The Hoosier Vocal Emotions Corpus: A validated set of North American English pseudo-words for evaluating emotion processing

Isabelle Darcy[1] and Nathalie M. G. Fontaine[2]

[1]Indiana University; [2]University of Montreal

Acknowledgments: We gratefully acknowledge the Department of Criminal Justice and the Office of Women's Affairs (Women in Science Program) at Indiana University for their financial support (grant to Dr. Fontaine) as well as the participants involved in this study. We also thank Gabriela Cepeda, Franziska Krüger, Pyoung-Hwa Han, Trisha Thomas, Chung-Lin Yang and Joshua Lee for their assistance with the pseudo-word creation, acoustic analyses, participant testing and data analysis. We are indebted to the actresses who took part in the recording sessions. We are further grateful to Dr. Beate Wendt for sharing some of her German stimuli, which were extremely useful in constructing our corpus. Dr. Fontaine is a Research Scholar, Junior 1, Fonds de recherche du Québec – Santé.

Conflicts of interest: None.

Corresponding author:

Dr. Isabelle Darcy,
Department of Second Language Studies, Indiana University
Morrison Hall 231
1165 E. 3rd St.
Bloomington, IN 47405
USA
Phone:+1 812 855 0033
E-mail: idarcy@indiana.edu

Abstract

This article presents the development of the "Hoosier Vocal Emotions Corpus", a stimulus set of recorded pseudo-words based on the pronunciation rules of English. The corpus contains 73 controlled audio pseudo-words uttered by two actresses in five different emotions (i.e., happiness, sadness, fear, anger, disgust) and in a neutral tone, yielding 1,763 audio files. In this article, we describe the corpus as well as a validation study of the pseudo-words. A total of 96 native English speakers completed a forced-choice emotion identification task. All emotions were recognized better than chance overall, with substantial variability among the different tokens. All recordings, including ambiguous stimuli are made freely available, and the recognition rates and the full confusion matrices for each stimulus are provided to assist researchers and clinicians in the selection of stimuli. The corpus has unique characteristics that can be useful for experimental paradigms requiring controlled stimuli (e.g., EEG or fMRI studies). Stimuli from this corpus could be used by researchers and clinicians to answer a variety of questions, including investigations of emotion processing in individuals with certain temperamental or behavioral characteristics associated with difficulties in emotion recognition (e.g., individuals with psychopathic traits), in bilingual individuals or non-native English speakers, in patients with aphasia, schizophrenia or other mental health disorders (e.g., depression), or in training automatic emotion recognition algorithms. The Hoosier Vocal Emotions Corpus is available at http://www.iub.edu/~psyling/hoosiervocalemotions.htm.

The Hoosier Vocal Emotions Corpus: A Data Collection of Recorded Pseudo-words for Evaluating Emotion Processing

The ability to process salient emotional and social cues is critical for adaptive behavior. A failure to process adequately expressions of emotion can have important negative and long-term effects on social behavior and can be a risk factor for adaptation problems, including aggressive and antisocial behavior (Herba & Phillips, 2004). The majority of studies on emotion processing have focused on facial expressions of emotion (e.g., Pollak & Sinha, 2002; Tottenham et al., 2009). There is less research on the vocal expressions of emotion, notably because of the difficulty in obtaining naturalistic recordings of vocal expressions of specific emotions (Scherer, Banse, Wallbott, & Goldbeck, 1991). Still, vocal cues play an important role in the expression of emotions. By "vocal", we refer to "everything that remains present in a spoken message after lexical and syntactic information has been removed" (van Bezooijen, 1984, p. 1). A growing number of studies conducted in the past decade indicate that humans, across languages and cultures, can infer emotion from vocal expression alone because of differential acoustic patterns (e.g., Banse & Scherer, 1996; Bänziger, Mortillaro & Scherer, 2012; Castro & Lima, 2010; Juslin & Laukka, 2003; Livingstone & Russo, 2018; Liu & Pell, 2012; Pell, Paulman, Dara, Alasseri, & Kotz, 2009; Sauter, Eisner, Ekman, & Scott, 2010; Scherer, Banse & Wallbott, 2001).

A number of emotion corpora have been produced (see Ververidis & Kotropoulos, 2006; Scherer, Clarke-Polner, & Mortillaro, 2011, for reviews). They all have their particular features and are composed of diverse vocal stimuli. Table 1 presents a sample of data collections of vocal expressions of emotion.

=== Please insert Table 1 about here ===

We developed and validated a set of pseudo-words based on the phonology and pronunciation rules of North American English, which we aim to make available to the research and clinical communities. The corpus, named the Hoosier Vocal Emotions Corpus (HVEC), includes important unique characteristics. First, it focuses on disyllabic pseudo-words, rather than meaningful words or sentences, to remove the semantic meaning and allow for the speech prosody to become the central attribute of emotion processing (Wendt & Scheich, 2002; Wendt et al., 2003). To our knowledge, only one other corpus (the 'Magdeburger Prosodie Corpus', a set of stimuli respecting the phonotactic and phonetic rules of the German language) includes isolated pseudo-words (Wendt & Scheich, 2002; Wendt et al., 2003). Our corpus' main features are based on this German corpus. Other corpora of vocal emotions contain pseudo-sentences (e.g., Castro & Lima, 2010; Liu & Pell, 2012). However, experimental paradigms can require shorter stimuli, which would be difficult to manually extract from sentences and subsequently validate separately. In addition, Rigoulot, Wassiliwizky, and Pell (2013) demonstrated in a gating paradigm study that the length of stimuli matters for the time course of emotion recognition, and that full sentences are recognized much more easily than truncated ones. Other corpora use affect bursts (e.g., "ah") or emotional sounds such as screams or laughter (e.g., Belin et al., 2008; Parsons, Young, Craske, Stein, & Kringelbach, 2014). Despite the high effectivity of such stimuli to convey specific emotions, they are also not necessarily suitable for experimental paradigms requiring controlled stimuli with medium or normal emotional intensity.

The Hoosier Vocal Emotions Corpus includes 73 controlled audio pseudo-words uttered by two actresses in five different positive or negative emotions (i.e., happiness, sadness, fear, anger, and disgust) and in a neutral tone, yielding 1,763 stimuli (some stimuli were pronounced more than two times). We selected the emotions based on the basic emotions identified by Ekman (1992), except for surprise, because this emotion can have any valence (it can be neutral, positive or negative). In addition, surprise utterances can be difficult to simulate experimentally (Pell, Paulmann, Dara, Alasseri, & Kotz, 2009). Although concerns have been raised about the use of acted rather than natural stimuli (Bachorowski & Owren, 2008), there are also arguments suggesting that actors can produce realistic portrayals and valid instances of vocal expressions of emotion (Ververidis & Kotropoulos, 2006). One important argument is that much of our verbal communication is subject to sociocultural censure and involves making impressions on others (Bachorowski & Owren, 2008; Banse & Scherer, 1996). Therefore, having people utter an emotion as if they were experiencing it may not be significantly different from a real-life communicative situation. Two female voices were preferred over having one male and one female voice, mainly for reasons of comparability and homogeneity between acoustic dimensions such as pitch range, and to facilitate their use in experimental paradigms requiring tight control of acoustic parameters of stimuli, such as event-related potential (ERP) studies. In this article, we describe the structure of the Hoosier Vocal Emotion Corpus, as well as the validation of the pseudo-words in terms of the emotion they portray. We also discuss potential applications of this set of stimuli.

