Journal of Personality and Social Psychology
2016, Vol. 111, No. 5, 686–705

0022-3514/16/$12.00 http://dx.doi.org/10.1037/pspi0000066

The Expression and Recognition of Emotions in the Voice Across Five Nations: A Lens Model Analysis Based on Acoustic Features

Petri Laukka
Stockholm University

Hillary Anger Elfenbein
Washington University in St. Louis

Nutankumar S. Thingujam
Sikkim University

Thomas Rockstuhl
Nanyang Technological University

Frederick K. Iraki
United States International University

Wanda Chui
University of California, Berkeley

Jean Althoff
University of Queensland

This study extends previous work on emotion communication across cultures with a large-scale investigation of the physical expression cues in vocal tone. In doing so, it provides the first direct test of a key proposition of dialect theory, namely that greater accuracy of detecting emotions from one's own cultural group—known as in-group advantage—results from a match between culturally specific schemas in emotional expression style and culturally specific schemas in emotion recognition. Study 1 used stimuli from 100 professional actors from five English-speaking nations vocally conveying 11 emotional states (anger, contempt, fear, happiness, interest, lust, neutral, pride, relief, sadness, and shame) using standard-content sentences. Detailed acoustic analyses showed many similarities across groups, and yet also systematic group differences. This provides evidence for cultural accents in expressive style at the level of acoustic cues. In Study 2, listeners evaluated these expressions in a 5 × 5 design balanced across groups. Cross-cultural accuracy was greater than expected by chance. However, there was also in-group advantage, which varied across emotions. A lens model analysis of fundamental acoustic properties examined patterns in emotional expression and perception within and across groups. Acoustic cues were used relatively similarly across groups both to produce and judge emotions, and yet there were also subtle cultural differences. Speakers appear to have a culturally nuanced schema for enacting vocal tones via acoustic cues, and perceivers have a culturally nuanced schema in judging them. Consistent with dialect theory's prediction, in-group judgments showed a greater match between these schemas used for emotional expression and perception.

Keywords: culture, dialect theory, emotion, in-group advantage, speech

Supplemental materials: http://dx.doi.org/10.1037/pspi0000066.supp

A long research tradition in psychology has pondered to what extent emotional expressions are universal versus culturally specific (e.g., Fiske, Kitayama, Markus, & Nisbett, 1998; Jack, Caldara, & Schyns, 2012; Mesquita & Frijda, 1992; Russell, 1994; Sauter, Eisner, Ekman, & Scott, 2010; Scherer, 1997). The current article attempts to expand this tradition by providing the first direct test of an interactionist theory that has been proposed to integrate existing empirical evidence, namely *dialect theory* (Elfenbein, 2013; Elfenbein & Ambady, 2002). In doing so, it presents a large-scale investigation that uses a *lens model* approach

This article was published Online First August 18, 2016.

Petri Laukka, Department of Psychology, Stockholm University; Hillary Anger Elfenbein, Olin Business School, Washington University in St. Louis; Nutankumar S. Thingujam, Department of Psychology, Sikkim University; Thomas Rockstuhl, Nanyang Business School, Nanyang Technological University, Singapore; Frederick K. Iraki, School of Humanities and Social Sciences, United States International University, Kenya; Wanda Chui, Haas School of Business, University of California, Berkeley; Jean Althoff, UQ Business School, University of Queensland, Australia.

Nutankumar S. Thingujam, Thomas Rockstuhl, Frederick K. Iraki, Wanda Chui, and Jean Althoff contributed equally, and appear in reverse-alphabetical order. We acknowledge Swedish Research Council 2006-1360 to Petri Laukka and U.S. National Science Foundation BCS-0617624 to Hillary Anger Elfenbein.

Correspondence concerning this article should be addressed to Petri Laukka, Department of Psychology, Stockholm University, 106 91 Stockholm, Sweden. E-mail: petri.laukka@psychology.su.se

(Brunswik, 1956) to examine evidence for universal and culturally specific schemas in both the expression and perception of emotion. The current research also expands on past work by presenting cross-cultural data on the communication of emotion via vocal tones, as a complement to the vast majority of existing studies, which focus on static photographs of facial expressions (Elfenbein & Ambady, 2002). By contrast, other channels of communication, such as vocal tones, have received relatively less attention (Juslin & Laukka, 2003).

Dialect Theory and In-Group Advantage

The current investigation is grounded in the framework of *dialect theory* (Elfenbein, 2013), which was originally developed to help explain the phenomenon of *in-group advantage*—whereby individuals more accurately judge emotional expressions from their own cultural group compared with expressions from foreign groups. In 1964, Tomkins and McCarter wrote that cultural differences in emotional expression are like "dialects" of the "more universal grammar of emotion" (p. 127). Just as linguistic dialects can differ subtly in their accents, grammar, and vocabulary—such as American versus British English—we argue in this article for the existence of paralinguistic dialects. Although dialect theory drew primarily from work that had been conducted using facial expressions (Elfenbein, 2013), its propositions are also likely to hold for emotion expressed through the voice, which typically shows in-group advantage as well (Juslin & Laukka, 2003). Like facial dialects, paralinguistic dialects should involve subtle yet systematic differences across cultures in the style of expressing emotion, in this case via acoustic cues in the voice. In addition, consistent with dialect theory, we argue that individuals tend to judge the vocal expressions of other people based on their own cultural style. Although the dialects of a language are still mutually intelligible, some of the meaning can get lost along the way. As in verbal language, it can be more challenging to understand the vocal tones of someone expressing emotions in a different emotional dialect, due to cultural differences in norms for expressing oneself and understanding others.

