysis are reported for this study. Post-hoc analyses to attempt to improve or understand this model are reported below.

There are a few variations that were tried after the model finished training and testing to attempt to understand why the accuracy was so low. The original model uses dropout and convolutions in an attempt to avoid overfitting the model. An alternate model with 1000 epochs had no better test set accuracy than the 100 epoch model, still plateauing in training around 60% accuracy. Another alternative was to use an even simpler network design, one that simply reduced the size of each subsequent layer, with no dropout or convolutions. This model did nearly reach 100% accuracy in training, but ended up with much lower accuracy in testing, at 24%, just a 4% improvement over chance assignment.

An exploration of individual classifications shows that extreme examples might have influenced the model greatly. Fear was overrepresented for classifications, and neutral was underrepresented. With less than 10% of recordings reaching longer than 25 seconds, these longer recordings may have influenced training. Most of the longer recordings were exclusive to sadness and fear recordings. Neutral recordings overlap

happiness in many features and therefore may have been misclassified as happiness.

Overlap in loudness in the recordings, as seen in Figure 7, could be another reason for the failure to build a predictive model. The spectrograms created as input for the neural networks model allow for the model to be sensitive to their three dimensions: duration, frequency, and power. Loudness affects the power domain, and if two recordings are similarly loud and similarly long, there is little to work with from the nuances in frequency for these recordings. Although a higher resolution spectrogram could capture these nuances in frequency, this was not possible in the current study due to these spectrograms being prohibitively large, so large that their resolution would have needed to be downsampled just to load them into VRAM while training the model, defeating the higher resolution's purpose. It is possible that a more efficient modeling method than the one employed in PyTorch could have worked, but it was beyond the scope of this study.

It is likely that the complexity of the pattern of features needed to differentiate a variety of emotions leads to needing higher resolution data, which in turn could lead to overfitting a model, which would defeat the ability to classify new recordings correctly. This difficulty should be further explored by computer scientists.

Overall, the linguistic approach, with its design informed by decades of similar studies, better discriminates between emotions than the neural network design. It does not, however, offer a predictive model for new data. The neural network model has this capability but only predicts accurately 39% of the time. Perhaps a combined approach, either informing a neural network design with theories relevant to recognizing each emotion or applying a clustering method and predictive model to enhance the linguistic MANOVA model could reach human levels of accuracy.

Hypotheses

Hypothesis 1: Emotions differ on acoustic measures. I reject the null hypothesis for the MANOVA model, which is that all four emotions (anger, fear, happiness, and sadness) will have equal means for each emotion, and that the overall pattern across all variables is the same for each emotion.

Hypothesis 2: Using some audio data to build a model, you can accurately identify the emotion in newly presented audio data. The prediction for the neural networks model is that the prediction accuracy of the test data will be greater than the rate of accuracy at a chance level (20% when using all five emotions). The results were better than this baseline, at 39% accuracy, but below the 90% accuracy for five emotions that a human listener would be estimated to achieve. Further manipulations of design of the input data resolution, network layers, dropout proportions, number of epochs, and other parameters did not yield higher accuracy. The lowest accuracy was found to be for neutral and fear emotions, but removing them one at a time did not yield better models, and it was deemed too far removed from the original intention to build a model on fewer than four emotions, so further revisions were halted.

It may be that a revised model that combines the benefits of features extracted from a theory-driven analysis with the complex modeling of neural networks may perform better. However, the few examples of previously published attempts to use that approach did not have encouraging results. One study only reached 50% accuracy in recognizing eight emotions, even after adjusting models to improve accuracy (Nicholson et al., 2000). As more techniques for modeling neural networks are developed, more possible solutions for modeling this complex task of vocal emotion recognition may emerge.

Hypothesis 3: There will be a clear mode for unique emotional expressions. The number of unique expressions in Figure 9 failed to show a very clear mode for each emotion, and the density of the frequency plot did not center along any particular value, resembling a flat distribution more than any other common distribution. From the statistics presented in Table 4, the chosen mode of five was not supported, I failed to provide strong evidence for this hypothesis.

While not all hypotheses were supported, the study accomplished its goals, offering a large corpus of recordings for researchers to use, with many new and exciting features to explore as well as plenty that replicate existing studies, allowing for comparisons to be made. The scripts and analyses offer insights into possibilities for

future generations of researchers to quickly design class projects using this corpus, or for established researchers to learn techniques that are rarely applied to emotion research.

Limitations

Anyone who uses the corpus presented here should be aware how limited it is in representing voices speaking American English. Other languages and most accents for English are not represented. A review by Scherer and his colleagues highlighted that there are cultural differences in production and perception of vocal emotion, but there is also evidence for universal understanding (Scherer et al., 2011). These researchers acknowledge that there are not enough acceptable datasets collected for understanding vocal emotion fully. They also advocate for a more theory-driven research paradigm that tests specific mechanisms of vocal emotion production, vocal emotion perception, or both. Focusing on listeners' perceptual experience of the portrayals, such as having them rate "breathiness" and "genuineness", could give clues to researchers for which acoustic cues were salient to listeners as they made judgments. The present study is designed so that future studies can use the corpus as stimuli in recognition and perception studies to further our knowledge of emotion recognition. Although the sample is not very diverse, collecting data must start somewhere, and having a more homogenous sample can aid in avoiding effects due to accents or culture instead of effects due to emotional expression.