Methods

Creation of the Stimuli

The stimulus set is composed of pseudo-words based on real English words. These pseudo-words were created by selecting common English disyllabic words using the COBUILD frequency information (per million) from the CELEX English Wordforms database (Baayen, Piepenbrock & Gulikers, 1995), and manipulating the order of segments within the word (see Wendt & Scheich, 2002, or Castro & Lima, 2010, for a similar procedure). For example, the pseudo-word "elby" was constructed from the noun *belly*. As a result, there is no clear phonetic relationship between the pseudo-words and their originals, but they are matched in terms of number of syllable and phonemes. Care was taken to ensure that the pseudo-words were phonotactically legal, that is, that the sequences of phonemes were permitted and easily pronounceable in English. Similarly, slight phonetic adjustments were made to comply with English pronunciation rules. For example, the pseudo-word "domner", based on *modern*, did not retain the flapped /d/ found in the North American English pronunciation of *modern*, since the flap is not found in word-initial position in English. Pseudo-words that were too clearly reminiscent of their original or of other real words were excluded. A final list of 73 pseudo-words was generated (see Table 2). Stress always fell on the first syllable but the vowel in the second syllable was not always fully reduced (indicated by the International Phonetic Alphabet [IPA] symbols in Table 2, where only "schwa" [ə] represents a reduced vowel). The transcription provided in Table 2 most closely reflects the actual pronunciation of *most* of the stimuli by both actresses. Since each actress pronounced a given pseudo-word 12 times (2 x 6 emotions), there are essentially 24 pronunciations of the same pseudo-word, thus displaying some variation from one token to the next. The transcription here reflects the most common pronunciation of the stimuli, and there might be some variation across specific stimuli, especially in terms of the vowels. Table 2 is provided here to give further guidance to researchers about the possible variations in pronunciation for the same pseudo-word, but we encourage researchers and clinicians who need an exact control of sound properties to check each stimulus they plan to use.

=== Please insert Table 2 about here ===

Elicitation and Recording Procedures

Two actresses were recruited to record the 73 pseudo-words in a neutral tone as well as in five different modal emotions: happiness, sadness, fear, anger, and disgust. Female voices were recorded as the basis of another experiment (i.e., an EEG paradigm involving young children; Hoyniak et al., 2018). Both actresses were native speakers of Midwestern United States English (North Midland dialect region, Clopper & Pisoni, 2004), and had lived exclusively in that region prior to the recording. They reported no fluency in any language other than English, and have not lived abroad. They were students in the Department of Theatre and Drama at a large Midwestern higher education institution (Indiana University, Bloomington, IN), and were 18 and 20 years old, respectively, at the time of recordings. Both actresses were paid, and gave consent to share the recordings in a publicly accessible database.

Each actress (henceforth AG and KM) was recorded individually in a single session of approximately 1.5 to 2 hours. The experimenter first briefly explained the general procedures to each actress, who was also given time to familiarize herself with the list of stimuli. Pronunciation of the pseudo-words was clarified as needed. The different emotions were discussed and explained. Stimuli were elicited using a short sentence preceding the pseudo-word: ‘it starts like /word/, I say /pseudo-word/, I say /pseudo-word/ again’ (see Table 3). This was done to help maintain consistent pronunciation of the pseudo-words, and enhance fluent delivery and more natural sounding speech. In addition, this form of elicitation was chosen to enable a similar delivery context for each pseudo-word across emotions and ensure high comparability. Each pseudo-word was thus pronounced at least twice (2 times per carrier sentence). For each actress, at least 146 stimuli were pronounced for each emotion, yielding a total of at least 876 stimuli per actress. However, some stimuli were pronounced more than two times when an actress chose to reattempt the emotion portrayal for a given carrier sentence, resulting in a total of 876 pseudo-words for AG and 887 pseudo-words for KM, for a grand total of 1,763 audio files. The stimuli are overall similar in terms of duration (*M* = 613 ms; *Median* = 608 ms; *SD* = 132 ms) and intensity (*M* = 62.29 dB; *Median* = 62.23 dB; *SD* = 3.849 dB).

Actresses were allowed to choose the order in which they preferred to utter each emotion. They were then seated in a recording booth, wearing Sennheiser HD515 Dynamic Stereo headphones, and before recording a set, were shown a short presentation of pictures and auditory examples of (non-English) pseudo-words spoken in the corresponding emotion (Wendt & Scheich, 2002). The pictures depicted situations in which examples of the specific emotion to be uttered were displayed. For example, various clip art pictures of angry individuals, arguing friends and knit eyebrows were shown to illustrate anger, and to clarify a general mood for each emotion. The experimenter demonstrated a few items in their carrier sentences (without modeling a particular emotion), to help with pronunciation of stimuli (fluency) and overall rhythm. Actresses were also encouraged to imagine situations/scenarios according to the emotion to be expressed. They were given as much time as they needed to “get into the character” of the emotion before proceeding with recordings. The experimenter also instructed the actresses not to exaggerate their expressions of the emotions, to achieve a “normal” level of emotional intensity, rather than “strong” (see Livingstone & Russo, 2018).

=== Please insert Table 3 about here ===

The stimuli were recorded in a noise-isolated recording booth, at a sampling rate of 44,100 Hz with a 16 bit resolution on a mono channel, using a Sennheiser e835 dynamic cardioid microphone and an Edirol UA25 USB stereo audio interface. The distance and orientation of the actresses to the microphone were held as constant as possible. Each stimulus (pseudo-word) was then manually cut from its sentence context and saved separately in a .wav format for presentation in the subsequent evaluation procedures.

We conducted a validation study with approximately 25 participants rating each sound file of the Hoosier Vocal Emotion Corpus, to estimate to what extent each recorded stimulus represents an acceptable rendition of the intended emotion. We included stimuli from both actresses into the corpus validation, that is, a total of 1,763 audio files. Given the large number of audio files, the time required for a single listener to evaluate all of them would have been prohibitively long. We therefore divided the files into four stimuli lists, which were presented to listeners for evaluation. All emotions were equally balanced in each list. However, we decided against mixing the two voices in each list (see Castro and Lima, 2010, for a similar design). Each list contained stimuli from only one speaker (lists 1 and 2: AG, list 3 and 4, KM). This was done in order to reduce comparison between voices, and to enhance the reliance on actual acoustic properties of the stimuli. An additional consideration was the cognitive load of this task, which is demanding for participants. Each participant rated only one list. The dataset accompanying the corpus contains ratings for each audio file from about 25 persons (see below for method details). All procedures were approved by the Indiana University Institutional Review Board.

Validation of the Stimuli

Procedure: In order to validate the stimuli of the corpus, we opted for a forced-choice identification task similar to the one used by van Bezooijen (1984) or Castro & Lima (2010). Stimuli were presented to listeners via headphones, using the Praat software (version 5.4.04; Boersma & Weenink, 2014) on computers running under Windows 7. Participants were tested individually and were seated at a computer station in a partitioned computer lab, wearing high-quality Sanako over-the-ear headphones at a self-chosen comfortable listening level. Their task was to listen to each sound file and identify what emotion they thought the speaker intended to convey. They were asked to choose one out of six possible emotions and indicate their choice by clicking on the correspondingly labeled button on the screen. The labels were ‘neutral’, ‘happy’, ‘sad’, ‘fear’, ‘angry’, and ‘disgust’. There was no “Other/None of the above” option (Livingstone & Russo, 2018). Participants were also asked to choose how confident they were in their choice, by clicking on a number on a scale ranging from 1 (not sure) to 5 (very sure). The instructions were displayed on the screen as follows:

> This is a judgment experiment about how actors convey emotions. You will hear an actress say non-words and your task is to choose what emotion you think it conveys. (Some non-words might be repeated a few times). Please don’t spend too much time on each non-word. Try to do it using your intuition.
>
> In addition, we ask that you indicate how confident you are with your choice on a scale of 1 (not sure) to 5 (very sure). There are several breaks.
>
> If you have questions, please ask now.

The buttons appeared as rectangles on a single line in the middle of the screen, and their order was randomly varied across list (but kept constant for any given participant) to avoid preference effects. The task was not timed, and listeners could replay the sound up to 8 times by clicking on a repeat button (Figure 1).

==== Please insert Figure 1 (top and bottom panels) about here ====

The presentation order of the sound file was randomized for each participant, and the script implemented a break after every 50 stimuli. No stimulus file was repeated. The average duration of the identification task was about 45 minutes. As explained above, the sound files were divided into four lists to keep the duration manageable for a single participant. Each of the 4 lists contained roughly the same number of stimuli: lists 1 and 2 (AG) contained 438 sound files; List 3 contained 443 sound files, and list 4 contained 444 sound files (KM). Participants were randomly assigned to one list upon arrival in the testing room. All participants also filled out a sociodemographic questionnaire (notably to assess age, sex, and languages spoken) administered through the Qualtrics survey software.