In the case of facial expression, previous studies strongly suggest—but do not conclusively prove—that systematic cultural differences in expression style can create the in-group advantage effect. One study examined composite facial expressions based on the left and right hemispheres of a face—that is, taking one photograph and turning it into two pictures, one that showed the left side twice and one that showed the right side twice (Elfenbein, Mandal, Ambady, Harizuka, & Kumar, 2004). In-group advantage was greater when participants judged the left hemisphere, which is more intense and mobile, compared with the right hemisphere, which is more prototypical. Expression style was the only plausible explanation, given that there was a fully within-subject design for both posers and judges. In another study, Elfenbein, Beaupré, Lévesque, and Hess (2007) identified differences in expressive style in terms of specific muscle movements that varied across the groups' posed facial expressions. Further, greater cultural differences in judgment accuracy were found for the emotions that had greater cultural differences in expression style. An additional source of evidence is that, in many studies of facial expression that do not show in-group advantage, the appearance of stimulus materials was deliberately constrained to be identical across cultures (e.g., Beaupré & Hess, 2005; Biehl et al., 1997). Taken together, these results suggest that in-group advantage results from cultural differences in the appearance of emotional expressions.

This article attempts to take a similar approach to the study of vocal tone. It has been argued that facial expressions are the most visible and controllable canvass for emotional expression, and that they can be made without necessarily being aware and without simultaneously communicating via other channels (Ekman & Friesen, 1969). By contrast, vocal expressions are often integrated into deliberate linguistic content. Because of these fundamental differences between communicative channels, we argue that it is worthwhile to examine to what extent past research on dialect theory replicates outside of the particular facial channel in which it has been primarily studied.

Beyond establishing that existing findings have greater generality, the current studies make a unique contribution by providing the first direct evidence for the core proposition of dialect theory—for which the previous work reviewed was strongly suggestive and yet the evidence was indirect. In particular, dialect theory argues that differences in the cultural style of producing emotion expressions map onto differences in the style of judging them, and that together they create the phenomenon of in-group advantage. Directly testing this proposition requires a systematic measurement of the physical cues that produce emotional expressions, a task for which vocal expression is ideal. There are a large number of well-documented acoustic parameters that define human vocal cues, related to the pitch, loudness, timbre, and rate of speech, which can be measured objectively during all moments of vocal production (e.g., Eyben et al., 2016). As such, the vocal channel is ideal for documenting the role of expressive style as a mechanism for in-group advantage. To achieve this, across the current studies we jointly measure both acoustic properties and human emotion judgments, and conduct a *lens model* analysis that integrates the expression and perception of vocal tone across cultures. The next section turns attention to this lens model approach.

The Lens Model Approach to Nonverbal Communication of Emotion

Research on social judgment has often drawn from Brunswik's (1956) classic *lens model* (e.g., Bernieri, Gillis, Davis, & Grahe, 1996; Gifford, 1994). The key insight of the lens model is that we can perceive the world only indirectly. There are cues in the environment that are probabilistically related to properties of the world, and we make use of these cues probabilistically. For example, the height of a building may not be measurable readily from the street, but there are useful yet imperfect cues available. We can observe the size of a shadow on the ground or count the number of stories. Perceivers use these observable cues in an attempt to understand the properties of their world. Some cues are more diagnostic than others and some are more easily detected than others. The accuracy of social perception is a function of both processes—that is, the presence of diagnostic cues as well as their effective use.

Figure 1 illustrates a lens model approach to studying nonverbal communication of emotion through vocal tone. The left side of the model focuses on expression, which is also called encoding or the emission of cues. Speakers experience or wish to display a particular emotional state, and they convey it using a variety of acoustic

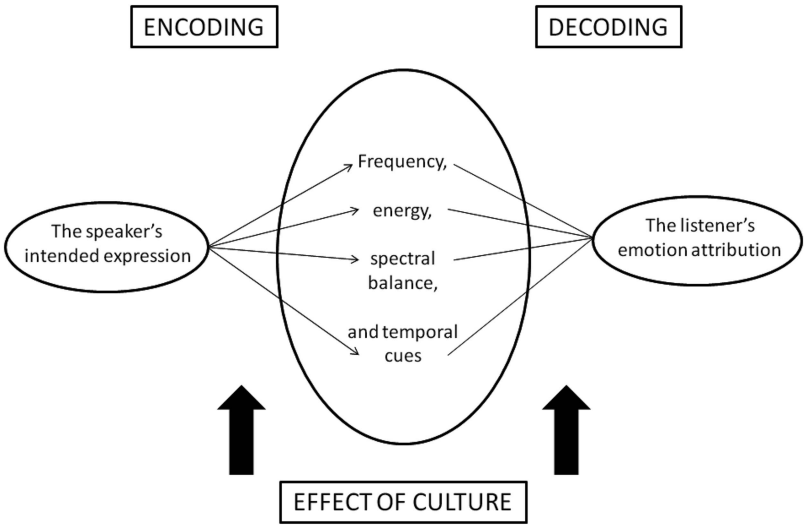

Figure 1. A lens model of cross-cultural communication of emotion via the voice.

