Categorical and Dimensional Ratings of Emotional Speech: Behavioral Findings From the Morgan Emotional Speech Set

Article *in* Journal of Speech Language and Hearing Research · October 2019

DOI: 10.1044/2019_JSLHR-S-19-0144

CITATIONS
2

READS
133

1 author:

Shae Morgan
University of Louisville
22 PUBLICATIONS **13** CITATIONS

SEE PROFILE

JSLHR

Research Article

Categorical and Dimensional Ratings of Emotional Speech: Behavioral Findings From the Morgan Emotional Speech Set

Shae D. Morgan[a,b]

Purpose: Emotion classification for auditory stimuli typically employs 1 of 2 approaches (discrete categories or emotional dimensions). This work presents a new emotional speech set, compares these 2 classification methods for emotional speech stimuli, and emphasizes the need to consider the entire communication model (i.e., the talker, message, and listener) when studying auditory emotion portrayal and perception.
Method: Emotional speech from male and female talkers was evaluated using both categorical and dimensional rating methods. Ten young adult listeners (ages 19–28 years) evaluated stimuli recorded in 4 emotional speaking styles (Angry, Calm, Happy, and Sad). Talker and listener factors were examined for potential influences on emotional ratings using categorical and dimensional rating methods. Listeners rated stimuli by selecting an emotion category, rating the activation and pleasantness, and indicating goodness of category fit.
Results: Discrete ratings were generally consistent with dimensional ratings for speech, with accuracy for emotion recognition well above chance. As stimuli approached dimensional extremes of activation and pleasantness, listeners were more confident in their category selection, indicative of a hybrid approach to emotion classification. Female talkers were rated as more activated than male talkers, and female listeners gave higher ratings of activation compared to male listeners, confirming gender differences in emotion perception.
Conclusion: A hybrid model for auditory emotion classification is supported by the data. Talker and listener factors, such as gender, were found to impact the ratings of emotional speech and must be considered alongside stimulus factors in the design of future studies of emotion.

An important aspect of social communication is a listener's ability to interpret the emotional state of an interlocutor. This communicative information enriches the interaction by telling the listener how the talker feels or how the talker wishes to be perceived as feeling. Such an interaction requires effective verbal encoding of the emotional state to be portrayed and its superimposition onto the verbal message. Both the message and the emotion must then be successfully transmitted to the listener, who is tasked with interpreting not only the message but also the emotional state of the talker. To study the complexities of this interaction, one must possess an understanding of speech production, auditory perception, and the psychological constructs of emotion; however, there seems to be a lack of consensus on appropriate methods for the investigation of perceptual auditory emotional experience in research, despite the recent increase in interest on the topic (as mentioned in Picou et al., 2018). Thus, we attempt to illuminate several methodological considerations (talker, listener, and stimulus factors) that will add to the body of literature examining the perception of auditory emotion. This is achieved through the introduction of a new set of emotional speech materials and the examination of behavioral ratings used in its validation.

[a]Program in Audiology, Department of Otolaryngology Head and Neck Surgery and Communicative Disorders, School of Medicine, University of Louisville, KY
[b]Department of Communication Sciences and Disorders, University of Utah, Salt Lake City
Correspondence to Shae D. Morgan: shae.morgan@louisville.edu

Editor-in-Chief: Bharath Chandrasekaran
Editor: Kate Bunton
Received April 1, 2019
Revision received May 28, 2019
Accepted August 2, 2019
https://doi.org/10.1044/2019_JSLHR-S-19-0144

What Is Emotion?

To discuss methodological considerations of auditory emotion perception, it is necessary to first examine what is meant by "emotion" as presented in this article. Many excellent discussions and definitions of emotion and its associated experiences have already been written (e.g., Ekman, 1992; Goldsmith, 1994; Gray & Watson, 2001; Russell, 2003; Scherer, 2005). Defining emotion can be

Disclosure: *The author has declared that no competing interests existed at the time of publication.*

difficult (Scherer, 2005); however, for our purposes, emotion will be defined (based on the above references) as a state of being or the seemingly instantaneous presence of feeling in response to some eliciting stimulus. Scherer (1987, 2005) suggested that the emotion-eliciting event or stimulus can be external (e.g., interacting with others or the environment) or internal (e.g., thinking about a past experience). For auditory emotion, the eliciting stimulus that triggers this emotional response is a sound. The range of feelings that occur in response to auditory emotion elicitors is vast and different, as are the opinions on how these feelings should be classified and studied.

Perceptual Model

Scherer (2003) recommended using a modified version of Brunswik's functional lens model (Brunswik, 1956) for understanding how emotions are communicated and then classified by listeners. Scherer's model considers all facets of a vocal emotional interaction; the model includes the encoding, transmission, and appraisal of emotions. An individual wishing to express or display an emotional state (real or feigned) selects a target emotional state and begins the transmission process. Transmission involves encoding the intended target feeling into distal indicators (suggested to be the acoustic characteristics of the voice), which are then received by the listener through sound transmission. Based on the distal indicators received, the listener makes perceptual judgments about the emotional state of the talker. There are two primary theories of how emotions are classified and, subsequently, how individuals judge a talker's emotional state: discrete and dimensional theories of emotion.

Theories of Emotion Classification

Discrete Categories

This theory of emotion classification was founded on the idea that humans have evolved to react to environmental stimuli in specific ways, resulting in a finite list of emotional states that are elicited by certain situations. This basic emotions theory (Darwin & Prodger, 1998) posits that emotions, as well as human experiences that lead to emotion, are discrete categories such as anger, happiness, fear, and sadness (Ekman, 1992; Ekman & Cordaro, 2011). Ekman (1992) proposed that emotional expressions (e.g., facial and vocal) fall into these basic categories, and the variability of emotion (and subsequent labels) comes from differing intensities of the emotion (e.g., frustration and fury are simply different intensities of the "anger" basic category). The list of specific emotion categories proposed by this theory could swell to a great number as emotional category labels become increasingly complex (e.g., tenderness, contentment, irritation). Additionally, limiting emotional expression to a finite set of emotion labels makes it difficult to account for the large host of specific feelings experienced by individuals. There is also a subjective component to emotion categorization, and what is perceived to be "joy" by one listener may be considered "bliss" by another and so on. Some emotions, such as contempt, are difficult for the layperson to conceptualize and consistently agree upon with others (Matsumoto & Ekman, 2004), resulting in categories that are not well defined. Thus, assignment of syntactic labels to subjective emotional experiences is a predominant limitation of using discrete models to classify auditory emotion.

Dimensions

Another approach to identifying and categorizing different emotions is conceptualized as the circumplex model of affect (Russell, 1980). This model suggests that emotional experiences result from the individual feeling some level of arousal or activation and some degree of valence or pleasantness. For example, an individual who wins the lottery (categorically described as pleasant surprise or joy) would likely be highly activated and feeling quite pleasant. In contrast, an individual who has been unemployed for several months after being unexpectedly terminated from employment (categorically described as sad or depressed) could feel very inactive and unpleasant during that experience. Thus, with the circumplex model of affect, all emotions are circumscribed by the dimensional combination of the activation and pleasantness induced by the eliciting stimulus. These dimensional qualities are reflected in the vocalizations of the talker (Laukka, Juslin, & Bresin, 2005), which are then interpreted by the listener to form a judgment of the talker's emotional state.

Some investigations suggest that the dimensions of activation and pleasantness alone are insufficient to describe all aspects of human emotion (e.g., Laukka et al., 2005; Schlosberg, 1954); however, adding other dimensions to explain more emotional experiences begs the question of whether their inclusion is only to explain additional discrete category labels not covered by fewer dimensions or whether they define an inherent property of the emotional experience. If dimensions are added to accommodate categories, one must wonder if discrete and dimensional models are simply different ways of viewing the same emotional phenomena or if they do in fact contribute uniquely to the understanding of emotional expression and perception. In response to this conundrum, recent studies of emotion have proposed hybrid models of emotion classification that involve both dimensional and categorical approaches.