Participants: 102 participants were tested. The testing took place between February 2016 and December 2016. Six participants were excluded for various reasons (not native speakers of English or did not grow up in the USA; multiple neuro-cognitive issues reported; incomplete dataset; technical failure; more than twice the average time needed to complete the task). In total, data from 96 participants (67% female) aged between 18 and 38 years old (M = 21.09, SD = 3.21) were included in the analysis (list 1: N = 24; list 2: N = 25; list 3: N = 24; list 4: N = 23). Most of the participants were college students and were predominantly Caucasian. Only one participant reported not knowing any language other than English. Twelve participants reported growing up bilingually using English and another language. About half of the participants (53.1%) reported knowledge of Spanish, 21.9% of French, and 6.3% of German, with 13 other languages mentioned by less than 4% of the participants (e.g., Japanese, 3.1%). A total of 34.4% of the participants reported knowing two languages besides English, and 12.5% reported knowing three languages besides English. Two participants reported knowing four or more languages besides English. Aside from the early bilinguals, three participants reported high proficiency in other languages learned after the first. None reported having any kind of uncorrected speech or hearing disorder. We recruited the participants using flyers posted in public areas (e.g., various departments at Indiana University) and word of mouth. Participants were compensated for their time.

Results

To ascertain the validity of the corpus, we used two dependent variables: emotion identification accuracy rates and confidence scores (how confident the participants were in their choice). Response time (RT) was collected on each trial but is not analyzed as a dependent variable given that the task was not speeded. Because there were six choice options on each trial, a random selection would yield an overall accuracy of 16.7%. Data were submitted to a chi-square analysis to estimate whether or not the participants were equally likely to choose among the six possibilities for a given stimulus. Table 4 provides the confusion matrix overall, across both speakers, and reveals that overall, emotion portrayals were recognized accurately. Figure 2 shows the overall median accuracy in emotion identification by the 96 participants, separated by speaker. Random performance level (~16%) is indicated by the dotted line.

=== Please insert Table 4 about here ===

=== Please insert Figure 2 about here ===

Figure 2 suggests that participants were able to identify each stimulus' intended emotion above chance[1]. The mean recognition accuracy is 45%. Sadness was recognized most accurately (*M* = 59%), followed by neutral (*M* = 51%), fear (*M* = 50%), disgust (*M* = 43%), and anger (*M* =

[1] The pattern of accuracy remained the same even after removing very slow and very fast trials (RT outliers, defined as data points that were more than 2.5 SD beyond all participants' mean RT, or faster than 100 ms; 3.24% of data were removed). The slow RTs on some trials are likely the result of the option of listening to the stimuli multiple times and of the fact that the task was not speeded.

38%). The emotion that was recognized least accurately was happiness (M = 31%). All emotions were recognized better than chance for both stimuli sets, except for happiness for KM stimuli, which was misidentified as neutrality more often than it was identified as happiness (see Table 5b).

A global chi-square analysis on the chosen response categories over all data points (across emotions and speakers) was significant ($\chi(25) = 29{,}429.29$; $p < .001$, Cramer's V = .37). This suggests that for each emotion, respondents did not randomly choose among the six options. Before evaluating whether this pattern holds for each emotion separately, we first examined whether there is a difference in accuracy between speakers, as suggested in Figure 2.

A one-way ANOVA comparing accuracy for each speaker (KM, AG) reveals that mean recognition accuracy for AG (M = 48%, 95%CI = 40 – 45) is significantly higher than for KM (M = 43%, 95%CI = 45 – 50), $F(1, 574) = 8.26$, $p = .004$). This significant effect of speaker indicates that raters were overall slightly more accurate at recognizing emotions portrayed by one speaker (AG) over the other (KM). However, such differences are to be expected among voice actors, and is unlikely to reflect an inherent difference among our listener groups. If one group of listeners were systematically less concentrated or accurate during the task, we would expect this difference to hold across emotions for a given speaker. To verify this, a mixed effects model declaring speaker and emotion as fixed factors (and participants as random factor) was conducted in SPSS 25. Multiple comparisons were adjusted with the Sidak correction. The type III tests of fixed effects shows a main effect of speaker ($F(1, 94) = 8.6$, $p = .004$), a main effect of emotion ($F(5, 470) = 37.7$, $p < .001$), and crucially, a significant interaction between the two factors ($F(5, 470) = 33.4$, $p < .001$). The interaction and pairwise comparisons reveals that for all emotions except Disgust and Neutral, AG's portrayals were recognized significantly more accurately than KM's; conversely, KM's portrayals of Disgust and Neutral were recognized significantly more accurately than AG's. The presence of an interaction suggests that it is unlikely to be the case that the KM listeners were systematically less accurate than the AG listeners (otherwise, one would have expected an absence of interaction).

Tables 5a and 5b provide the confusion matrix obtained for our stimulus set (emotion portrayal by participants' choices; *n* = 42,306 data points), separated by speaker.

=== Please insert Table 5a and 5b about here ===

Given the significant effect of speaker and the speaker by emotion interaction, we further conducted a series of Chi-square analyses (non-parametric goodness-of-fit tests) in SPSS 25 for each speaker and emotion separately, which confirms the global analysis. The results of the tests for each emotion and each speaker are provided in table 5a and 5b. They show that the tests were significant for all speakers and all emotions, indicating that listeners were not responding randomly.

The examination of the patterns of misidentifications in Tables 5a and 5b revealed the following tendencies. For AG, all emotions except fear were most often misidentified as neutral, which represents the second highest proportion of choices in these cases. In the case of fear, items were misidentified most often as happiness. However, even though for instance happiness was misinterpreted as neutral in 25% of the cases for AG, the reverse is not true: neutral items only were misinterpreted as happiness in 7% of the cases, and were more commonly misinterpreted as anger or sadness, each in roughly 15% of cases (see Table 5a). The error patterns for KM stimuli stand out in that happiness stimuli were most often recognized as neutral, which is the dominant, modal response. Happiness choices were given in 25% of cases, neutral choices in 32%. For the other emotions, unlike for the AG stimuli, neutral is the second choice after the correct identification only for anger and sadness. Disgust is misidentified as anger in 21% of cases, more often than neutral, and fear is confused with sadness in 27% of cases (see Table 5b).

This overall high proportion of neutral choices is possibly due to the fact that the stimuli were created at a medium/normal intensity level, without emotional exaggeration, rendering the identification task potentially more difficult. In order to help researchers evaluate how ambiguous a given recording is, we also provide the full confusion matrix for each stimulus in the database (see Bänziger et al., 2012, supplemental materials, for a similar approach).

Some particular items were identified at very high accuracy rates by all participants who rated it, and conversely, others were almost never correctly identified. Figure 3 shows the accuracy variance obtained for each stimulus (each sound file in the corpus represents one dot). The boxplots 3a and 3b show the distribution and median accuracy for each emotion (3a for AG, 3b for KM). The figures reveal that a proportion of items (particularly for happiness) fell below the random performance level (i.e., 16.7%) – suggesting that these particular stimuli are ambiguous and not ideal representations of the intended emotion, at least for the participants who rated the stimuli.

=== Please insert Figures 3a-b about here ====

We also obtained confidence ratings for each stimulus rated (i.e., how confident the participants were in their choice). The correlations coefficients (Pearson *r*) are provided at the top of each figure. Only one relationship (neutral for AG) was not significant. Figure 4 shows the correlations between accuracy of identification and the confidence ratings of the participants for each emotion and each speaker separately.