properties in the voice. The right side of the model focuses on perception, which is also called decoding or the interpretation of cues. Listeners wish to understand the speaker's emotional state, and they attempt to do so by interpreting the acoustical properties that are available for them to hear. Based on these cues, they integrate both components of the lens model by attempting to infer the speaker's underlying emotion.

As such, conducting a lens model analysis of vocal expression involves several chronological stages. First, a study needs to identify a range of relevant acoustic cues that might indicate emotion, and then examine to what extent those cues are present versus absent in vocal tones expressing emotion. This is what Brunswik (1956) called *cue validity*. Second, a study needs to examine to what extent the acoustic cues are associated with perceivers' emotion judgments, which is what Brunswik (1956) called *cue utilization*. Third, the overall accuracy of perceivers' judgments is measured, in terms of the correspondence between their judgments and the intended emotional state. Cue validity and cue utilization can be described as the schemas that people use to express and judge properties of their social worlds, respectively. It is important to emphasize, as per Brunswik's (1956) original description, that both of these mirror-image processes are probabilistic. Individuals vary in their expression styles and perception styles, and there is room for flexibility and error throughout. From a lens model perspective, the accuracy of communication depends on the degree of correspondence between the speaker's style of expression and the style expected by the perceiver.

As argued by Scherer, Clark-Polner, and Mortillaro (2011), expanding the use of the lens-model approach would be particularly valuable for cross-cultural research because cultural effects may exist in (a) the expression of emotion, which results in culturally specific patterns of acoustic cues; (b) the perception of emotion, which results from cultural differences in the way individuals attend to, perceive, and interpret various acoustic cues; or (c) both. The lens model has been an influential framework for the study of vocal expression (Scherer, 2003). However, relevant work has been conducted exclusively in within-cultural settings (e.g., Bänziger, Hosoya, & Scherer, 2015), with the only cross-cultural study to date being on the acoustic cues within music (Laukka, Eerola, Thingujam, Yamasaki, & Beller, 2013). The current studies attempt to fill this gap, by extending research using the lens model across cultures to study human emotional states expressed through the voice. We use the lens model approach to generate evidence for the key proposition of dialect theory, namely that cultural differences in the production and perception of specific expression cues combine to create the phenomenon of in-group advantage.

Existing Cross-Cultural Studies on Vocal Expression and the Current Research

In this section we first review existing cross-cultural studies on vocal expression, and then describe how the current research attempts to extend the previous body of work. We start with a discussion of expression (also called encoding) studies and then proceed with perception (also called decoding) studies.

Expression of Emotion

Research has investigated the acoustic properties associated with various emotions (e.g., Banse & Scherer, 1996; Hammerschmidt & Jürgens, 2007; Juslin & Laukka, 2001; Sauter, Eisner, Calder, & Scott, 2010). These studies have been conducted with several different languages, and findings across the studies suggest that there is a basic consistency to the acoustic properties associated with basic emotions such as anger, fear, happiness, and sadness (Juslin & Laukka, 2003; Scherer, 2003). However, relatively little work has examined vocal tones from more than one culture within the same study (e.g., Braun & Heilmann, 2012; Fonagy & Magdics, 1963; Pell, Paulmann, Dara, Alasseri, & Kotz, 2009; Ross, Edmondson, & Seibert, 1986). These studies usually do not provide data on systematic tests of cultural differences in acoustic properties, potentially due to the relatively small sample sizes. The number of participants serving as speakers in these

studies tends to range from two to five per culture, or is not reported. A notable exception is Anolli, Wang, Mantovani, and de Toni (2008), who compared vocal expressions portrayed by 30 Chinese and 20 Italian undergraduate students. They identified similarities and differences between how the two groups modulated some acoustical parameters in their vocal expressions, and did not investigate emotion perception with these vocal stimuli.

The closest study to date was based on machine learning, and also did not involve human data on emotion perception. Laukka, Neiberg, and Elfenbein (2014) used a corpus of stimuli from Australia, India, Kenya, Singapore, and the United States (Laukka et al., 2010). They trained acoustic-based classifier programs to recognize emotional expressions. For each model, training was based on stimuli from one cultural group, and recognition was based either on stimuli from the same group on which it was trained or on stimuli from a different group. Accuracy was at levels greater than what would be expected by chance, regardless of whether classifiers were trained and tested on stimuli from the same versus different cultures. This provides evidence for basic universals in the acoustic cues for emotional expression. However, accuracy was higher when classifiers were trained and tested on stimuli from the same versus different cultures. This finding demonstrated systematic acoustic differences across cultures because, logically, no human bias or other influences could have been responsible for better accuracy. The present research builds on this work, using stimuli from the same corpus of vocal tones. We extend this work by conducting acoustical analysis to document which specific paralinguistic cues are responsible for the cultural differences. Further, we collect judgment data from participants versus machine learning models. That is, we know from Laukka et al. (2014) that cultural differences exist and that they can create in-group advantage under certain simulated circumstances. However, we do not yet know what particular differences these may be, or whether they would create in-group advantage with real human listeners.