Hybrid Models

Returning to our definition, we have specified that emotion is a feeling or experience that comes as a result of some eliciting stimulus (for our purposes, an auditory stimulus). The emotional reaction results in measurable phenomena (e.g., facial movements or vocal changes) that are perceptually dimensional, and the assessment of these dimensional changes results in the assignment of a categorical label (Mehu & Scherer, 2015). This hybrid approach has been evaluated recently in visual stimuli by measuring the categorical perception of stimuli along with their dimensional attributes using facial emotion portrayals that are morphed continua between two emotion category extremes

(Fujimura, Matsuda, Katahira, Okada, & Okanoya, 2012). Fujimura et al. concluded that the two methods of classification can co-occur during perceptual tasks, supporting the hybrid model. Other research has since extended this finding to classification of emotion in music (Eerola & Vuoskoski, 2011), and (to our knowledge) no studies have intentionally investigated the hybrid theory with speech materials and auditory perception. It should be noted that Livingstone and Russo (2018) measured category selection and some ratings of dimensionality (intensity and genuineness) during the validation of their speech and song materials database; however, the relationship between dimensional and discrete ratings was not the focus of their work. Thus, evidence supporting a hybrid model of emotion classification for speech stimuli has not been reported previously.

Considerations

In search of evidence to support a hybrid model of auditory emotion classification, one must consider the many variables that influence classification of vocal emotion. We examine some of these variables by evaluating talker, stimulus, and listener factors that influence the production and perception of emotional speech stimuli.

Talker Experience

Some researchers argue that speech stimuli ought to be controlled as much as possible, requiring acted productions of prototypical expressions of the emotions under consideration. This method is the most popular for collecting vocal productions of emotion (e.g., Goudbeek & Scherer, 2010; Laukka, Audibert, & Aubergé, 2012; Laukka & Elfenbein, 2012; Ruffman, Henry, Livingstone, & Phillips, 2008; Williams & Stevens, 1972). This method produces expressions of emotion that are prototypical but still subject to the interpretation of the actor or actress. The inherent issue with using acted speech is that we rarely hear prototypical expressions of emotion in our everyday conversation.

In another point of view, the expressions of emotions may be more ecologically valid if produced spontaneously and in natural settings that induce those emotions organically in talkers (e.g., Arimoto, Ohno, & Iida, 2011; Laukka, Neiberg, Forsell, Karlsson, & Elenius, 2011; Schmidt, Janse, & Scharenborg, 2016). Databases formed from these types of samples generally come from TV show recordings or other large publicly available resources (e.g., radio broadcasts or online gaming interactions). Spontaneous speech, although more ecologically valid than read lists of sentences, adds significant difficulty when creating a new corpus. Extracting the few emotional expressions that may be interspersed through hours of recordings is time consuming, and natural recordings of spontaneous speech may be contaminated with extraneous noise in the environment. Another limitation to these recordings is that the reliability of the emotion label assignment is compromised, as the labels are assigned after the recorded event is completed (Cowie et al., 2011; Scherer, Clark-Polner, & Mortillaro, 2011).

Stimulus Elicitation

To elicit vocal production of emotion, talkers are typically asked to produce a set of stimuli (affective bursts, nonsense words, words, sentences, passages, etc.) in their own subjective conceptualization of an emotion. The inclusion of emotion categories to be produced is typically determined by the experimenter, with no standard for which emotions should be included, although Ekman (1992) does propose a list of "basic" emotions that are regularly considered. Another method for eliciting emotional productions is induction (e.g., Bachorowski, 1999). Emotion induction provides the talker with a scenario, a script of a scene, or an experience that is intended to induce a specific emotion. A third method of emotion elicitation is similar to induction but is more personal, asking the talkers to recall a time in their life when they experienced the target emotion and then to immerse themselves in that recollection as they produce the stimuli.

Listener Factors

Some listener factors, such as gender, may influence emotional ratings or accuracy of emotion identification. For example, female listeners may demonstrate superior emotion recognition accuracy compared to male listeners (Lambrecht, Kreifelts, & Wildgruber, 2014). This advantage is attributed to evolutionary biology and females' historical need to tend to the emotional productions of their children (see also "the primary caretaker hypothesis"; Babchuk, Hames, & Thompson, 1985).

Few studies have examined gender differences in the portrayal and perception of emotion in speech. A meta-analysis by Hall (1978) provided accuracy for males and females decoding nonverbal, emotional communication. The results from the analysis of 75 different studies revealed a female advantage in emotion identification. More recently, Lambrecht et al. (2014) confirmed this advantage in the auditory modality. Another study showed that females gave more negative ratings of facial stimuli than males did (Natale, Gur, & Gur, 1983), suggesting gender differences in the processing of pleasantness or valence. Using affective bursts, Belin, Fillion-Bilodeau, and Gosselin (2008) found that female talkers received higher ratings of activation and intensity and lower ratings of valence (pleasantness) compared to male talkers. They also found that male listeners gave higher ratings of emotional intensity than female listeners, and there was no interaction between talker and listener gender.

To effectively assess these influences on vocal emotion recognition and classification, emotional speech samples appropriate to the aim of the study or process must be used. For example, to study gender effects, a database must include male and female talkers and listeners; however, some sets of emotional speech are produced by a single talker or talkers of one gender (e.g., Dupuis & Pichora-Fuller, 2014). Thus, the database to be used should be carefully selected based on the aims of the study, or a new database should be constructed suitable to achieve those aims.

Existing Emotional Speech Corpora for Auditory Testing

Several corpora of emotional speech for use in auditory testing are currently available with validation data and other information (e.g., acoustic analyses) about the recordings. For a review of these databases, we refer you to excellent reviews provided by Koolagudi and Rao (2012); El Ayadi, Kamel, and Karray (2011); and, more recently, Picou et al. (2018).

Rationale for the Creation of a New Database

While these reviewed emotional speech corpora may be suitable for testing listener emotion recognition abilities, they are designed in a way that limits their usefulness in studying other aspects of auditory processing (such as competing speech scenarios; e.g., Brungart, 2001; Helfer & Freyman, 2009). The available databases provide no prime or cue in their materials that listeners might use to select a target emotion or target sentence from a mixture of other stimuli. Priming cues are commonly employed when evaluating a listener's ability to localize a target talker and suppress irrelevant auditory information (e.g., the coordinate response measure corpus; Bolia, Nelson, Ericson, & Simpson, 2000). These auditory skills are essential in listening scenarios with competing speech sources (e.g., the cocktail party effect; Cherry, 1953) and are particularly affected in individuals with hearing loss (e.g., Helfer & Staub, 2014). An emotional speech set that contains a priming cue would allow for investigation of how emotionality affects auditory processing in these relevant and complex listening scenarios. Additionally, many of the reviewed databases include only one talker per target group (e.g., male and female, older adult and younger adult), limiting the ability to explore talker differences that likely exist with emotion production. To address the lack of priming cues and talker variability in the existing emotional speech sets, we created a new database that includes several male and female talkers and a priming cue within the sentences, which will be used in future experiments testing auditory processes that may be influenced by emotion in the voice.

Purpose and Hypotheses

The purposes of this work are to (a) present the validation and characteristics of a new set of emotional speech that may be used in future research on auditory perception, (b) contrast listeners' use of categorical and dimensional ratings of emotion using this database, and (c) illustrate talker and listener factors that impact these ratings of emotional speech.

We hypothesize that dimensional ratings will group stimuli into emotion categories that fit a category label. We further hypothesize that listener confidence in their category selection will increase as dimensional ratings approach extremes of the rating scale. Such a finding would suggest that dimensionality of speech stimuli contributes to the categorical assignment of emotional speech, providing evidence for a hybrid model of emotion classification with speech stimuli. Finally, we hypothesize that talkers and listeners will vary in their production and perception of emotional speech, demonstrating the need to carefully design studies with appropriate constraints on talker, stimulus, and listener factors most relevant to the research question.

Development of the Morgan Emotional Speech Set

A corpus of emotional speech was created and then validated using both methods of emotion classification while considering a complete model of vocal emotion (Scherer, 2003).

Method: Stimuli Creation and Recording

Talkers

All recruiting was carried out using procedures approved by the University of Utah Institutional Review Board (IRB). Six participants (three men, three women) for recording were recruited from the Department of Theater, University of Utah. Participants were all Caucasian young adult students (aged 19–21 years, $M = 20$ years, $SD = 0.8$ years) who were native speakers of American English and had completed all coursework in vocal production offered by their program. This population was targeted for recruitment to ensure that they had received some formal training on vocal emotion production and would be able to accurately and consistently produce the emotions desired for the study. Preference was given in this instance to the experimental control of the emotional productions rather than to the naturalness of the stimuli (as could have been obtained with untrained talkers or spontaneous speech productions).