The number of recordings per subject in the regular instructions portion of the experiment indicates several things: whether the subject made many mistakes or not, whether the subject spoke slowly or quickly, whether there were technical delays or not, and whether the subject was late to the appointment or not. These effects are not controlled for effectively, so a limitation of the study is that while the equality of time spent for each block of recordings was preserved, the number of recordings could still vary per block. This imbalance will somewhat affect any results of analyses that use audio without averaging them per subject.

One limitation for acted emotion remains, which is knowing what the cues for an emotion are and recognizing them does not always translate to being able to act convincingly using those cues. Additional analyses could identify outliers among

speakers, specifically speakers who do not understand emotions or who cannot act convincingly. If these outliers are identified, they can be studied further for future detection and removal from studies seeking to understand typical emotion recognition. A listener study design could also be used to rate speakers in their acting ability.

A limitation of using a silent recording room to control for noise is that subjects are more hesitant to speak loudly. Although subjects were instructed that they could be as loud or quiet as they felt would be genuine, the loudness as captured by the intensity measures in Figure 7 indicate that overall they were quiet. Many recordings were in the whisper threshold of 30 dB, and a majority of recordings were below the threshold of a normal conversation of two people in a room, which is 60 dB. Although there were several subjects who chose to act anger as particularly loud, breaching 80 dB, overall the genuineness of the recordings might be questioned because of the unusual quietness of the subjects. A follow-up study with listeners would clarify if this quietness is a problem for recognition and genuineness or not.

Future Research

Psychology has traditionally used a top-down approach to research, with theories logically formulated and then tested through empirical findings, while computer science often uses a bottom-up approach, focusing on collecting large empirical datasets and inferring theories from the patterns in the data. Each approach could solve the emotion recognition problem. However, all previous attempts in each approach by researchers have failed to produce discriminability of multiple emotions in a variety of environments at the skill level of an average adult. Finding statistically significant differences between expressions of vocal emotion is easy to do, but creating a predictive model of vocal emotion has proven elusive. There are two research paths that could extend the methods in this study: a top-down approach that seeks to refine emotion theory and a bottom-up approach that seeks more data to build refined models.

The design of this study had no definitive way to verify what constitutes a distinct pattern among recordings per subject per emotion. Subjective ratings by listeners in a follow-up study could help distinguish subjects who attempted the same pattern for every

recording of an emotion and subjects who chose to vary their performance over the course of portraying an emotion. Anecdotally, most subjects seemed to use an approach of expressing the emotion the same way every time during the regular instructions portion of the study. I perceived only a few subjects who varied their loudness, speech rate, and pitch from portrayal to portrayal during the regular instructions portion of the study. A study could verify how people approached the acting task as well as how unique their intentionally varied expressions are.

Although the word-level linguistic features did not offer much of an advantage over the recording-level features, they did add to the overall complexity of the model, allowing for nuanced differences between similar emotions to be explored. Although not a word-level or recording-level feature, the forced pause feature was a first of its kind for this type of study, and did prove useful for differentiating emotions. Future research could continue to explore small features such as this pause to start compiling a list of useful features that people may use in aggregate to make determinations of what emotion is being expressed.

Research progress in specific languages may be able to generalize to all languages once enough studies make their data widely available. With enough research in enough languages, the commonalities between languages can start to be established. So far, studies are often limited to language-specific concerns, in an attempt to first understand one language at a time. An alternate approach would be to take the data from this or other available audio corpora and to compare their acoustic features to look for commonalities. With a sizable sample size in multiple studies with controlled conditions, a researcher may be able to identify general patterns.

The neural network approach might need even more data. Using 50 hours of data may not have been enough for a robust model, especially considering that the dataset included several subjects that could be considered outliers due to their exceptionally long recordings. In addition to this, using only one script with 42 words makes it difficult to generalize to other words or syntax structures. Future studies could combine this corpus with other corpora that include these five emotions to create a more generalizable model.

In addition to varying approaches, future research can probe the depth of features within the script. This script features a question as the last sentence. The end of the sentence marked with a question mark is likely not as interesting as initially formulated since many subjects stated the question in a declarative way at the end instead of rising intonation, which is how a typical question is marked in English speech. However, the beginning of the sentence, anchored by “Why”, offers a tentative, hesitant, or questioning intonation that was expressed similarly by many subjects. A clearly declarative statement, “I went for a walk”, is offered as contrast to the beginning of the question, “Why don’t you come with me”. A study could investigate how the question intonation varies from emotion to emotion. Question-based intonation changes for emotion recognition could yield another feature for discriminating emotion. Additionally, the content of the script may lend to different features compared to other studies’ scripts. The ambiguity of statements like “You wouldn’t believe what happened” allow for many possible dramatic interpretations. The variety instructions dataset could also hold many insights into nuanced expressions of emotion for these ambiguous statements.