Perception of Emotion

Table S1 in the online supplemental material contains a comprehensive description of 40 cross-cultural studies of vocal emotion recognition, and we briefly review this literature below. In these cross-cultural studies, vocal expressions are recorded from at least one national or ethnic group, and afterward they are judged by members of their own in-group and at least one national or ethnic group outside of their origin. In the typical design, participants listen to vocal stimuli, and after each stimulus provide a forced-choice judgment among the emotion categories being tested. Within this basic design, there have been three main variations. First, in the most common variant, the *many-on-one* design, individuals from multiple groups judge stimuli originating from a single group (e.g., Altrov & Pajupuu, 2015; Beier & Zautra, 1972; Graham, Hamblin, & Feldstein, 2001; Scherer, Banse, & Wallbott, 2001; Van Bezooijen, Otto, & Heenan, 1983; Waaramaa, 2015). Second, in the *one-on-many* design, individuals from a single group judge stimuli originating from multiple groups (e.g., Kramer, 1964; Pell, Monetta, Paulmann, & Kotz, 2009; Thompson & Balkwill, 2006). In the third main variation, the *balanced design*, stimuli from multiple groups are judged by members of each of these groups. Most balanced studies examine emotion recognition accuracy with two cultural groups (e.g., Albas, McCluskey, & Albas, 1976; Paulmann & Uskul, 2014; Sauter et al., 2010), and a few studies have examined three cultural groups (Braun & Heilmann, 2012; Davitz, 1964; Shochi, Rilliard, Aubergé, & Erickson, 2009).

Overall, these studies tend to provide evidence for at least minimal universality (Russell, 1994), in that participants recognized vocal stimuli from foreign groups more accurately than one would expect from chance guessing alone. In this respect, some of the most revealing studies have recently examined physically isolated cultural groups with little exposure to mass media (e.g., Bryant & Barrett, 2008; Sauter et al., 2010; Cordaro, Keltner, Tshering, Wangchuk, & Flynn, 2016). The finding that these groups are accurate in understanding foreign vocal tones suggests the possibility of innate biological influences on the emotion perception process, because physical isolation limits the plausibility of alternative explanations such as cultural learning. We note that—beyond this minimal universality—all but three of the 40 studies listed in Table S1 in the online supplemental material show some degree of in-group advantage (cf., Albas et al., 1976; McCluskey, Albas, Niemi, Cuevas, & Ferrer, 1975; Zhu, 2013).

In evaluating evidence for in-group advantage, it is worth emphasizing data from balanced designs, which provide the greatest amount of information. It is important to note that they are able to control for the possible extraneous influence of main effects across expresser and perceiver cultural groups, and they test cultural differences in the form of an interaction effect. After all, any two samples can vary along dimensions other than cultural group. For example, emotion recognition accuracy is influenced by gender composition, age, education level, and socioeconomic status (e.g., Hall, Andrzejewski, & Yopchick, 2009; Izard, 1971; Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979), among other factors that can vary across national samples. Likewise, in spite of attempts to create equivalent stimuli from multiple groups, the recorded expressions could also vary due to differences in recording conditions and equipment, or the skill levels of the individual expressers. Expression and recognition can also be influenced by comfort with the laboratory setting, which tends to be greater among Western university students because they more frequently serve as research participants. Taken together, the ability to control for these known and potentially unknown factors is worthwhile, and the current investigation therefore uses a large-scale balanced design.

The Current Research: Testing Dialect Theory Using a Lens Model Analysis

Dialect theory proposes that in-group advantage results from a match between culturally specific schemas in emotional expression style and culturally specific schemas in emotion recognition (Elfenbein, 2013). In the current article, we test this proposition directly using three steps. In Study 1, we examine the expression, or encoding, side of the lens model (see Figure 1). We hypothesize that there may be systematic differences across cultures in the style of expressing emotion through the voice, and we test this hypothesis by documenting at a micro level the presence of specific nonverbal cues that are used similarly or differently across five cultures. In an effort to maximize the quality of vocal portrayals in producing diagnostic cues, we employ professional actors in each country. We also sample a larger than usual number of speakers

from each culture, in order to reduce the possibility of specific item effects that are idiosyncratic to particular individuals. In addition, we include a large number of emotion categories, compared with what has been typically used in previous work, including more positive emotions than usual. Finally, we use an extensive number of acoustic cues, with a set of parameters that has been newly proposed as a potential standard for research on human vocal acoustics (Eyben et al., 2016).