=== Please insert Figure 4 about here ====

Discussion

The goal of this project was to create a corpus of auditory pseudo-words uttered in different emotions. The corpus includes 73 controlled audio pseudo-words uttered by two actresses in five different emotions (i.e., happiness, sadness, fear, anger and disgust) and in a neutral tone, yielding at least 876 stimuli per actress. In addition, the pseudo-words are based on the pronunciation rules of North American English and they are not a caricature or an exaggeration of the emotions portrayed. Each recording has been validated by native English listeners in terms of recognition accuracy of the intended emotion portrayal. Overall, the emotions were recognized at accuracy levels that were clearly higher than chance (M = 45% across emotions and speakers, for a chance level at about 16%). The recognition proportions obtained for our data were most accurate for fear, neutral, and sadness, and least accurate for happiness and disgust, consistent with previous data in other languages (Banse & Scherer, 1996; Castro & Lima, 2010; Liu & Pell, 2012; Pell, Paulmann, et al., 2009; Scherer et al., 1991; van Bezooijen, 1984). The one exception is anger. In our stimuli, anger was recognized with surprisingly low accuracy (38%). It is often among the best recognized emotions (e.g., Scherer, Clark-Polner, & Mortillaro, 2011; Bänziger et al., 2012; Wendt & Scheich, 2002). This effect is possibly due to the fact that our pseudo-words were produced with a medium/normal emotional intensity level, possibly making them more confusable with neutral stimuli. Indeed, for both speakers (and particularly for KM), anger was most often confused with neutrality. The resulting accuracy in our dataset is globally similar to the ones reported in previous studies on vocal emotion (hovering in the 40-60% range, see Scherer et al., 2011), in particular among the studies that used similar stimuli (words or short sentences, such as Rigoulot et al., 2013) and a similar number of response options.

The audio stimuli were created as high-quality recordings in a .wav format, which allows experimenters to run more detailed acoustic analyses in order to match stimuli for specific experimental purposes. For instance, intensity (as loudness, in dB) and duration measurements (in ms) are provided in the corpus database, but other acoustic parameters can be extracted, such that matched stimuli could be selected for the needs of an EEG study for instance.

The corpus is available at http://www.iub.edu/~psyling/hoosiervocalemotions.htm. The website provides basic information about the corpus and how to request access to the sound files and the database. For each item, recognition accuracy and confusion patterns, as well as speaker, filename, and a number of acoustic details, are provided in an accompanying database, in order to allow researchers to select items specifically for their needs. The list of attributes provided for each sound file in the corpus is detailed in the appendix.

A number of methodological issues need to be considered. First, the validation of the stimuli was based on data collected in a laboratory setting using a forced-choice methodology with six response alternatives. Even though this methodology is commonly used across studies, its ecological validity for real-time interactions in social situations remains limited. It is unclear to what extent these results would generalize to real life situations outside the laboratory, or to experimental paradigms where a given stimulus is presented only once without any available "categorization labels," because forced-choice procedures produce better performance than free-choice tests (see Bachorowski & Owren, 2008). Second, a similar consideration involves the specific linguistic context in which an emotion is heard, and the type of linguistic materials used. Hearing a short (2-syllables) pseudo-word to identify an emotion is likely much more difficult than identifying it via a longer, meaningful sentence (see Rigoulot et al, 2013), and is likely to lead overall to lower recognition accuracy. Similarly, medium/normal emotional intensity (as opposed to high, such as in affect bursts) is likely to make emotion recognition less straightforward. Taken together, the identification accuracy we obtained in our study is the product of the forced-choice methodology on the one hand and of the medium/normal intensity of the stimuli, the fact that they are pseudo-words presented in isolation, and the context-free format of their presentation in the recognition task.

Third, the corpus includes a limited set of emotions (i.e., happiness, sadness, fear, anger, and disgust) and a neutral tone. Other emotions could have been included (i.e., surprise and contempt). We selected the emotions to be included in the corpus based on the basic emotions identified by Ekman (1992) and based on whether they can have either a positive or a negative valence. Therefore, surprise was not included because it can have any valence (it can be neutral, positive or negative) and also because this emotion can be difficult to simulate in the laboratory (Pell et al., 2009). Researchers and clinicians should then consider this limitation when selecting this corpus for their work as well as the fact that that only one positive emotion (i.e., happiness) is included, which would impede systematic analyses of valence effects and the examination of different positive emotions. Forth, they should also consider that the corpus contains pseudo-words uttered by two females (i.e., it does not include male voices). Finally, because the validation of the stimuli was based on a between-design (i.e., each participant rated the pseudo-words from one actress only), it is hard to establish differences in the validation between the two speakers. Although this design may be seen as a limitation due to logistics, the information we provide in the corpus should enable the researchers and the clinicians to make informed decisions as to what stimuli to select for their work.

Potential Applications of the Hoosier Vocal Emotions Corpus

The Hoosier Vocal Emotions Corpus was specifically developed for the requirements of EEG research on emotion processing. Stimuli from this corpus were first used in a study on the neural responses (using EEG techniques) to vocal emotion processing and their associations with temperamental traits and behavioral problems in young children (Hoyniak et al., 2018). The corpus has unique characteristics that are useful for experimental paradigms requiring controlled stimuli (e.g., EEG or fMRI studies), namely disyllabic pseudo-words (i.e., short stimuli without a semantic meaning) that are overall similar in terms of duration and loudness, and that represent medium/normal emotional intensity.

To the best of our knowledge, the Magdeburger Prosodie Corpus (Wendt & Scheich, 2002; Wendt et al., 2003) is the only other corpus that includes isolated disyllabic pseudo-words. However, this corpus is composed of stimuli respecting the phonotactic and phonetic rules of the German language. Although there are data suggesting that emotions can be recognized across languages and cultures, there is still an in-group advantage in the processing of emotional vocalizations (Sauter et al., 2010). We therefore developed new emotional vocalizations based on the phonology and pronunciation rules of North American English for research and clinical work requiring English-based stimuli. The use of the corpus does not need to be limited to English speakers, however. For instance, studies of emotion or prosodic processing in monolingual or in multilingual individuals, or in non-native English speakers, could be easily conducted using stimuli from this corpus (e.g., Dewaele, 2004; Min & Schirmer, 2011; Paulmann & Uskul, 2014).

Stimuli from the corpus could also be used to investigate emotion processing in individuals with certain temperamental or behavioral characteristics associated with difficulties in emotion recognition (e.g., individuals with psychopathic traits or alexithymia). In addition, the stimuli could be used to study the extent to which patients with aphasia, schizophrenia or other mental disorders (e.g., depression) are able to process prosodic/vocal emotion information.

The HVEC's short, disyllabic pseud-words, which are acoustically more homogenous than longer sentences, can also be useful to researchers performing acoustic analyses. Investigations that seek to characterize the prosodic and acoustic features of different emotions would benefit from this kind of tightly controlled and not exaggerated materials, since they can help isolate specific acoustic parameters for emotion recognition more precisely. Also, the fact that our stimuli are produced with normal emotional intensity (as opposed to high, such as in affect bursts) contributes to creating more ambiguity in the corpus, and makes emotion recognition not only less straightforward, but possibly also more ecologically valid. Ambiguous or subtle acoustic characteristics can be studied with a corpus that preserves this variability like ours, and because we provide the full confusion matrix for each stimulus, researchers seeking to determine the acoustic parameters of various emotions will have a large range of clear, ambiguous, and misclassified stimuli to choose from. This variability and the range of stimulus uncertainty is also very useful for the field of automatic emotion recognition. Training paradigms would thus be able to first use the non-ambiguous stimuli (see Brendel et al., 2010) and progressively incorporate more subtle stimuli, ultimately leading to robust recognition scores.

Finally, the neutral tone stimuli can be used on their own for research applications other than emotional processing. For instance, they could be used for pseudo-word or voice recognition tasks in investigations of individual differences in auditory, phonetic or phonological processing or learning.

Conclusion

In this paper, we presented the Hoosier Vocal Emotion Corpus, a set of controlled disyllabic pseudo-words in five basic emotions and in a neutral tone. This corpus is one of the few databases of pseudo-words vocal expressions for North American English. The corpus consists of 1,763 high-definition audio recordings by two female speakers at a medium/normal emotional intensity level. The validation of the corpus with a forced-choice recognition paradigm revealed high rates of emotional validity. The recognition accuracy for each item as well as the full confusion matrix are provided in an accompanying database, which allows researchers to explore the full range of stimulus uncertainty. Despite some of the limitations discussed above, this corpus presents a valuable resource for a wide variety of researchers and clinicians.

Open Practices Statement

The validation study was not pre-registered. The data and materials for all experiments are available at http://www.iub.edu/~psyling/hoosiervocalemotions.htm.