Stimuli

The 1,080 Theo–Victor–Michael (TVM) sentences used by Helfer and Freyman (2009) were used in the present work. These sentences were originally designed to investigate effects of speech-on-speech auditory processing, which requires a cue name to which the listener may be primed to identify a target talker amidst other distracting talkers using similar speech materials. For example, a listener would be instructed to listen to the sentence that starts with the cue name "Theo," while three sentences are presented, each with a unique cue name. These sentences are structured as follows: "[Cue name] discussed the [word 1] and the [word 2] today," where the cue name was replaced with the name "Theo," "Victor," or "Michael," and the word placeholders were filled with one- or two-syllable common nouns. The specific structure of these sentences lends to experiments related to auditory perception, and these sentence features may be utilized in future research endeavors studying the effects of emotion on other aspects of auditory perception (e.g., competing speech paradigms with emotional targets and distractors). At present, no other emotional speech database contains a priming cue to accompany target keywords,

making this database, its structure, and its potential applications unique.

The talkers produced TVM sentences in four emotional styles. The emotion categories Angry, Happy, Sad, and Calm each represent a quadrant of the activation/pleasantness dimensional space, similar to the arousal/valence plane (Russell, 1980). Specifically, Angry is a high-activation, low-pleasantness emotion; Happy is a high-activation, high-pleasantness emotion; Sad is a low-activation, low-pleasantness emotion; and Calm is a low-activation, high-pleasantness emotion. Previous studies typically employ Neutral as a fourth category instead of Calm; however, neutral speech has been demonstrated to contain a negative valence (Scherer, Banse, Wallbott, & Goldbeck, 1991), and so Calm was included in place of Neutral as an emotion that contrasts from the others in activation and pleasantness. Other recent work has also proposed the use of Calm as a reference emotion category similar to Neutral with which other emotions may be compared (Livingstone & Russo, 2018).

Selecting one emotion from each quadrant allows for examination of data for each individual emotion category, as well as for groups of emotion categories along a specific emotion dimension. For example, comparing emotion recognition scores for Calm and Sad categories compared to Angry and Happy categories allows for a comparison between low- and high-activation emotions.

The 1,080 TVM sentences include 360 sentences for each talker of a given gender. The 360 sentences were divided into 90 sentences for each emotion, resulting in 30 sentences for each cue name, emotion, and talker (30 sentences × 3 cue names × 4 emotions × 3 talkers = 1,080 sentences). The initial recordings were designed to contain more stimuli than were necessary to allow for five stimuli for each cue name–emotion–talker combination to be removed in the event of errors or inaccuracies in production or recording. In the event that no (or fewer than five) inaccuracies were discovered, the tokens least representative of their emotion category for each combination of talker, emotion, and cue name were excluded. These tokens were identified as those with the largest Euclidean distance from the grand mean of activation and pleasantness (determined using the validation study detailed below) for each emotion category. After removal of these stimuli, the final database contained 1,800 stimuli (25 sentences × 3 cue names × 4 emotions × 3 talkers × 2 talker genders).

All sentences were semantically neutral relative to the emotions examined in this study (i.e., they carried no happy, sad, angry, or calm emotional meaning). Several scoring words used in the original TVM sentences were determined by the experimenter to prime a particular emotion based on their semantic meaning (e.g., "joy") and were replaced with more emotionally neutral common words (e.g., "lint") that have no obviously apparent emotional meaning. Male and female talkers recorded the same sets of sentences due to the limited number of one- to two-syllable nouns that are common enough to be used as identifiable words for the average listener. However, within each gender group, each talker recorded unique sentences. That is, the first, second, and third male talkers all recorded unique sentences, but those sentences were also recorded by the first, second, and third female talkers, respectively.

Procedure

Talkers participated in two sessions each. During the first session, they consented to the study and agreed to the practice and recording procedures. They were given the lists of sentences that would be recorded in each emotion, which they were instructed to review and practice prior to their second session. At this time, the talkers were also instructed on the dimensional differences (activation and pleasantness) between the desired emotional productions and were encouraged to practice their production of the target emotions with these dimensional qualities in mind during their rehearsal. Talkers returned on a different day for a second session after having rehearsed the sentences at least one time through. During the second session, the experimenter recorded the talkers as they produced the rehearsed sentences. The recordings were carried out in a sound-treated room. The room was lined almost completely with sound-absorbing foam squares of varying size and density to help reduce reverberation within the room.

Participants were seated at a table in the room with a list of sentences in front of them. The sentences for each emotion were in a different randomized order than they had practiced to avoid any order familiarity, ensuring that practicing the sentences in a specific order would not affect their productions. Participants wore a Shure SM10 head-mounted microphone, which was connected directly to a Marantz digital audio recorder. Each emotion set was recorded as a separate sound file. At the beginning of each set, before the rehearsed sentences were spoken, participants read a practice list of 12 TVM-style sentences (4 sentences × 3 cue names; all scoring words were replaced with one- or two-syllable colors) to engage their rehearsed production of the specified emotion for that task set. The order in which emotions were recorded was randomized for each talker. Participants were offered breaks in between lists to help them disengage from the emotional style of the completed recording and to prepare to record the next emotional style. The experimenter monitored production of the sentences during the recording using Sennheiser HD550 headphones connected to the Marantz audio recorder. During the practice sentences, the recording level was adjusted on the Marantz recorder to maximize the recorded input without peak clipping to ensure a high-fidelity recording of each emotion. All recordings had a 36-kHz sampling rate, which maintained signal fidelity of spectral information through 18 kHz (sufficient for audible speech information). If at any time a sentence peak-clipped, was misread, or was otherwise unfavorably produced or recorded (as determined by the experimenter; e.g., observed extraneous noise), the sentence was noted and copied to a page to be read by the talker at the end of the recording. Determination of unfavorable productions was based on aspects of the production or recording environment that were unrelated to the emotional expression of the talker (e.g., noise

introduced by the talker shifting in the chair, unnatural pauses, starting a word over midsentence). In three instances, an entire emotion section was reproduced either at the request of the participant (for feeling they had not given a true portrayal of the intended emotion) or due to some error in recording that was inaudible during monitoring but evident post hoc upon examination of the audio files. For the repetitions requested by the talkers, the recordings used as part of the database were those that the talker felt were their truest representation of the intended emotion (in both cases, the repeated section was selected to replace the original).

Stimulus Preparation and Availability

Recorded files were segmented into individual sentence files using ELAN (Version 5.0.0-alpha [ELAN, 2017]; to mark boundaries) and MATLAB (Release 2017a [MATLAB, 2017]; to cut the stimuli). Rough boundaries for individual stimuli were set manually by listening to each recording and marking the breaks between sentences in ELAN. These boundaries were exported into a text file that was utilized by MATLAB to rapidly segment the long recording file and rename the cut stimulus according to its stimulus code. Once rough cut, the stimuli were trimmed using a MATLAB program that presented the rough stimuli, allowed the user to mark the true stimulus onset and offset, identified the nearest zero-crossing, and then saved the finalized stimuli. The experimenter was able to toggle between the waveform and spectrogram of the stimuli, as well as preview how the stimulus sounded prior to finalizing the cut. The stimuli were all scaled to the same root-mean-squared amplitude to ameliorate level differences among talkers and recording sessions.

The final stimuli with their associated ratings are available for use as a public tool for researchers interested in studying emotional speech. Interested parties may download the set for free on Zenodo (doi:10.5281/zenodo.2662514), an online repository. Validation data including raw category selections and individual dimensional ratings for each listener are available from the author upon request.

Validation Method: Perceptual Ratings

Listeners

The perceptual ratings task was carried out at the University of Utah following procedures approved by the IRB. The task involved listening to the stimuli and then making several subjective judgments about them. Eleven listeners were recruited; data from one male subject were excluded due to his withdrawal prior to completing the validation task. Thus, 10 participants (five men, five women) aged 19–28 years (M = 24 years, SD = 2.9 years) completed the validation. Participants were native speakers of American (n = 9) or Canadian (n = 1) English.