We next focus on the decoding side of the lens model. Study 2 examines the effect of culture on emotion recognition accuracy. We test both for cultural universality, in the form of better-than-chance accuracy across groups, as well as cultural specificity in the form of in-group advantage. In doing so, we use a balanced design in which vocal expressions from five cultural groups are judged by individuals from each of these groups. The advantage of using such a large design is that it provides a large amount of data using comparable methods from multiple cultures, which provides more generalizable evidence than that based on a smaller number of groups. It also provides the opportunity to examine how relative cultural distance can influence cross-cultural differences in emotion recognition. For example, there may be lesser in-group advantage when speakers and listeners are from Western cultures versus when one group is from the West and another is from East Asia or Africa. Including a relatively large number of cultural groups is challenging but necessary to make such a comparison.

In the final step to test dialect theory, we connect encoding and decoding together into a lens model analysis. More specifically, we compare the schemas that speakers use to express their emotion via acoustical cues with the schemas that listeners use to judge others' vocal expressions. Based on dialect theory, we hypothesize that the correspondence between these two types of schemas—that is, for expression and for perception—should be greater when speakers and listeners come from the same versus different cultural backgrounds. This paper provides the first reported test of how the expressive styles for producing and judging cues map onto each other across cultures.

Study 1: Effect of Culture on the Expression of Emotion in the Voice

The first part of this investigation involves documenting the acoustic cues that speakers use across cultures in conveying their nonverbal expressions of emotion through the voice. The current study is the first to examine systematically to what extent the schemas for using acoustic cues to express emotion have both commonalities and systematic differences across cultural groups. In terms of the lens model discussed above (see Figure 1), this is the stage of cue validity, and serves as the first step of testing dialect theory.

Method

Stimulus materials. Vocal stimuli were selected from the Vocal Expressions of Nineteen Emotions across Cultures (VENEC) database, which consists of emotion portrayals by 100 professional actors across five cultures (Australia, India, Kenya, Singapore, and the United States; Laukka et al., 2010). To prevent confounds across linguistic backgrounds, each of the included cultures is from a country where English is an official language. The English language was chosen due to its use in a wide range of countries, and the specific countries were selected to sample English speakers from diverse geographic locations. The selected countries also vary greatly with respect to Hofstede's (2001) cultural dimensions. Considering the five dimensions in Hofstede's model (www.geert-hofstede.com; Hofstede, 2001), (a) Australia and the United States are often considered *individualistic* societies, whereas India, Kenya, and Singapore are more *collectivistic*, (b) India, Kenya, and Singapore are considered high in *power distance*, whereas the United States and Australia are low, (c) Singapore is low in *uncertainty avoidance*, with the other four cultures having moderate levels, (d) in terms of *masculinity*, levels from lowest to highest are Singapore, India, Kenya, the United States, and Australia, and (e) *long-term orientation* is lower in Australia, Kenya, and the United States, and higher in India and Singapore.

For the sake of consistency, a "professional actor" was defined in each nation as an individual who had been paid on at least one occasion for their acting. The database contains recordings from 20 actors from each culture (50% women; ages = 18–30 years), each of whom was born and raised in their respective country and spoke English since birth or early childhood. In order to increase cultural homogeneity within the sample, recruitment was limited to members of the Marathi ethnic group in India, Chinese ethnic group in Singapore, and Caucasian individuals in Australia and the United States. Each actor conveyed 18 affective states (affection, anger, amusement, contempt, disgust, distress, fear, guilt, happiness, interest, lust, negative surprise, positive surprise, pride, relief, sadness, serenity, and shame), each with three levels of emotion intensity (below average, moderately high, and very high). For comparison, each actor recorded emotionally neutral expressions. In order to maintain consistency across portrayals, in each case the verbal material consisted of short phrases with emotionally neutral standard content ("Let me tell you something", "That is exactly what happened"). This resulted in 1,100 stimuli per culture (20 actors × 18 intended emotions × 3 levels of intensity, plus 1 neutral expression per actor), for a total of 5,500 stimuli from five cultures. As a whole, the VENEC corpus can be expected to contain a diversity of expressive styles due to speaker, culture, and emotion intensity effects.

The actors were first provided with scenarios based on the appraisal theory of emotion (e.g., Ellsworth & Scherer, 2003; Lazarus, 1991; Ortony, Clore, & Collins, 1988). These scenarios describe typical situations in which each of the above emotions may be elicited, and actors were instructed to try to enact finding themselves in these situations. The scenarios presented to the actors are available in the online supplemental material. The protocol further asked the speakers to try to remember similar situations that they had experienced personally and that had evoked the specified emotions. They were asked, if possible, to try to put themselves into the same emotional state of mind. The classic method of acting out emotional episodes by reactivating past emotional experiences is common among actors (Stanislavski, 1936). These instructions were used to encourage consistency in the methods employed by the individuals providing stimuli, and to limit idiosyncratic or cultural differences in interpreting the meaning of each emotional category label. The actors were also instructed to try to express the emotions as convincingly as possible, but without using overtly stereotypical expressions. Experimenters

did not provide the actors with feedback or recommendations of any kind.