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Bachorowski, J-A., & Owren, M. J. (2008). Vocal expressions of emotion. In: M. Lewis, J. M. Haviland-Jones, & L. Feldman Barrett (Eds.), *Handbook of emotions* (3rd Ed.) (pp. 196-210). NY: The Guildford Press.

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal Expression Corpus for experimental research on emotion perception. *Emotion*, *12*, 1161-1179.

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

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*, 531-539.

Bezooijen, R. A. M. G. van (1984) *Characteristics and recognizability of vocal expressions of emotion*. Dordrecht: Foris Publications. Retrieved on January 30, 2018 from https://repository.ubn.ru.nl /handle/2066/114117

Boersma, P., & Weenink, D. (2014). *Praat: doing phonetics by computer* (Computer program). Version 5.3.35.

Brendel, M., Zaccarelli, R., Schuller, B., & Devillers, L. (2010). Towards measuring similarity between emotional corpora. *Proceedings of the International Conference on Language Resources and Evaluation*. European Language Resources Association, Valletta, Malta, pp. 58-64.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A database of German emotional speech. *Proceedings of the Interspeech*, Lisbon, Portugal, pp. 1517-1520.

Castro, S. L., & Lima, C. F. (2010). Recognizing emotions in spoken language: A validated set of Poruguese sentences and pseudosentences for research on emotional prosody. *Behavior Research Methods*, *42*, 74-81.

Clopper, C. G., & Pisoni, D. B. (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of phonetics*, 32, 111-140.

Costantini, G., Iadarola, I., Paoloni, A., & Todisco, M. (2014). Emovo corpus: An Italian emotional speech database. *Proceedings of the International Conference on Language Resources and Evaluation*. European Language Resources Association, Reykjavik, Iceland, pp. 3501-3504.

Dewaele, J.-M. (2004). The emotional force of swearwords and taboo words in the speech of multilinguals. *Journal of Multilingual and Multicultural Development, 25*, 204-222.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*, 169-200.

Herba, C., & Phillips, M. (2004). Annotation: Development of facial expression recognition from childhood to adolescence: Behavioral and neurological perspectives. *Journal of Child Psychology and Psychiatry*, *45*, 1185-1198.

Hoyniak, C.P., Bates, J.E., Petersen, I.T., Yang, C.-L., Darcy, I., & Fontaine, N.M.G. (2018). Reduced neural responses to vocal fear: A potential biomarker for callous-uncaring traits in early childhood. *Developmental Science*, *21*, e12608.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: different channels, same code? *Psychological Bulletin, 129*, 770-814.

Laukka, P., Elfenbein, H. A., Chui, W., Thingujam, N. S., Iraki, F. K., Rockstuhl, T., & Althoff, J. (2010). Presenting the VENEC corpus: Development of a cross-cultural corpus of vocal emotion expressions and a novel method of annotating emotion appraisals. *Proceedings of the International Conference on Language Resources and Evaluation*. European Language Resources Association, Valletta, Malta, pp. 53-57.

Lima, C.F., Castro, S. L., & Scott, S. K. (2013). When voices get emotional: A corpus of nonverbal vocalizations for research on emotion processing. *Behavior Research Methods*, *45*, 1234-1245.

Liu, P., & Pell, M. D. (2012). Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal emotional stimuli. *Behavior Research Methods*, *44*, 1042-1051.

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*, e0196391.

Min, C. S., & Schirmer, A. (2011). Perceiving verbal and vocal emotions in a second language. *Cognition and Emotion, 25*, 1376-1392.

Parsons, C. E., Young, K. S., Craske, M. G., Stein, A. L., & Kringelbach, M. L. (2014). Introducing the Oxford Vocal (OxVoc) Sounds database: a validated set of non-acted affective sounds from human infants, adults, and domestic animals. *Frontiers in Psychology, 5*(562).

Paulmann, S., & Uskul, A. K. (2014). Cross-cultural emotional prosody recognition: Evidence from Chinese and British listeners. *Cognition and Emotion, 28*, 230-244.

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics*, *37*, 417-435.

Pollak, S. D., & Sinha, P. (2002). Effects of early experience on children's recognition of facial displays of emotion. *Developmental Psychology*, *38*, 784-791.

Rigoulot, S., Wassiliwizky, E., & Pell, M. D. (2013). Feeling backwards? How temporal order in speech affects the time course of vocal emotion recognition. *Frontiers in Psychology*, *4*(367).

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences of the United States of America, 107*, 2408-2412.

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

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, *32*, 76-92.

Scherer, K. R., Clarke-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*, 401–435.

Tottenham, N., Tanaka, J., Leon, A.C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D. J., Westerlund, A., Casey, B. J., & Nelson, C. A. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, *168*, 242-249.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, *48*, 1162-1181.

Wendt, B., Hufnagel, Brechmann, A., Gaschler-Markefski, B., Tiedge, J., Ackermann, H., & Scheich, H. (2003). A method for creation and validation of a natural spoken language corpus used for prosodic and speech perception. *Brain and Language*, *87*, 187.

Wendt, B., & Scheich, H. (2002). The “Magdeburger Prosodie-Korpus”. P*roceedings Speech Prosody Conference* (pp. 699-701). Aix-en-Provence, France.

Table 1. Sample of data collections of vocal expressions of emotion

References	Name/description of the data collection	Language/	Speakers	Type and number of vocal stimuli	Kind of speech	Emotions (in terms used by authors)	Other perceptual modalities
Bänziger, Mortillaro, & Scherer (2012)	Geneva Multimodal Emotion Portrayals Core Set (GEMEP-CS)	Non-language (pseudospeech sentences and a nonverbal vocalization; "*aaa*" by French speakers)	5 women and 5 men (professional French-speaking theater actors)	145 emotion expressions (pseudospeech sentences)	Acted speech	17 emotions (e.g., amusement, despair, hot anger, fear/panic, joy/elation, sadness, contempt, disgust, surprise)	Video (i.e., presentation of dynamic picture without sound) and audio-video (i.e., presentation of dynamic picture and sound)
Belin, Fillion-Bilodeau, & Gosselin (2008)	Montreal Affective Voices (MAV)	Nonverbal affect bursts using the French vowel "*ah*"	10 different actors (5 women and 5 men)	90 nonverbal affect bursts	Acted speech	Anger, disgust, pain, sadness, surprise, happiness, pleasure and neutral	-

Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss (2005)	Berlin Emotional Speech Database (EMO-DB)	German	5 women and 5 men	10 meaningful sentences by 6 emotions (plus the neutral state) by 10 actors, in addition to some second versions (*n* = about 800 sentences)	Acted speech	Anger, fear, joy, sadness, disgust, boredom and neutral	-
Castro & Lima (2010)	Set of Portuguese sentences and pseudosentences	European Portuguese	2 women	16 Portuguese sentences and 16 pseudosentences by 6 emotions (plus the neutral state) Mean length = 8 syllables (range 6-11)	Acted speech	Happiness, sadness, anger, fear, disgust, surprise and neutral	-
Costantini, Iadarola, Paoloni, & Todisco (2014)	EMOVO Corpus	Italian	6 actors (3 women and 3 men)	14 sentences by 6 emotions (plus the neutral state) by 6 actors (588 sentences)	Acted speech	Disgust, joy, fear, anger, surprise, sadness and neutral	-
Laukka et al. (2010)	Vocal Expressions of Nineteen Emotions across Cultures (VENEC)	English	100 professional actors from 5 English speaking cultures (USA,	About 6500 vocal expressions (mainly short phrases with emotionally	Acted speech	19 emotions (e.g., amusement, anger, contempt, disgust, distress, fear, guilt, happiness, shame) and	-