Stimuli

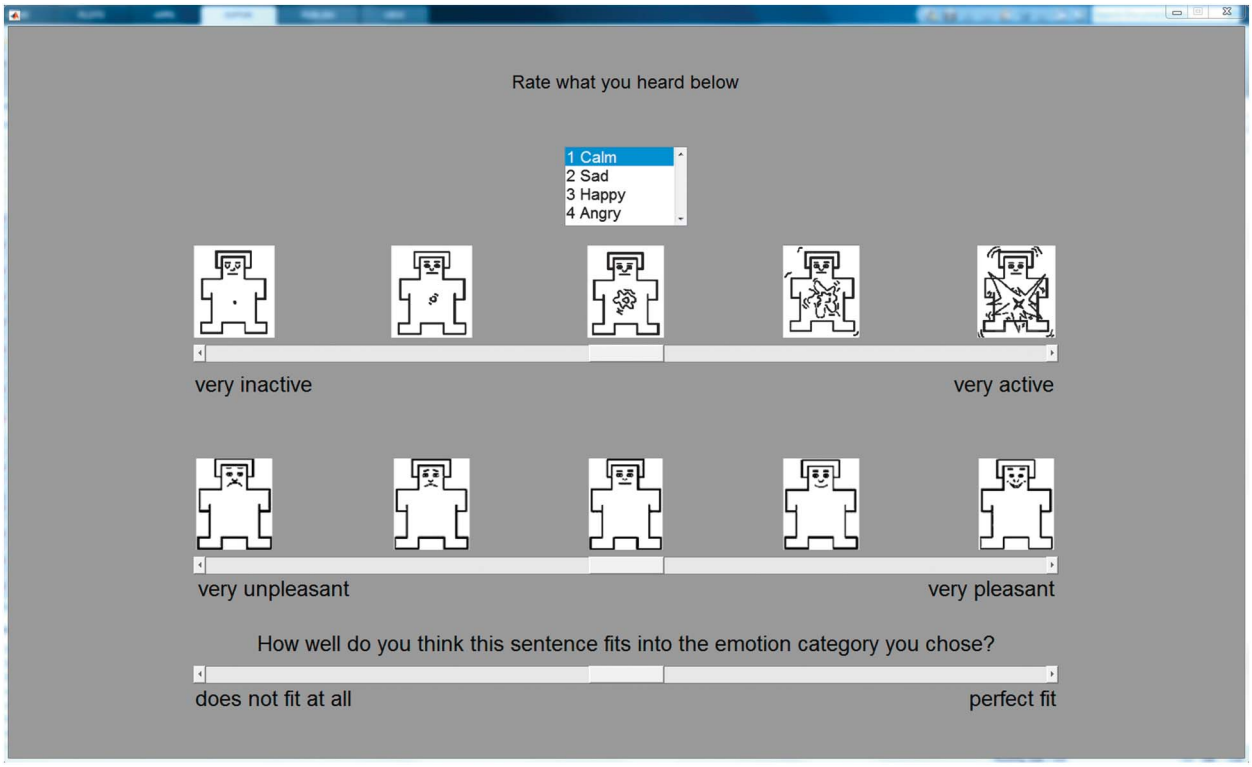
All stimuli recorded for the development of the Morgan Emotional Speech Set (MESS) were evaluated. The inclusion of all stimuli (i.e., discarded stimuli in addition to the final database) was to capture additional variability in dimensional ratings of stimuli, rather than basing results exclusively on the most representative stimuli for each emotion. Thus, listeners heard all 2,160 sentences (30 sentences × 3 cue names × 4 emotions × 3 talkers × 2 genders). In addition to the speech stimuli, several nonspeech sounds that have been previously rated using emotion dimensions were included in the validation task as a check of dimensional emotion ratings. These stimuli came from the International Affective Digital Sounds (IADS) corpus, and a full description of their creation and validation can be found in Bradley and Lang (2007). Twenty-four stimuli were selected from the IADS (six for each quadrant of the activation/pleasantness space). Listener ratings of both MESS materials and IADS in the present validation provided a way to confirm whether listener perceptions of activation and pleasantness for speech stimuli were congruent with ratings made by other groups of listeners in previous work.

Procedure

The task consisted of six sessions lasting approximately 90 min each. During the first session, participants signed consent forms and received a hearing screening at 20 dB HL for octave frequencies from 250 to 8000 Hz (American National Standards Institute, 2010). All participants passed the hearing screening. Next, participants were given instructions and performed the experimental task. The task consisted of making four judgments of each of the 364 stimuli (360 MESS sentences + 4 IADS stimuli) in each session (364 stimuli × 6 sessions = 2,184 total stimuli). Figure 1 provides an example of the graphical user interface (GUI) that was seen by the listeners and that was used to collect the data. Following the rating task, listeners filled out a Positive and Negative Affect Scale (Watson, Clark, & Tellegen, 1988) to assess their level of affect over the past week (consistent with traditional administration of the scale). Ratings of positive and negative items were summed to yield a composite positivity score and a composite negativity score. The average positive and negative ratings (along with the standard deviations) given by listeners in the present work were consistent with the normative data for the scale, suggesting that the present stimuli were rated by a sample of listeners that were similar in affective state to the general population. The mean positivity and negativity ratings from Watson et al. (1988) along with the means and standard deviations from this study are presented in Table 1. The subsequent five sessions were identical to the first, with the exception that consent procedures and hearing screening were not performed.

As mentioned, the participants completed four judgments of each stimulus. The first was to select which emotion the listener heard in the sentence from a list of options (Angry, Happy, Sad, and Calm). Collecting judgments of emotion category for many stimuli allowed the creation of a confusion matrix to evaluate listener accuracy in emotion category judgments. This method of data collection has been employed by other studies (Dupuis & Pichora-Fuller, 2011; Luo, Fu, & Galvin, 2007) and provides an evaluation of

Figure 1. The graphical user interface used in the validation of the Morgan Emotional Speech Set.



whether or not the talkers in the present work produced categorically identifiable emotional portrayals.

For the second judgment, the listener assessed the amount of activation in the talker's voice from a scale of 0 = *very inactive* to 100 = *very active*. For the third rating, the listener assessed the amount of pleasantness in the stimulus on a scale from 0 = *very unpleasant* to 100 = *very pleasant*. Although other descriptors have been used for these emotion dimensions, the terms *pleasantness* and *activation* were expected to most likely maximize the listeners' understanding of each label's intended meaning. Additionally, the Self-Assessment Manikin (SAM; Bradley & Lang, 1994) was included in the GUI to aid listeners' rating of activation and pleasantness. The SAM is a visual tool that depicts a rudimentary manikin figure expressing activation or pleasantness (see Figure 1). To strengthen the internal validity of the present work, pictures of the SAM were included in the GUI in equal intervals above the slider bar (with the middle "neutral" representation of the SAM centered on the halfway point of the slider) for rating activation and pleasantness. Listeners were encouraged not to center their ratings on a particular picture of the SAM but to use the entire scale with the pictures as a guide. Since ratings of SAM have been used to obtain accurate judgments of these two emotion dimensions (e.g., Bradley & Lang, 1994), adding them to the GUI was intended to help listeners make selections of activation and pleasantness that are comparable with previous work using this measure.

Finally, the fourth rating was a judgment of how well the listener thought the presented stimulus fit into the category selected in the first task. This "goodness-of-category-fit" measure was obtained to allow listeners to express that a stimulus may have been placed into a category only because no other category options were provided. Thus, listeners were able to express whether a stimulus was or was not characteristic of the category they chose. This measure also allowed an evaluation of correlations between dimensional attributes and category selection confidence, which would bolster the argument in favor of a hybrid model of emotion classification.

Listeners were seated in a sound-treated booth (as previously described for the database recordings) in front of a monitor, a keyboard, and a mouse. The keyboard had

Table 1. Means and standard deviations (*SD*s) for the Positive and Negative Affect Scale questionnaire obtained during the original validation study and the present work.

Item rating	Watson et al. (1988)	Morgan (2018)
Positive		
M	33.3	30.9
SD	7.2	6.0
Negative		
M	17.4	16.3
SD	6.2	5.2

a cover so that the only exposed key was the "Enter" key. The participants were instructed to use the mouse to select the emotion they heard in each sentence, to drag the top slider to make a judgment of activation, to drag the middle slider to make a judgment of pleasantness, and to drag the bottom slider to select how well the stimulus fit into the category that was selected. Each stimulus was only presented one time. The program offered participants a break after each quarter of the session.

Stimuli were played through the custom MATLAB GUI using Tucker-Davis Technologies (TDT) System III signal processing hardware. The program randomly selected (from a list of options for each session) the stimulus for presentation on a given trial and passed it to a TDT RP2.1 enhanced real-time processor (used to communicate with MATLAB and interface with the other TDT equipment), a TDT PA-5 programmable attenuator (used to adjust signal level), and, finally, a TDT HB-7 headphone driver. Sennheiser HD550 headphones were connected to the headphone driver for binaural, diotic presentation of the stimuli. The system was calibrated regularly throughout the validation so that the stimuli would be presented to listeners at a level of 60 dBA (consistent with provisions in the IRB protocol and similar to other emotion studies, e.g., Luo et al., 2007). The calibration was carried out using a Larson-Davis Technologies sound-level meter coupled to a preamplifier and a circumaural faceplate with appropriate weight to simulate the use of the headphones. It should be noted that, although calibration was performed monaurally to a level of 60 dBA, the listeners heard the stimuli binaurally, which adds a perceptual benefit of approximately 3 dB under headphones, and therefore, the actual effective listening level was approximately 3 dB higher. Presentations of the stimuli were monitored by the experimenter during the experiment.