The recordings were conducted on location in each country (Brisbane, Australia; Pune, India; Nairobi, Kenya; Singapore, Singapore; and Los Angeles, United States). Conditions were standardized across locations, in order to maximize consistency in the methods. All recordings took place in localities with dampened acoustics, and the actors' speech was recorded directly onto a computer with 44 kHz sampling frequency using a high-quality microphone (sE Electronics USB2200A, Shanghai, China). To enable a wide dynamic range while avoiding clipping, each actor first produced sample stimuli, and recording levels were optimized and held constant based on the loudest sample.

Selection of vocal stimuli. We sampled 550 expression portrayals from the VENEC corpus. Doing so first involved selecting 10 of the 18 affective states (anger, contempt, fear, happiness, interest, sexual lust, pride, relief, sadness, and shame). In each case, to reduce the number of stimuli and to limit floor effects, we used only the vocal tones expressed in the condition with instructions for a moderately high level of emotion intensity. We also included the emotionally neutral utterances, which provide a baseline to compare emotionally expressive stimuli. Further, in Study 2 with human listeners, the inclusion of neutral voices provides an option that effectively serves as a non-of-the-above option if participants do not believe that any item is appropriate from the forced-choice list of emotional categories.

In reducing the number of emotion categories, we relied on the same items as in Laukka et al. (2014). This selection included well-studied emotional states that are largely agreed to represent basic emotions (anger, fear, happiness, and sadness; e.g., Ekman, 1992). This also included several less well-studied emotions that are considered basic emotions by some but not all theorists: contempt, interest, lust, and relief (see Tracy & Randles, 2011, for a summary of the current debate). For variety and to increase the available data, we also included two self-conscious emotions (Tangney & Tracy, 2012)—pride and shame—which have rarely figured in studies of vocal expression. Notably, and in contrast to most previous empirical work that has focused on negative emotions (see Sauter, 2010), our selection included equal numbers of positive and negative emotions. This allows us to explore which positive emotions have recognizable vocal signals across cultures.

We further reduced the number of stimuli using a pretest to indicate those among the portrayals that were best recognized and, thus, most likely to contain diagnostic acoustical cues. In a small within-cultural pilot test of $N = 121$, all moderate-intensity portrayals from each culture were judged in a forced-choice format by individuals from the same cultural group as each set of stimuli. The pilot test included all 19 affective states that are contained in the larger VENEC corpus, and was not limited to the emotional states included in the current study. This resulted in 380 stimuli per culture (20 actors $\times$ 19 intended expressions = 380 stimuli per culture). Due to the cumbersome nature of including 19 options in a forced-choice design, separate tests were conducted for positive and negative expressions. Listeners were asked to choose one emotion label that best described the expressed emotion for each stimulus. They used an 11-alternative forced-choice task, which included the 9 states from that valence as well as neutral. In addition, the task included a none-of-the-above option which was labeled as "other emotion" in participant instructions. The response options for the negative valence test were anger, contempt, disgust, distress, fear, guilt, neutral, sadness, shame, (negative) surprise, and "other emotion". The options for the positive valence test instead were affection, amusement, happiness, interest, lust, neutral, pride, relief, serenity, (positive) surprise, and "other emotion". Each response option was explained to pilot participants by giving them the same scenarios as presented to the actors (see online supplemental material), to ensure that emotion words were interpreted similarly across individuals and cultures. Participants listened to the stimuli through headphones, and the experiments were run individually using MediaLab software (Jarvis, 2008) to present stimuli and record responses.

The pilot tests were conducted on location in each respective country (Brisbane, Australia; Pune, India; Nairobi, Kenya; Singapore, Singapore; and Berkeley, United States). Pilot participants judged all stimuli in Australia ($N = 8$, 6 women, age $M = 29.1$), India ($N = 24$, 9 women, age $M = 25.2$), and the United States ($N = 25$, 14 women, age $M = 31.0$). Due to time constraints, a small number of participants could only judge one valence or the other in Kenya (total $N = 30$, 15 women, age $M = 22.0$; negative emotions $N = 28$, positive emotions $N = 30$), and Singapore (total $N = 34$, 18 women, age $M = 22.8$; negative emotions $N = 31$, positive emotions $N = 30$). This led to a total of 121 participants involved, who were all born and raised in their respective country and spoke English since birth or early childhood.

The pilot test results were used to select the portrayals with the highest recognition accuracy. This included separately for each of the five cultures: five male and five female stimuli for each of the 11 selected expressions, for a final total of 550 vocal stimuli. The overall proportion of correct recognition for the selected stimuli was 42.5%, which is 4.7 times higher than chance performance (chance level = 9.1%). The rates for the individual emotions were: anger (63.8%), neutral (59.4%), sadness (51.5%), lust (50.3%), relief (45.8%), fear (37.8%), happiness (36.2%), contempt (34.7%), interest (33.0%), pride (29.3%), and shame (26.1%). Note that, for these moderate intensity-level stimuli, selecting the most highly recognized items did not lead to ceiling effects. The number of portrayals that each individual actor contributed to this selection varied.