			India, Kenya, Singapore and Australia) (50% women)	neutral content, expressed in three levels of intensity)		neutral	
Liu & Pell (2012)	A database of Chinese vocal emotional stimuli	Pseudo-sentences (semantically meaningless and relatively plausible as Chinese sentences)	10 native Mandarin speakers (5 women and 5 men)	35 pseudosentences by 6 emotions (plus the neutral state)		Anger, disgust, fear, sadness, happiness, pleasant surprise and neutral	-
Lima, Castro, & Scott (2013)	A corpus of nonverbal vocalizations	Nonverbal vocalizations by European Portuguese native speakers	4 speakers (2 women and 2 men) who did not have formal acting training.	121 sounds (no guidance was provided as to the specific kind of sounds the speakers had to make)	Acted speech	4 positive states (achievement/triumph, amusement, sensual pleasure and relief) and 4 negative states (anger, disgust, fear and sadness)	-
Livingstone & Russo (2018)	The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)	English	24 North American English-speaking professional actors (12 women and 12 men)	English sentences (total of 7,356 recordings)	Acted speech and song	Speech: calm, happy, sad, angry, fearful, surprise and disgust Song: calm, happy, sad, angry and fearful Each expression was produced at two levels	Face-and-voice and face-only

						of emotional intensity with an additional neutral expression.	
Parsons, Young, Craske, Stein, & Kringelbach (2014)	Oxford Vocal Sounds database (OxVoc)	Non-verbal sounds	Infant vocalizations (4 girls and 5 boys) Adult vocalizations (19 clips by women only for distress vocalizations, 15 women and 15 men for laughter vocalizations and 15 women and 15 men for neutral vocalizations) Animal vocalizations (pet cats and dogs)	Total of 173 stimuli Infants: cry vocalizations (*n* = 21); laughter vocalizations (*n* = 18); neutral babbles (*n* = 25) Adults: distress vocalizations (*n* = 19); laughter (*n* = 30); neutral (*n* = 30) Animals: distress (*n* = 30)	Infants: sounds from video recordings of infants filmed in their own homes Adults and animals: sounds found from online resources	Happy (laughter vocalizations), sad (cry and distress vocalizations) and neutral	-
Rigoulot, Wassiliwizky,& Pell (2013)	Database of emotionally-inflected pseudo-utterances	Pseudo-utterances by native speakers of Canadian English	4 speakers (2 women and 2 men)	120 pseudo-utterances (7 syllables in length)	Acted speech	Anger, disgust, fear, happiness, sadness, and neutral	-

Wendt & Scheich (2002); Wendt et al., (2003)	Magdeburger Prosodie-Korpus	German	2 actors (woman and man)	Linguistically meaningful words (n > 3,000) and disyllabic pseudo-words (n = 200)	Acted speech	Anger, disgust, fear, happiness, sadness and neutral	-

Table 2. List of the 73 pseudo-words included in the corpus in roman and in IPA. Boldface in the orthographic representation indicates the syllable carrying the main stress.

Item Number	Orthographic representation	IPA transcription	Item Number	Orthographic representation	IPA transcription
1	**ner**vack	/'nɜɹvæk/	38	**vig**ging	/'vɪgɪŋ/
2	**lo**rack	/'loɹæk/	39	**vo**ker	/'voʊkəɹ/
3	**lai**ret	/'lɛɹət/	40	**vo**kered	/'voʊkəɹd/
4	**vo**kered	/'voʊkəɹd/	41	**vo**lers	/'voʊləɹs/
5	**tai**rack	/'tɛɹək/	42	**win**nith	/'wɪnɪθ/
6	**dom**ner	/'dɑmnəɹ/	43	**zi**ddy	/'zɪdi/
7	**na**mmy	/'næmi/	44	**zi**lard	/'zɪləɹd/
8	**ta**nnock	/'tænək/	45	**ver**coed	/'vɜɹkoʊd/
9	**a**gerth	/'ægəɹθ/	46	**for**ny	/'foɹni/
10	**ar**midge	/'ɑɹmɪdʒ/	47	**ad**mage	/'ædmɪdʒ/
11	**bu**rish	/'bʊɹɪʃ/	48	**aff**ning	/'ɑfnɪŋ/
12	**der**nom	/'dɜɹnəm/	49	**el**by	/'ɛlbi/
13	**re**vo	/'ɹɛvoʊ/	50	**er**vy	/'ɜɹvi/
14	**fin**gill	/'fɪŋgəl/	51	**in**fess	/'ɪnfɛs/
15	**jou**less	/'dʒoʊlɛs/	52	**you**ssle	/'jusəl/
16	**le**bby	/'lɛbi/	53	**ker**vo	/'kɜɹvoʊ/
17	**low**men	/'loʊmən/	54	**ker**voed	/'kɜɹvoʊd/
18	**ma**dage	/'mædədʒ/	55	**lar**py	/'lɑɹpi/
19	**me**nno	/'mɛnoʊ/	56	**lek**nodge	/'lɛknədʒ/
20	**me**rrus	/'mɛɹəs/	57	**mod**ner	/'mɔdnəɹ/
21	**mo**wan	/'moʊwən/	58	**mo**kers	/'moʊkəɹs/
22	**na**bick	/'næbɪk/	59	**mus**ser	/'mʌsəɹ/
23	**ne**mmy	/'nɛmi/	60	**na**ffing	/'næfɪŋ/
24	**ni**dder	/'nɪdəɹ/	61	**ni**fish	/'nɪfɪʃ/
25	**ni**llen	/'nɪlən/	62	**ni**pher	/'nɪfəɹ/
26	**no**mel	/'nɔməl/	63	**oth**ening	/'ɔθ(ə)nɪŋ/
27	**no**mey	/'noʊmi/	64	**ra**ckies	/'ræki:z/
28	**ra**midge	/'ɹæmɪdʒ/	65	**sco**pies	/'skoʊpi:z/
29	**sha**vil	/'ʃævɪl/	66	**shi**fin	/'ʃɪfɪn/
30	**shi**bur	/'ʃɪbəɹ/	67	**vack**ner	/'væknəɹ/
31	**slo**ver	/'sloʊvəɹ/	68	**va**shil	/'væʃɪl/
32	**te**rrel	/'tɛɹəl/	69	**vi**shal	/'vɪʃəl/
33	**tha**ger	/'θægəɹ/	70	**wed**ick	/'wɛdɪk/
34	**tho**mer	/'θoʊməɹ/	71	**win**thy	/'wɪnθi/
35	**val**ish	/'vælɪʃ/	72	**you**shing	/'ju:ʃɪŋ/
36	**ven**ner	/'vɛnəɹ/	73	**zu**ber	/'zubəɹ/

37	**ver**ney	/'vɜɹni/

Table 3. Example of the materials used to elicit the pseudo-words for each emotion

/'sloʊvəɹ/	It starts like ‘slow’	I say	**slo**ver	I say	**slo**ver	again
/'loɹæk/	It starts like ‘lord’	I say	**lor**ack	I say	**lor**ack	again

Table 4. Classification counts of vocal emotion portrayals by the participants' responses and the overall proportion of accurate responses (%) within each emotion, across both speakers.

		Responses of participants						
Emotion portrayed		Anger	Disgust	Fear	Happiness	Neutral	Sadness	Total
Anger	Count	**2662**	1026	563	876	1413	515	7055
	% within emotion	37.7	14.5	8.0	12.4	20.0	7.3	100.0
Disgust	Count	1209	**3021**	247	466	1244	821	7008
	% within emotion	17.3	43.1	3.5	6.6	17.8	11.7	100.0
Fear	Count	479	168	**3563**	762	977	1129	7078
	% within emotion	6.8	2.4	50.3	10.8	13.8	16.0	100.0
Happiness	Count	852	591	509	**2159**	2003	894	7008
	% within emotion	12.2	8.4	7.3	30.8	28.6	12.8	100.0
Neutral	Count	862	719	465	409	**3664**	1030	7149
	% within emotion	12.1	10.1	6.5	5.7	51.3	14.4	100.0
Sadness	Count	98	188	927	217	1428	**4150**	7008
	% within emotion	1.4	2.7	13.2	3.1	20.4	59.2	100.0

Note.
Modal response is indicated in boldface/highlighted (n = 42,306 data points).