Results

Data Analysis

For each stimulus, the category selection was noted and ratings of activation, pleasantness, and goodness of category fit were averaged across all listeners to create a composite score for each rating. Analyses were carried out in Stata 14 (StataCorp, 2015). Linear mixed-effects models (LMMs) and Pearson product–moment correlations (PPMCs) were used to evaluate the listener judgments. LMMs have been shown to be particularly beneficial when performing tests involving speech materials by accounting for variance associated with different talkers (Quené & van den Bergh, 2004). Talker was included in each LMM as a random intercept to model variability associated with the six talkers in the study. Random slopes were included for talkers by emotion category. PPMCs were used to demonstrate the relationships between ratings of talker confidence in their category selection and ratings of activation or pleasantness. This measure allows a unique evaluation of a listener's perceptions of and relationships between categorical selections and dimensional ratings. Linear combinations of the model estimations for activation and pleasantness groups (low vs. high) were also assessed post hoc to confirm whether emotion groups were significantly distinct based on their dimensional ratings. Specifically, model estimates were aggregated by emotion dimension using the lincom function in Stata 14 to compare high- and low-activation emotion groups (angry and happy vs. calm and sad, respectively) and pleasant and unpleasant emotion groups (calm and happy vs. angry and sad, respectively). These analyses assessed whether the talkers effectively produced the intended emotions and whether the intended emotions were dimensionally distinct.

Reliability

Interrater reliability for category selection was assessed using Fleiss' kappa. Kappa values for each emotion implied moderate interrater agreement for Calm and Sad category selection (.44 and .59, respectively) and substantial agreement for Angry and Happy category selection (.80 and .75, respectively), based on criteria by Landis and Koch (1977). All reliability calculations for category selection were statistically significant ($p < .001$), suggesting that rater agreement was not due to chance. Intraclass correlation coefficients (ICCs) assessed the rater agreement for ratings of activation and pleasantness. Since all listeners rated all stimuli, two-way, random-effect, single-item ICCs (2, 1) were calculated, and listeners demonstrated good agreement for ratings of activation, ICC(2, 1) = .69, $F(2183, 19647) = 25.69$, $p < .001$, and pleasantness, ICC(2, 1) = .69, $F(2183, 19647) = 24.87$, $p < .001$, following interpretation guidelines by Cicchetti (1994). Furthermore, two-way, random-effect ICCs were calculated and averaged for multiple stimuli, ICC(2, k), and listeners demonstrated excellent agreement for ratings of activation, ICC(2, k) = .96, $F(2183, 19647) = 25.69$, $p < .001$, and pleasantness, ICC(2, k) = .96, $F(2183, 19647) = 24.87$, $p < .001$. Additionally, all ratings of the IADS stimuli were within 1 *SD* of those obtained during their original validation, suggesting that the listeners in this study gave ratings similar to those given during other validations of emotional stimuli.

Category Selection

Raw accuracy scores, accompanied by unbiased hit rates (Wagner, 1993), and statistics of rater agreement on category selection are typically employed to assess the validity of the category selections (e.g., Livingstone & Russo, 2018). Listeners were quite accurate in selecting the intended emotion category for each stimulus, with performance for the identification of all emotion categories well above chance. A confusion matrix detailing the raw accuracy (in percentage of listener judgments) and unbiased hit rates for stimuli from each emotion category is shown in Table 2. Errors of category assignment were consistent with those observed in other studies that used similar methods to evaluate their emotional speech corpora (e.g., Dupuis & Pichora-Fuller, 2011). The most common confusion was between

Table 2. Confusion matrix showing the proportion of listener judgments for each target emotion that were given to each of the four emotion categories.

Target emotion	Categorical judgments			
	Calm	Sad	Happy	Angry
Calm	.88 (.54)	.07	.04	.01
Sad	.28	.70 (.63)	.01	.01
Happy	.14	.00	.85 (.79)	.01
Angry	.12	.01	.01	.86 (.83)

Note. Unbiased hit rates (Wagner, 1993) are in parentheses.

calm and sad stimuli, as also manifested by the lower inter-rater agreement on those items; however, it may be noted that similarly reduced recognition accuracy for these categories has been observed in other studies (e.g., Livingstone & Russo, 2018).

Perceived Activation

Activation was assessed using a slider bar ranging from 0 to 100 to approximate a continuous scale. The scale ranged from *not very activated* to *very activated*. In the LMM used to assess activation, perceived activation (averaged across listeners) was the dependent variable, with emotion category (Angry, Calm, Happy, Sad) as the predictor variable and a random intercept for talkers with random slopes for talkers by emotion category. The results from this study show a significant difference in activation among emotion categories ($z > -2.25$, $p < .05$). Post hoc analyses (adjusted for multiple comparisons using Bonferroni's method) showed significant differences in arousal for all comparisons except for Happy and Angry ($z = 1.27$, $p = .613$) and Calm and Sad ($z = -2.33$, $p = .06$). Linear combination analysis of the ratings for high-activation (Angry and Happy) and low-activation (Sad and Calm) emotion groups confirmed these findings by revealing higher activation ratings for the high-activation emotion group compared to the low-activation emotion group ($\beta = 30.72$, $z = 178.28$, $p < .001$). Table 3 contains ranges, means, and standard errors for activation ratings of the four emotion categories.

Perceived Pleasantness

Pleasantness was also assessed using a slider bar ranging from 0 to 100. The scale ranged from *very unpleasant* to *very pleasant*. The same model described above was used to assess perceived pleasantness as the dependent variable. The results show a significant difference in ratings of pleasantness among emotion categories ($z = -4.95$, $p < .001$). Post hoc analyses (with appropriate adjustment using Bonferroni's method) revealed significant differences in pleasantness among all emotion categories (all $|zs| > 3.89$, all $ps < .001$). Linear combination analysis of the ratings for high-pleasantness (Calm and Happy) and low-pleasantness (Angry and Sad) emotion groups revealed higher pleasantness ratings for the high-pleasantness emotion group compared to the low-pleasantness emotion group ($\beta = 31.45$, $z = 62.00$, $p < .001$). Table 3 also contains ranges, means, and standard errors for pleasantness ratings of the four emotion categories.

Table 3. Ranges, means, and standard errors (*SE*s) of activation and pleasantness ratings obtained for Morgan Emotional Speech Set stimuli.

Emotion category	Range	Mean	*SE*
Activation			
Calm	19–68	44.08	0.31
Sad	16–67	33.17	0.34
Happy	52–79	68.34	0.13
Angry	51–83	70.35	0.16
Pleasantness			
Calm	34–75	55.55	0.23
Sad	19–54	34.63	0.17
Happy	50–86	72.76	0.14
Angry	13–51	30.79	0.19

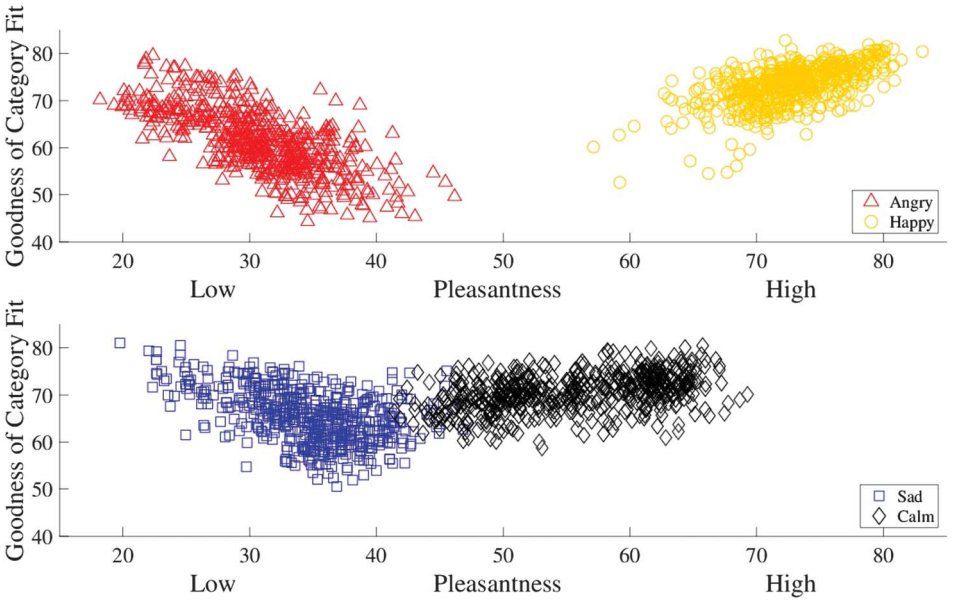
Note. Ratings were made using a scale ranging from 0 to 100.