Acoustic analysis. Vocal stimuli were analyzed with regard to 65 acoustic parameters. For overviews on voice physiology, acoustics, and perception, see Kreiman and Sidtis (2011) and Raphael, Borden, and Harris (2011). First, we first utilized openSMILE software (Eyben, Weninger, Gross, & Schuller, 2013) to extract the parameters included in the Geneva Minimalistic Acoustic Parameter Set (GeMAPS; see Eyben et al., 2016, for a detailed description). GeMAPS provides a standardized state-of-the-art method to measure a baseline set of acoustic cues containing frequency, energy, spectral balance, and temporal descriptors. Frequency related cues include aspects of the fundamental frequency (F0) of the voice—which represents the rate of vocal fold vibration and is strongly correlated to the perception of pitch—and the vocal tract resonance frequencies called formants. Energy cues refer to aspects of the intensity of the voice signal, which is subjectively heard as loudness and reflects the effort required to produce the speech. The third class of cues reflects the spectral balance of the voice signal. These cues are related to perceived voice quality, or the timbre of the voice, and are influenced by a variety of laryngeal and supralaryngeal features. Temporal cues, finally, include infor-

mation about the rate and duration of voiced and unvoiced speech segments. We used a prerelease version of GeMAPS that contained 58 cues.

The GeMAPS features were selected by an international expert group, based on (a) their potential to reflect physiological changes in voice production (e.g., Sundberg, Patel, Björkner, & Scherer, 2011), (b) the frequency and success with which they have been used in previous studies, and (c) their theoretical significance (e.g., Scherer, 1986). Experiments have demonstrated that classification models can be applied to vocal expressions using GeMAPS and achieve good recognition accuracy (Eyben et al., 2016). Indeed, this work shows that these particular cues provide recognition accuracy on par with that of much larger parameter sets.

Second, we added seven cues not in GeMAPS. This included six cues containing dynamic frequency and energy information (i.e., delta regression coefficients and proportion of frames with rising or falling frequency/energy). For the final cue, we added utterance duration (which provides a measure of speech rate for standard content utterances) to the parameter set. This led to a total of 65 acoustic cues. The additional parameters were extracted using Praat software (Boersma & Weenink, 2008), and we refer readers to Laukka, Neiberg, Forsell, Karlsson, and Elenius (2011) for details.

Because 65 cues can provide a potentially unwieldy data set, results for the 550 stimuli were included in a principal components analysis with varimax normalized rotation. This allowed us to reduce the number of acoustic cues to include in the subsequent statistical analyses. Parallel analysis indicated the number of factors to retain, as implemented in the "paran" package in R (Dinno, 2009). This revealed a 14-factor solution. Based on the PCA results, 16 cues were retained based on the highest factor loadings and/or interpretability for each class of cues. Table 1 lists these acoustic parameters. A description of the full set of cues, and their factor loadings, is available in the online supplemental material in Table S2.

It is valuable to control for individual differences in baseline values between actors while enabling the comparison of the direction and magnitude of acoustic variation across conditions (Banse & Scherer, 1996; Juslin & Laukka, 2001; Patel, Scherer, Björkner, & Sundberg, 2011). For example, a person's physical size or gender can influence the size and shape of their throat, vocal folds, and laryngeal and supralaryngeal features. Individual differences of these types are nuisance factors that are worthwhile to control when examining emotional expression cues. For this reason, the raw values for each cue were normalized using a Fisher z transformation across all portrayals produced by each actor.

Results

Analyses were conducted to examine evidence for cultural universals and differences in the acoustical patterns associated with vocal expressions of emotions across five groups. For each of the 16 acoustic cues used in analysis, a separate 5 (expresser culture: Australia, India, Kenya, Singapore, and the United States) × 11 (emotion: anger, contempt, fear, happiness, interest, lust, neutral, pride, relief, sadness, and shame) between-groups analysis of variance (ANOVA) examined speakers' use of the cue. Results of these analyses appear in Table 2.

All cues showed a main effect for emotion, which indicated that actors made use of these cues to distinguish their portrayals differently across emotional states. The trends for this emotion main effect are summarized in Table 2, in terms of which emotional categories tend to have high, medium, or low levels for each cue. More details are available in the supplemental materials Table S3, including mean values and 95% confidence intervals. The acoustic profiles are generally in line with previous research (e.g., Juslin & Laukka, 2003) for the well-studied emotions (anger, fear, happiness, and sadness), and also provide new data on the acoustic patterns associated with less well-studied emotions. Relatively little previous data are available for the vocal expression of contempt, interest, lust, pride, relief, and shame.