Table 5a, confusion matrix for AG stimuli, with Chi-square Goodness-of-Fit test per emotion

	response							
emotion (speaker AG)	A	D	F	H	N	S	Total	Chi-square Goodness-of-Fit test
anger	**1564**	559	268	267	624	295	3577	(χ(5) = 2089; p < .001
disgust	496	**1168**	227	266	724	696	3577	(χ(5) = 1021; p < .001
fear	362	62	**2092**	599	279	183	3577	(χ(5) = 4779; p < .001
happiness	217	195	329	**1313**	896	627	3577	(χ(5) = 1645; p < .001
neutral	539	294	398	262	**1550**	534	3577	(χ(5) = 1944; p < .001
sadness	42	99	238	55	550	**2593**	3577	(χ(5) = 8328; p < .001

Table 5b, confusion matrix for KM stimuli, with Chi-square Goodness-of-Fit test per emotion

	response							
emotion (speaker KM)	A	D	F	H	N	S	Total	Chi-square Goodness-of-Fit test
anger	**1098**	467	295	609	789	220	3478	($\chi(5) = 925$; $p < .001$
disgust	713	**1853**	20	200	520	125	3431	($\chi(5) = 4033$; $p < .001$
fear	117	106	**1471**	163	698	946	3501	($\chi(5) = 2664$; $p < .001$
happiness	635	396	180	<u>*846*</u>	**1107**	267	3431	($\chi(5) = 1124$; $p < .001$
neutral	323	425	67	147	**2114**	496	3572	($\chi(5) = 4870$; $p < .001$
sadness	56	89	689	162	878	**1557**	3431	($\chi(5) = 3052$; $p < .001$

Note.

The underlined number indicates that for these stimuli, happiness was not chosen as the modal response for intended Happy stimuli; Neutral was the most frequently chosen response.

Figures

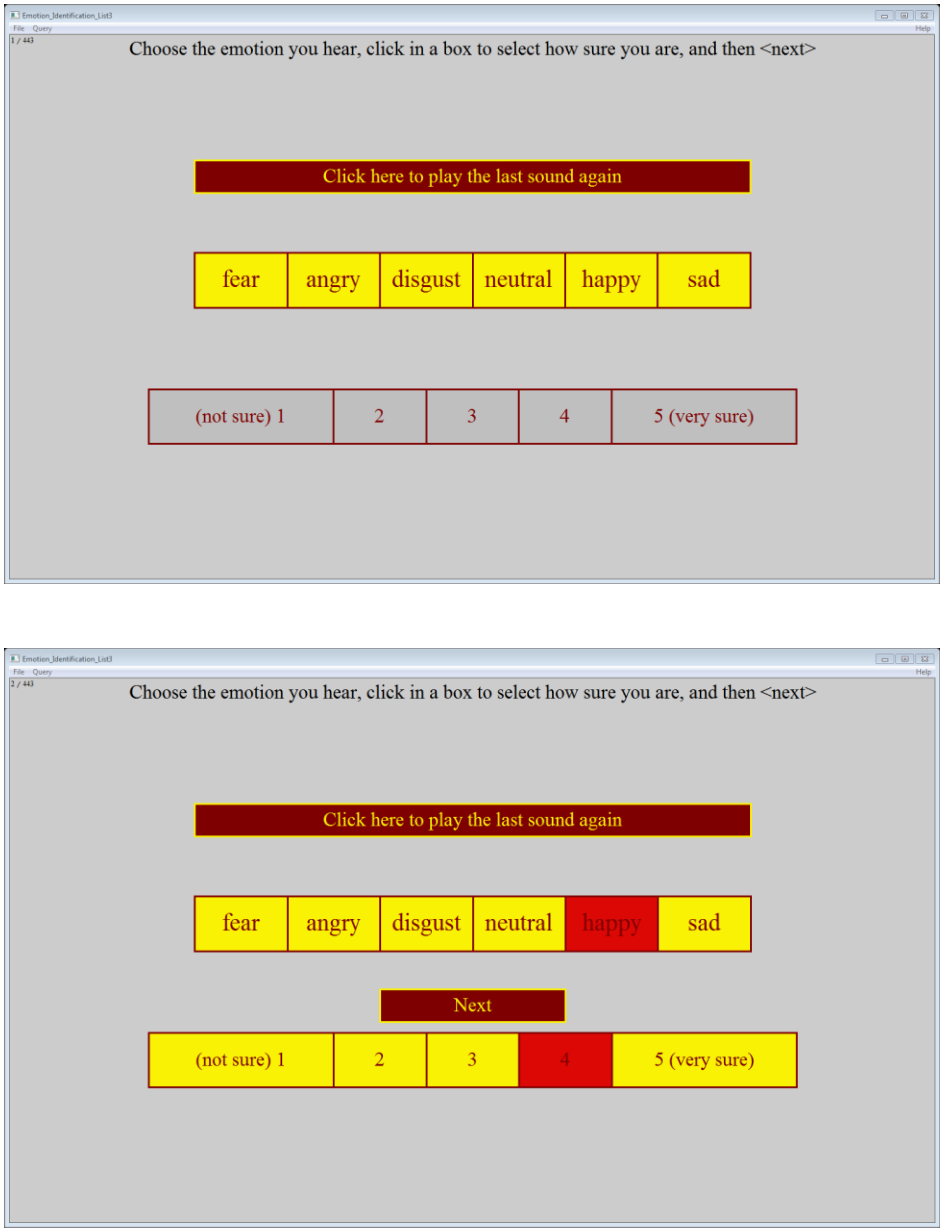

Figure 1. Screenshots of the Praat script interface for the recognition task. The top panel shows the first screen in a trial, where the emotion labels are highlighted (clickable). The bottom panel shows the second screen in a trial, with the confidence scale now also highlighted. The respondent's choices appear highlighted in red (dark grey) and a next button is displayed for participants to move to the next trial. The task was self-paced. Up to 8 replays were allowed.

Figure 2. Overall median accuracy in emotion identification by the 96 participants, separated by speaker. Random performance level (~16%) is indicated by the dotted line.