Perceived Goodness of Category Fit

Listeners provided a measure of satisfaction with their category selection by indicating how well each sentence fit into the category they selected. The scale ranged from *does not fit at all* to *perfect fit*, thus allowing listeners to express whether the categories available to choose from were or were not accurate matches for the stimuli using a more continuous response variable. These ratings were also performed on a visual analog scale ranging from 0 to 100 with two–decimal-point precision. PPMCs were calculated to assess relationships between goodness of fit and both activation and pleasantness for each emotion category. The critical value for determining statistical significance was adjusted using Bonferroni correction for multiple comparisons.

As expected, more unpleasant emotions (Angry and Sad) showed a significant negative correlation between pleasantness and goodness-of-fit data, and more pleasant emotions (Happy and Calm) showed a significant positive correlation between pleasantness and goodness-of-fit data (see Figure 2). Also quite expectedly, high-activation emotions (Happy and Angry) showed a significant positive correlation between activation and goodness of fit, while low-activation emotions (Calm and Sad) showed a significant negative correlation between activation and goodness of fit (see Figure 3). These correlations, while all statistically significant, ranged from slight to strong, with the weakest correlation between activation and goodness of fit for the Sad stimuli ($r = -.17$, $p < .001$) and the strongest between pleasantness and goodness of fit for the Angry stimuli ($r = -.65$, $p < .001$). Table 4 further details the correlations between goodness of category fit and activation or pleasantness for all emotion categories.

Figure 2. Ratings of pleasantness and goodness of fit for high-activation (top panel) and low-activation (bottom panel) emotions.



Talker and Listener Gender

As a preliminary investigation into whether similar differences exist in this database, an analysis was performed to identify any talker and listener gender differences for ratings of activation and pleasantness. In this model, ratings of activation and pleasantness were the dependent variables (of separate models), with talker gender, listener gender, and the interaction between talker and listener gender as fixed effects. Individual talkers were included as a random intercept. Inclusion of random slopes yielded no significant contribution to the model fit, so they were excluded. There were significant main effects of talker gender ($\beta = -8.60$, $z = -7.93$, $p < .001$) and listener gender ($\beta = -4.05$, $z = -5.45$, $p < .001$) on ratings of activation, but no significant main effects were found for talker and listener gender for ratings of pleasantness (both $|zs| < 1.34$, all $ps > .18$). Table 5 provides the means and

Figure 3. Ratings of activation and goodness of fit for low-pleasantness (top panel) and high-pleasantness (bottom panel) emotions.

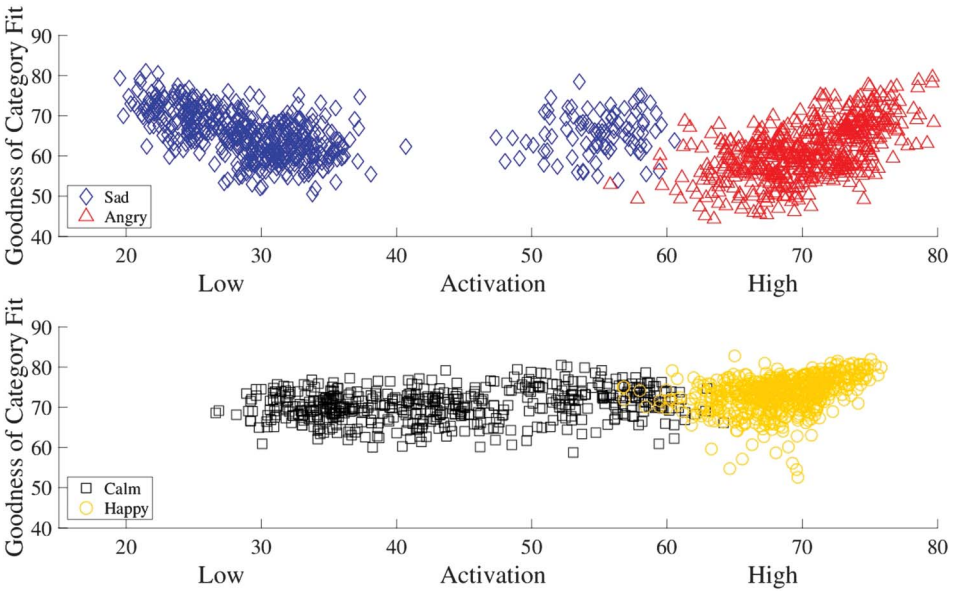


Table 4. Pearson product–moment correlations and associated *p* values for listeners' ratings of goodness of category fit with activation and pleasantness for the four emotion categories.

Emotion category	*r*	*p*
Activation		
Calm	.20	< .001
Sad	–.17	< .001
Happy	.37	< .001
Angry	.49	< .001
Pleasantness		
Calm	.40	< .001
Sad	–.46	< .001
Happy	.56	< .001
Angry	–.65	< .001

standard errors for ratings made by male and female listeners for stimuli produced by male and female talkers. No significant Talker Gender × Listener Gender interaction was present for either ratings of activation or ratings of pleasantness (both *p*s > .475).

Discussion

Emotion Classification Method

Our first hypothesis was that dimensional ratings would group stimuli into emotion categories that fit a consistent category label. In general, discrete, categorical ratings of emotional stimuli were consistent with their dimensionally perceived boundaries. Listeners demonstrated the ability to accurately classify these emotional speech productions into their intended category with accuracy well above chance performance. The results indicate that the four emotions included in the study were categorically and dimensionally distinct, and they formed groupings based on unique combinations of activation and pleasantness. We, therefore, propose the MESS as a valid emotional speech set that contains perceptually appropriate emotional speech stimuli for use in future experiments.

Interestingly, while there were no differences found between activation ratings for emotion categories within the high- or low-activation groups (i.e., activation ratings were not different between Angry and Happy or between Calm and Sad), pleasantness ratings were distinct for emotion groups (high vs. low) and each emotion within each group (Calm vs. Happy and Angry vs. Sad). Listeners seemed to rate the activation of speech stimuli using a two-alternative approach, either low or high activation. In contrast, they rated the pleasantness using more of a continuum from low to high. Based on these findings, it is possible that listeners may dichotomize the activation dimension first and then make inferences as to the emotional state of the talker based on the nuances detected in the pleasantness domain. That is to say, perhaps listeners follow a structure of processing auditory emotional information in speech to arrive at the final appraisal of the talker's emotional state. In this structure, listeners may process the most salient cue information first (whether the speech indicates high or low talker activation), followed by the less salient, but more defining, information of pleasantness (whether high or low within that activation category). Further investigation is required to confirm this proposal and better understand factors related to the portrayal and perception of emotion that may result in the differential use of the two emotional scales.

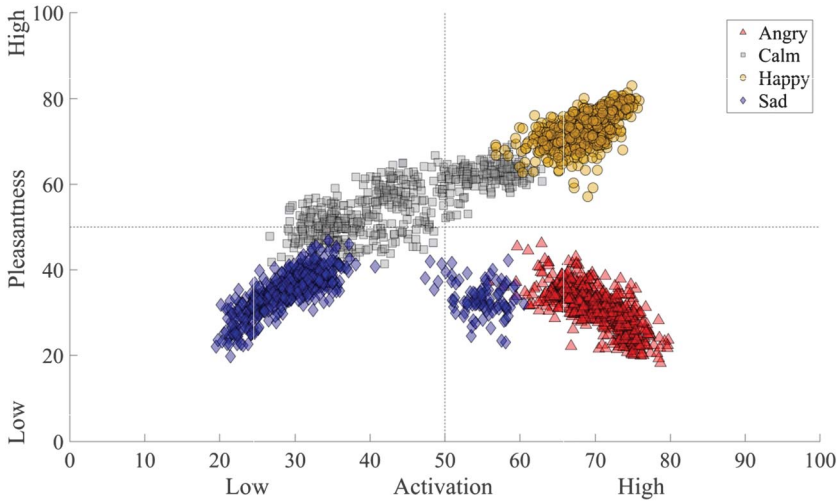
Table 5. Means and standard errors (*SE*s) of activation and pleasantness ratings obtained for Morgan Emotional Speech Set stimuli.