References

Allgood, R., & Heaton, P. (2015). Developmental change and cross-domain links in vocal and musical recognition performance in childhood. *British Journal of Developmental Psychology, 33*, 398-403.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596.dsm01

Banse, R., & Scherer, K. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality & Social Psychology, 70*(3), 614-636.

Bhullar, N. (2007). *Effect of facial and vocal emotion on word recognition in 11-to-13-month-old infants.* (Doctorate), Virginia Tech, 3313183.

Boersma, P., & Weenink, D. (2019). *Praat (Version 6.0.48)*. Amsterdam, Netherlands: University of Amsterdam. Retrieved on 10 July 2020 from http://www.praat.org/

Brigham Young University. (2020). *Facts & Figures.* https://www.byu.edu/numbers

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17*(2), 124-129.

Ekman, P. (2016). What scientists who study emotion agree about. *Perspectives on Psychological Science, 11*(1), 31-34.

Feldstein, S. (1964). *A profile analysis of simulated affective vocal behavior*. Paper presented at the British Psychological Society, Leicester, England.

Jiang, D. N. & Cai, L. H. (2004). Classifying emotion in Chinese speech by decomposing prosodic features. *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing*, Jeju Island, Korea.

Jürgens, R., Drolet, M., Pirow, R., Scheiner, E., & Fischer, J. (2013). Encoding conditions affect recognition of vocally expressed emotions across cultures. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00111

Jürgens, R., Grass, A., Drolet, M., & Fischer, J. (2015). Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected. *Journal of Nonverbal Behavior, 39*, 195-214. https://doi.org/10.1007/s10919-015-0209-5

Juslin, P. N., Laukka, P., & Bänziger, T. (2018). The mirror to our soul? Comparisons of

spontaneous and posed vocal expression of emotion. *Journal of Nonverbal Behavior, 42*, 1-40. https://doi.org/10.1007/s10919-017-0268-x

Kowallis, L. (2018). *Are vocal expressions of emotion unique to each person?* Unpublished manuscript.

Laukka, P., & Juslin, P. N. (2007). Similar patterns of age-related differences in emotion recognition from speech and music. *Motivation and Emotion, 31*, 182-191. https://doi.org/10.1007/s11031-007-9063-z

Lauritzen, M. (2009). *Acoustic mediation of vocalized emotion identification: Do decoders identify emotions idiographically or nomothetically?* (Doctor of Philosophy), Brigham Young University, Brigham Young University.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). *Montreal Forced Aligner* (Version 0.9.0) [Computer software]. Retrieved from http://montrealcorpustools.github.io/Montreal-Forced-Aligner/

Morningstar, M., Verity, Y., Feldman, L., & Dirks, M. (2018). Mid-adolescents' and adults' recognition of vocal cues of emotion and social intent: Differences by expression and speaker age. *Journal of Nonverbal Behavior, 42*, 237-251. https://doi.org/10.1007/s10919-018-0274-7

Nicholson, J., Takahashi, K., & Nakatsu, R. (2000). Emotion recognition in speech using neural networks. *Neural Computing & Applications, 9*, 290–296. https://doi.org/10.1007/s005210070006

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N. Antiga, L., Desmaison, A. Kopf, A., Yang, E. Devito, Z., Raison, M. Tehani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, *32*, 8024–8035. Curran Associates, Inc. Retrieved on 10 July 2020 from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Saratxaga, I., Navas, E., Hernáez, I., & Aholab, I. (2006). Designing and recording an emotional speech database for corpus based synthesis in Basque. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC '06)*, Genoa, Italy.

Sauter, D., Panattoni, C., & Happé, F. (2013). Children's recognition of emotions from vocal cues. *British Journal of Developmental Psychology, 31*, 97-113. https://doi.org/10.1111/j.2044-835X.2012.02081.x

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99*(2), 143-165.

Scherer, K. R. (2013). Vocal markers of emotion: Comparing induction and acting elicitation. *Computer Speech and Language, 27*, 40-58. https://doi.org/10.1016/j.csl.2011.11.003

Scherer, K. R., Clark-Polner, E., & Mortillaro, M. (2011). In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion. *International Journal of Psychology, 46*(6), 401-435. https://doi.org/10.1080/00207594.2011.626049

Spackman, M., Brown, B., & Otto, S. (2009). Do emotions have distinct vocal profiles? A study of idiographic patterns of expression. *Cognition and Emotion, 23*(9), 1565-1588.

Vasconcelos, M., Dias, M., Soares, A. P., & Pinheiro, A. P. (2017). What is the melody of that voice? Probing unbiased recognition accuracy with the Montreal affective voices. *Journal of Nonverbal Behavior, 41*, 239-267. https://doi.org/10.1007/s10919-017-0253-4

Zhou, H., Huang, M., Tianyang, Z., Xiaoyao, Z., & Bing, L. (2018). *Emotional chatting machine: Emotional conversation generation with internal and external memory.* Paper presented at the Thirty-Second Association for the Advancement of Artificial Intelligence.