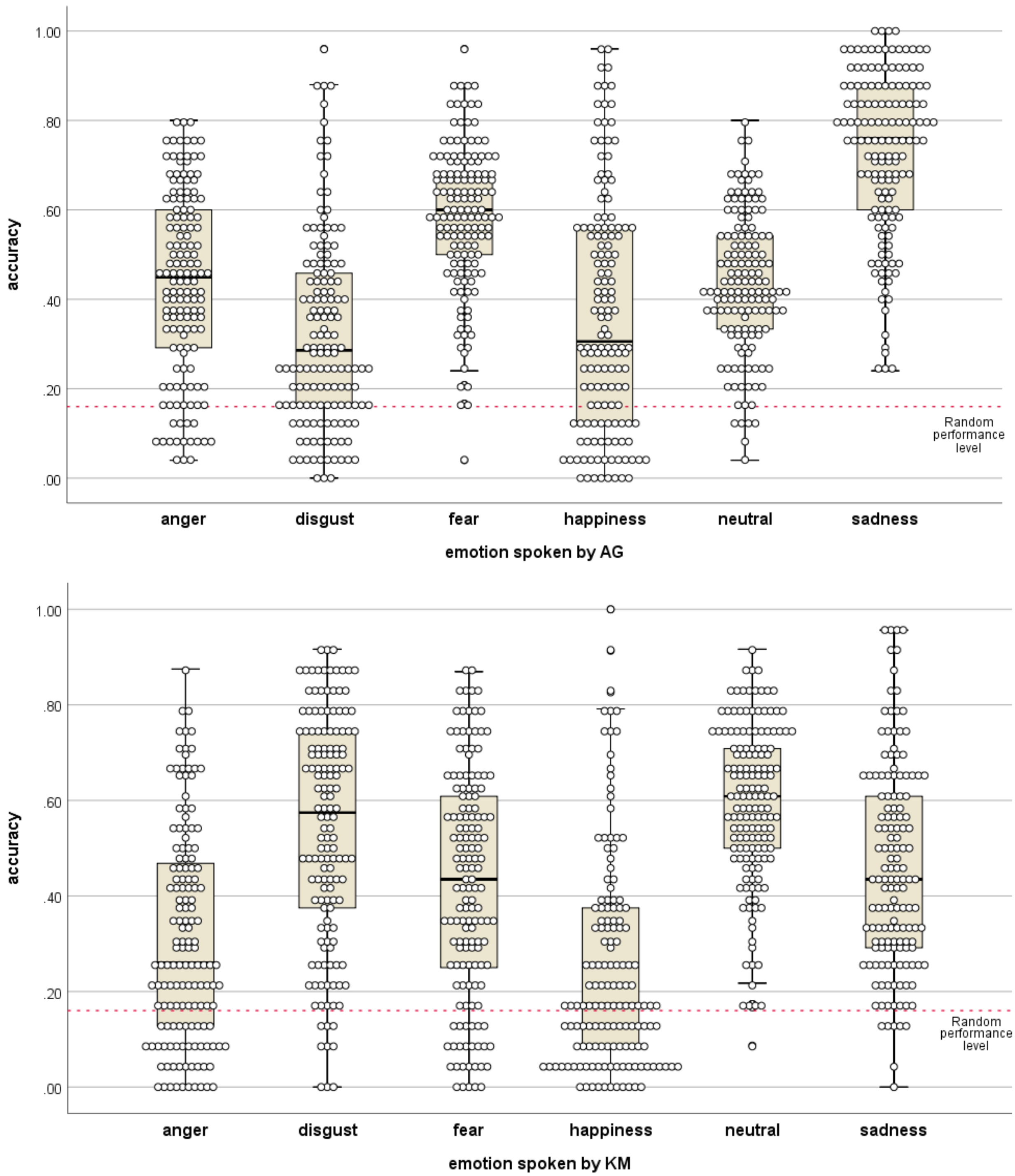

Figure 3a-b. Box plot with overlaid dot plot for each emotion's identification accuracy. Each dot represents one stimulus (i.e., one sound file in the corpus). Horizontal lines represent the medians, boxes show the Interquartile range (IQR) representing 50% of the cases, whisker bars extend to 1.5 times the IQR. Fig. 3a plots AG stimuli, Fig. 3b plots KM stimuli.

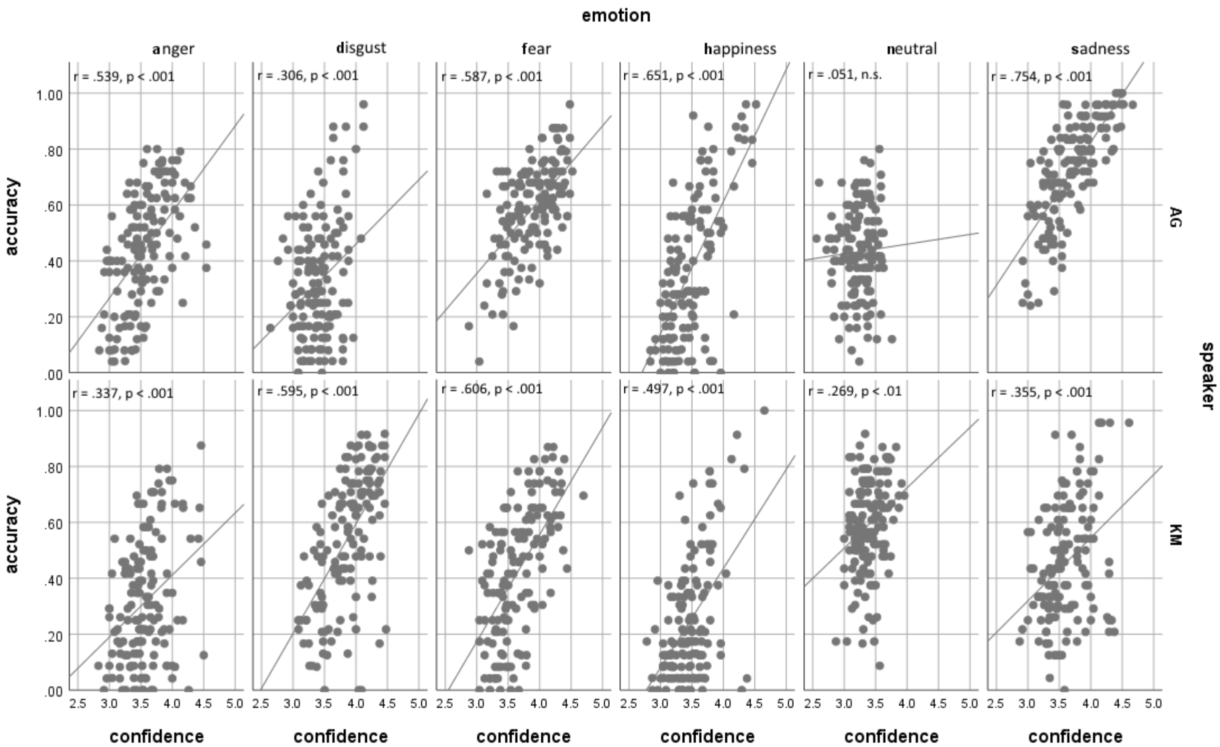

Figure 4. Correlations between the mean identification accuracy of each stimulus and the mean confidence ratings by speaker (top panel: AG; bottom panel: KM).

Appendix

Structure of the corpus (see http://www.iub.edu/~psyling/hoosiervocalemotions.htm). Row 1 and row 2 refer to the corresponding rows in the Excel file (see website link). Each line is a column header in the Excel file or in the comma-delimited spreadsheet (csv). Explanation provides a brief outline of the column content.

<table>
<tr><td>Row 1</td><td>Row 2</td><td>Explanation</td></tr>
<tr><td></td><td>ipa</td><td>International Phonetic Alphabet transcription</td></tr>
<tr><td></td><td>spelling</td><td>Item in English roman alphabet</td></tr>
<tr><td></td><td>item</td><td>Item number</td></tr>
<tr><td></td><td>token</td><td>Token number</td></tr>
<tr><td></td><td>file_name</td><td>Audio file name with extension</td></tr>
<tr><td></td><td>duration_ms</td><td>File duration in milliseconds</td></tr>
<tr><td></td><td>intensity_average_dB</td><td>Average intensity in dB</td></tr>
<tr><td></td><td>intensity_min</td><td>Minimum Intensity</td></tr>
<tr><td></td><td>intensity_max</td><td>Maximum intensity</td></tr>
<tr><td></td><td>voice</td><td>Speaker</td></tr>
<tr><td></td><td>list</td><td>List number</td></tr>
<tr><td></td><td>n_listeners</td><td>Number of listeners who rated this list</td></tr>
<tr><td></td><td>emotion</td><td>Emotion</td></tr>
<tr><td></td><td>accuracy_mean</td><td>Mean accuracy over all trials</td></tr>
<tr><td></td><td>confidence_mean</td><td>Mean confidence score over all trials</td></tr>
<tr><td rowspan="6">confusion_matrix_cnt</td><td>A</td><td rowspan="6">Confusion matrix: raw count of trials where the emotion was chosen, over all trials</td></tr>
<tr><td>D</td></tr>
<tr><td>F</td></tr>
<tr><td>H</td></tr>
<tr><td>N</td></tr>
<tr><td>S</td></tr>
<tr><td rowspan="6">confusion_matrix_prct</td><td>A</td><td rowspan="6">Confusion matrix: % of trials where the emotion was chosen, over all trials</td></tr>
<tr><td>D</td></tr>
<tr><td>F</td></tr>
<tr><td>H</td></tr>
<tr><td>N</td></tr>
<tr><td>S</td></tr>
<tr><td></td><td>accuracy_mean_validrt</td><td>Mean accuracy over selected trials only (RT outliers removed)</td></tr>
</table>

	confidence_mean_validrt	Mean confidence score over selected trials only (RT outliers removed)
confusion_matrix_cnt_validrt	A	Confusion matrix: raw count of trials where the emotion was chosen, over selected trials only
	D	
	F	
	H	
	N	
	S	
confusion_matrix_prct_validrt	A	Confusion matrix: % of trials where the emotion was chosen, over selected trials only
	D	
	F	
	H	
	N	
	S	
	rt_mean	Mean RT over all trials
	rt_median	Median RT over all trials
	rt_mean_validrt	Mean RT over selected trials (RT outliers removed)
	rt_median_validrt	Median RT over selected trials (RT outliers removed)

Note:
a: Anger
d: Disgust
f: Fear
h: Happiness
n: Neutral
s: Sadness