Dimensional ratings	Male talkers		Female talkers	
	M	*SE*	*M*	*SE*
Activation				
Female listeners	51.58	0.57	60.18	0.48
Male listeners	48.05	0.59	56.13	0.46
Pleasantness				
Female listeners	47.06	0.53	49.13	0.59
Male listeners	48.12	0.51	49.42	0.55

The second hypothesis was that listener confidence in their category selection would increase as dimensional ratings approached extremes of the rating scale. The correlations between category appropriateness and dimensional ratings observed in this study support the assertion that, as the sentences changed in perceived activation or pleasantness toward the predicted quadrant in the activation–pleasantness dimensional space, the listeners were more satisfied with their category selection. Indeed, the stimuli with the lowest reported goodness-of-fit values in each category were generally those with incorrect categorical assignment. These findings agree with the hybrid model of emotion classification found for facial (Mehu & Scherer, 2015) and music (Eerola & Vuoskoski, 2011) stimuli. These studies demonstrated that categorical and dimensional perceptions of emotion likely co-occur and that they are related. Specifically, Mehu and Scherer found that, as ratings of emotional dimensions for facial stimuli changed, there was a corresponding shift in categorical perception of the emotion. The findings of the present work extend this to auditory emotion perception. As dimensional ratings of stimuli became more extreme and better defined, the listeners in this study were more confident with their category assignments.

This evidence suggests that perceived auditory emotional experience is distinguishable by the level of perceived activation and pleasantness of a stimulus. Also in agreement with other work, as stimuli increased in perceived activation, the perceived pleasantness between emotions became more distinct (see Figure 4). Emotions that differ in pleasantness were easily categorized into pleasant or unpleasant groupings (especially when activation was high). Bradley and Lang (2007) observed a similar trend in the ratings of the IADS corpus, with a convergence of lower activation affective sounds along the valence dimension. It is important to note that level differences were normalized in this study. The overall intensity of a sound is an indicator of the amount of emotional activation (Laukka et al., 2005),

Figure 4. Dimensional ratings (activation and pleasantness) averaged across listeners for all stimuli in each intended emotion.



and thus differences in perceived activation found in this study may be conservative estimates when compared to ratings obtained using stimuli whose level differences between emotion conditions have been preserved.

Gender Comparisons

Our last hypothesis was that talkers and listeners would vary in their production and perception of emotional speech. Examination of the talker gender effects showed that female talkers were rated as sounding more activated than male talkers. Notably, Sad stimuli from one female talker received ratings of pleasantness that were similar to her peers but received ratings of activation that were much higher when compared to the other female talkers (see Figure 4), which may have influenced these results. However, the Sad sentences were still distinctly grouped from Angry sentences despite the increase in perceived activation, suggestive of good categorical separation, even with the variability in dimensional ratings. Our findings are in agreement with Belin et al. (2008), who reported that female talkers received higher ratings of activation or arousal compared to men; Belin et al. used affective bursts, while the present work utilized speech stimuli, suggesting that the gender bias in perceived activation is robust across different types of stimuli.

Belin et al. (2008) also found listener gender effects only in their ratings of emotion intensity (where females gave higher ratings of intensity compared to males). In this work, we found a similar effect. As previously mentioned, female listeners may demonstrate an advantage when identifying and evaluating emotional stimuli (Hall, 1978; Lambrecht et al., 2014), which could account for their more distinctive dimensional ratings compared to male listeners. While some evidence suggests that females tend to make more negative judgments of negative stimuli than men do (Natale et al., 1983), no evidence of this was found in these stimuli. Ratings of pleasantness were consistent among raters regardless of gender. While dimensional rating differences were, in some cases, small, future work should carefully consider the effects of dimensional ratings based on talker and listener gender.

Limitations

One purpose of this study was to elucidate issues and acknowledgments that must be disclosed when studying auditory emotion. There were limitations of this study in its design and in each of the areas of the emotional communication model (Scherer, 2003). First, the analyses performed employed the raw validation data, rather than using a previously validated source. Thus, the observed relationship between the two emotion classification models may indeed be an artifact of the database stimuli themselves, as they have not been previously assessed categorically or dimensionally. However, we note that the data served their primary purpose of validating this emotional speech set by revealing distinct emotion categories that are well characterized by ratings of activation and pleasantness.

In addition, the study employed a four-alternative forced-choice paradigm without offering listeners an alternative option in the event that none of the suggested emotions was present (Frank & Stennett, 2001; Livingstone & Russo, 2018). This may have resulted in the listeners selecting a "default" emotion each time they were uncertain or unsatisfied with the options presented (e.g., "calm" as it was first on the list and was most commonly selected). This overselection of calm is evident by examining the difference between the raw and unbiased hit rates in Table 2. As such, it is impossible to know with certainty the true recognition rate for calm sentences using these data. However, this limitation was addressed, at least in part, by the inclusion of the unbiased hit rate measure, which accounts for biases in selection of certain categories over others.

Actors, rather than untrained talkers, produced the materials in the MESS. These talkers were instructed on the intended emotional dimensions of the portrayed emotions, which may have introduced bias to their productions and resulted in less natural productions. Preference was given in this instance to the experimental control of the emotional productions rather than to the naturalness of the stimuli (as could have been obtained with untrained talkers or spontaneous speech productions). Future work should consider the importance of experimental control and generalizability when choosing the population of talkers for emotional speech productions.

The stimuli were emotionally neutral in their linguistic content. It is an uncommon occurrence that sentences with such a rigid frame and structure are produced as emotionally as were these stimuli. Again, we reiterate that this was employed for the sake of experimental control. We acknowledge that the semantic meaning of the words in the sentence plays a significant role in the perception of an emotional utterance. Future work must consider the relevance of the stimulus materials for their individual experimental needs.

Lastly, participants in this study were young adults all with normal hearing. Differences in the perception of emotion (whether by category assignment or dimensional ratings) may differ for children and older adults or for those with hearing loss or other auditory limitations (e.g., Dupuis & Pichora-Fuller, 2011, 2015; Luo et al., 2007). This work was meant to lay a foundation for a comparison between two emotional perception models using a sample of healthy younger adults. Future research should consider other listener characteristics when designing their study. Preference may be given to the experimental control of two facets of the communicative model of emotion in order to more fully study the other. For example, it is requisite to strictly control the talker and listener populations when studying the impact of the stimulus type. Allowing for too many talker, stimulus, and listener factors to covary limits the ability to attribute observed effects to any one part of the emotional communicative experience, as effects may be confounded between these three parts of the process.

Summary and Conclusions

Investigators should consider multiple aspects of emotional processes and communication when designing studies involving the production and perception of emotional speech stimuli. Presently, most studies have examined the categorization of emotions into either discrete labels or along some continuum of two or more dimensions. Models of vocal emotion treat the encoding, transmission, and decoding of vocal emotion as connected processes, and researchers should attend to and consider each portion of these processes when designing experiments.

Categorical and dimensional ratings of emotional stimuli were found to be consistent in this work, providing evidence for a hybrid model of emotion classification using speech stimuli. Category selection was deemed by listeners to be most appropriate as the stimuli were perceived to be more dimensionally distinct, suggesting that the assignment of broad emotion category labels is related to the perceived dimensional ratings of a stimulus. Talker and listener factors (i.e., gender) were found to impact the ratings of emotional speech and must be considered in the design of future studies.

We recommend that future researchers carefully consider the aim(s) of their studies as they prepare to develop materials, as the method of emotion elicitation will greatly affect the generalizability and application of their results. Furthermore, we recommend caution when selecting talker, stimulus, and listener variables for a given study. Differences in emotion perception may exist in listeners of different demographic or health backgrounds. Careful control and consideration of talker, stimulus, and listener variables will be imperative when making assertions and additional recommendations based on the results of any study.

Acknowledgments

This study was supported by National Institute on Deafness and Other Communication Disorders Grant R01DC012315, awarded to Eric Hunter. I gratefully acknowledge Rebecca Labowe, Sarah Hargus Ferguson, Skyler Jennings, Zac Imel, Susan Naidu, and Brian Baucom for their assistance in the various aspects of this project from data collection to statistical approach.

References

American National Standards Institute. (2010). *Specifications for audiometers (ANSI S3.6-2010)*. New York, NY: Author.

Arimoto, Y., Ohno, S., & Iida, H. (2011). Assessment of spontaneous emotional speech database toward emotion recognition: Intensity and similarity of perceived emotion from spontaneously expressed emotional speech. *Acoustical Science and Technology, 32*(1), 26–29. https://doi.org/10.1250/ast.32.26

Babchuk, W. A., Hames, R. B., & Thompson, R. A. (1985). Sex differences in the recognition of infant facial expressions of emotion: The primary caretaker hypothesis. *Ethology and Sociobiology, 6*(2), 89–101.

Bachorowski, J.-A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science, 8*(2), 53–57. https://doi.org/10.1111/1467-8721.00013

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The Montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods, 40*(2), 531–539.

Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America, 107*(2), 1065–1066.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry, 25*(1), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Bradley, M. M., & Lang, P. J. (2007). *The International Affective Digitized Sounds (IADS-2): Affective ratings of sounds and instruction manual*. Gainesville: University of Florida.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America, 109*(3), 1101–1109.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley: University of California Press.

Cherry, C. (1953). The cocktail party effect. *The Journal of the Acoustical Society of America, 25*(5), 975–979.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284.

Cowie, R., Cox, C., Martin, J.-C., Batliner, A., Heylen, D., & Karpouzis, K. (2011). Issues in data labelling. In R. Cowie, C. Pelachaud, & P. Petta (Eds.), *Emotion-oriented systems* (pp. 213–241). Berlin, Germany: Springer.

Darwin, C., & Prodger, P. (1998). *The expression of the emotions in man and animals* (3rd ed.). New York, NY: Oxford University Press.

Dupuis, K., & Pichora-Fuller, M. K. (2011). Recognition of emotional speech for younger and older talkers: Behavioural findings from the Toronto emotional speech set. *Canadian Acoustics, 39,* 182–183.

Dupuis, K., & Pichora-Fuller, M. K. (2014). Intelligibility of emotional speech in younger and older adults. *Ear and Hearing, 35*(6), 695–707. https://doi.org/10.1097/aud.0000000000000082

Dupuis, K., & Pichora-Fuller, M. K. (2015). Aging affects identification of vocal emotions in semantically neutral sentences. *Journal of Speech, Language, and Hearing Research, 58*(3), 1061–1076. https://doi.org/10.1044/2015_JSLHR-H-14-0256

Eerola, T., & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music, 39*(1), 18–49.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion, 6*(3–4), 169–200.

Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion Review, 3*(4), 364–370.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572–587.

ELAN. (2017). *ELAN (Version 5.0)*. Nijmegen, the Netherlands: Max Planck Institute for Psycholinguistics. Retrieved from https://tla.mpi.nl/tools/tla-tools/elan

Frank, M. G., & Stennett, J. (2001). The forced-choice paradigm and the perception of facial expressions of emotion. *Journal of Personality and Social Psychology, 80*(1), 75.

Fujimura, T., Matsuda, Y.-T., Katahira, K., Okada, M., & Okanoya, K. (2012). Categorical and dimensional perceptions in decoding emotional facial expressions. *Cognition & Emotion, 26*(4), 587–601.

Goldsmith, H. (1994). Parsing the emotional domain from a developmental perspective. In R. J. Davidson & P. Ekman (Eds.), *The nature of emotion* (pp. 68–73). New York, NY: Oxford University Press.

Goudbeek, M., & Scherer, K. R. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America, 128*(3), 1322–1336. https://doi.org/10.1121/1.3466853

Gray, E. K., & Watson, D. (2001). Emotion, mood, and temperament: Similarities, differences, and a synthesis. In R. L. Payne & C. L. Cooper (Eds.), *Emotions at work: Theory, research and applications for management* (pp. 21–43). Chichester, United Kingdom: Wiley.

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin, 85*(4), 845–857. https://doi.org/10.1037/0033-2909.85.4.845

Helfer, K. S., & Freyman, R. L. (2009). Lexical and indexical cues in masking by competing speech. *The Journal of the Acoustical Society of America, 125*(1), 447–456. https://doi.org/10.1121/1.3035837

Helfer, K. S., & Staub, A. (2014). Competing speech perception in older and younger adults: Behavioral and eye movement evidence. *Ear and Hearing, 35*(2), 161.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: A review. *International Journal of Speech Technology, 15*(2), 99–117.

Lambrecht, L., Kreifelts, B., & Wildgruber, D. (2014). Gender differences in emotion recognition: Impact of sensory modality and emotional category. *Cognition & Emotion, 28*(3), 452–469. https://doi.org/10.1080/02699931.2013.837378

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Laukka, P., Audibert, N., & Aubergé, V. (2012). Exploring the determinants of the graded structure of vocal emotion expressions. *Cognition & Emotion, 26*(4), 710–719. https://doi.org/10.1080/02699931.2011.602047

Laukka, P., & Elfenbein, H. A. (2012). Emotion appraisal dimensions can be inferred from vocal expressions. *Social Psychological and Personality Science, 3*(5), 529–536. https://doi.org/10.1177/1948550611428011

Laukka, P., Juslin, P., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition & Emotion, 19*(5), 633–653. https://doi.org/10.1080/02699930441000445

Laukka, P., Neiberg, D., Forsell, M., Karlsson, I., & Elenius, K. (2011). Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation. *Computer Speech & Language, 25*(1), 84–104. https://doi.org/10.1016/j.csl.2010.03.004

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE, 13*(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Luo, X., Fu, Q.-J., & Galvin, J. J., III. (2007). Vocal emotion recognition by normal-hearing listeners and cochlear implant users. *Trends in Hearing, 11*(4), 301–315. https://doi.org/10.1177/1084713807305301

MATLAB. (2017). *MATLAB Release 2017a*. Natick, MA: The MathWorks, Inc.

Matsumoto, D., & Ekman, P. (2004). The relationship among expressions, labels, and descriptions of contempt. *Journal of Personality and Social Psychology, 87*(4), 529–540. https://doi.org/10.1037/0022-3514.87.4.529

Mehu, M., & Scherer, K. R. (2015). Emotion categories and dimensions in the facial communication of affect: An integrated approach. *Emotion, 15*(6), 798–811. https://doi.org/10.1037/a0039416

Morgan, S. D. (2018). Informational masking and emotion in the speech signal (Doctoral dissertation). The University of Utah, Salt Lake City.

Natale, M., Gur, R. E., & Gur, R. C. (1983). Hemispheric asymmetries in processing emotional expressions. *Neuropsychologia, 21*(5), 555–565. https://doi.org/10.1016/0028-3932(83)90011-8

Picou, E. M., Singh, G., Goy, H., Russo, F., Hickson, L., Oxenham, A. J., … Launer, S. (2018). Hearing, emotion, amplification, research, and training workshop: Current understanding of hearing loss and emotion perception and priorities for future research. *Trends in Hearing, 22,* 2331216518803215. https://doi.org/10.1177/2331216518803215

Quené, H., & van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech*

Communication, 43(1–2), 103–121. https://doi.org/10.1016/j.specom.2004.02.004

Ruffman, T., Henry, J. D., Livingstone, V., & Phillips, L. H. (2008). A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging. *Neuroscience & Biobehavioral Reviews, 32*(4), 863–881. https://doi.org/10.1016/j.neubiorev.2008.01.001

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*(1), 145.

Scherer, K. R. (1987). Toward a dynamic theory of emotion: The component process model of affective states. *Geneva Studies in Emotion, 1,* 1–96.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*(1), 227–256. https://doi.org/10.1016/S0167-6393(02)00084-5

Scherer, K. R. (2005). What are emotions? And how can they be measured. *Social Science Information, 44*(4), 695–729.

Scherer, K. R., Banse, R., Wallbott, H., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion, 15*(2), 123–148.

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.

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological Review, 61*(2), 81.

Schmidt, J., Janse, E., & Scharenborg, O. (2016). Perception of emotion in conversational speech by younger and older listeners. *Frontiers in Psychology, 7,* 781. https://doi.org/10.3389/fpsyg.2016.00781

StataCorp. (2015). *Stata Statistical Software: Release 14* (Version 14). College Station, TX: StataCorp LP.

Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior, 17*(1), 3–28. https://doi.org/10.1007/BF00987006

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America, 52*(4), 1238–1250.