Table 1
Summary of Acoustic Cues Measured in Study 1

Feature type	Description	Factor loading
Frequency-related cues		
F0M	Mean fundamental frequency (F0) on a semitone frequency scale	Factor 2: .90
F0PercRange	Range of the 20th to the 80th percentile of F0	Factor 5: .59
F0SlopeRise	Mean slope of signal parts with rising F0	Factor 9: .95
F0SlopeFall	Mean slope of signal parts with falling F0	Factor 7: .92
F0FracRise	Percentage of frames with rising F0	Factor 11: .72
F1FreqM	Mean of first formant (F1) centre frequency	Factor 5: .83
F1FreqSD	Standard deviation of first formant (F1) centre frequency	Factor 12: .80
Energy-related cues		
IntM	Mean voice intensity estimated from an auditory spectrum	Factor 3: .81
HNR	Mean harmonics-to-noise ratio, i.e., the relation of energy in harmonic vs. noise-like components	Factor 2: .91
IntFracRise	Percentage of frames with rising voice intensity	Factor 10: .82
IntFracFall	Percentage of frames with falling voice intensity	Factor 8: −.74
Spectral-balance cues		
F1Amplitude	Relative energy of the spectral envelope in the first formant region	Factor 1: .97
Hammarberg	Hammarberg index, i.e., the ratio of the strongest energy peaks in the 0–2 kHz vs. 2–5 kHz regions	Factor 4: .76
H1-A3	Ratio of energy of the first F0 harmonic vs. the highest harmonic in the third formant range	Factor 4: .63
Temporal cues		
VoicedSegM	Mean length of continuously voiced regions	Factor 6: .69
UnvoicedSegM	Mean length of unvoiced regions (approximating pauses)	Factor 1: −.79

Note. For a more comprehensive description of the acoustic cues, see Eyben et al. (2016) and Laukka et al. (2011). See text for description of the models yielding the factor loadings listed.

Table 2
Analysis of Variance of Acoustic Cues Contained in Vocal Emotion Stimuli Across Five Cultures and 10 Affective States (Study 1)

Acoustic cue	Emotion effect			Emotion × Culture interaction			Trends for the emotion effect		
	$F(10, 495)$	p	η_p^2	$F(40, 495)$	p	η_p^2	High (↑)	Medium (=)	Low (↓)
Frequency-related cues									
F0M	38.79	.001	.44	1.90	.001	.13	Fe, Ha, In, An	Sa	Pr, Re, Co, Sh, Ne, Lu
F0PercRange	5.60	.001	.10	1.77	.003	.12	In, Lu, Co	Pr, Ha, An, Sa, Re, Sh	Fe, Ne
F0SlopeRise	3.81	.001	.07	1.65	.008	.12	Fe, Sa	Ne, Sh, Pr, Lu, Co, Re, An	Ha, In
F0SlopeFall	1.90	.043	.04	1.05	.394	.08	Lu	Sa, In, Sh, Co, Pr, An, Fe, Ha, Ne	Re
F0FracRise[a]	10.75	.001	.18	1.90	.001	.13	In, Ha, Pr	An, Fe, Lu, Sa	Co, Ne, Re, Sh
F1FreqM	5.74	.001	.10	1.74	.004	.12	Fe, An	Ha, In, Lu, Re, Sa, Co, Sh, Pr	Ne
F1FreqSD	3.79	.001	.07	.93	.599	.07	An, Ha, Sa	Fe, Pr, Co, Re, In, Lu, Sh	Ne
Energy-related cues									
IntM	60.60	.001	.55	3.13	.001	.20	An, Ha, Fe	In, Co, Pr, Re	Sa, Ne, Sh, Lu
HNR	16.44	.001	.25	1.79	.003	.13	In, Fe, Sa	Ha, Sh, Re	Pr, An, Ne, Co, Lu
IntFracRise[b]	5.29	.001	.10	1.10	.309	.08	An	Fe, Sa, Co, In, Pr, Ha, Ne, Sh	Re, Lu
IntFracFall[b]	9.37	.001	.16	2.10	.001	.15	An, Ha, Fe, In	Sa, Ne, Pr, Co	Sh, Re, Lu
Spectral-balance cues									
F1Amplitude	19.54	.001	.28	2.06	.001	.14	Ha, Ne, An, In	Pr, Co, Fe, Sa	Sh, Re, Lu
Hammarberg	25.09	.001	.34	3.26	.001	.21	Ne, Sa, Sh, In	Lu, Pr	Co, Fe, Ha, Re, An
H1-A3	14.28	.001	.22	2.09	.001	.14	Ne, Sh, Sa	In, Co, Lu, Pr, Re	Ha, Fe, An
Temporal cues									
VoicedSegM	3.66	.001	.07	2.27	.001	.15	Ha	Co, Pr, An, Fe, In, Sa, Lu, Ne, Re	Sh
UnvoicedSegM	17.70	.001	.26	1.61	.012	.12	Lu, Re, Sh	Co, Pr, Sa, An, Fe	Ha, In, Ne

Note. Underlined F values remained significant $p < .05$ after Bonferroni adjustment for multiple testing. High (↑) and low (↓) denotes cues with z values above 0 or below 0, respectively; as indicated by 95% CI not overlapping with 0. Medium (=) denotes cues with z values around 0, as indicated by 95% CI that includes 0. An = anger; Co = contempt; Fe = fear; Ha = happiness; In = interest; Lu = lust; Ne = neutral; Pr = pride; Re = relief; Sa = sadness; Sh = shame. See Table 1 for an explanation of cue abbreviations.
[a] $F(10, 493)$. [b] $F(10, 494)$.

Note that the main effect for the culture of speaker is not included, because it is not meaningful when analyzing z scores that control for individual differences across speakers.

Within these ANOVAs, the effect of primary interest is the Emotion × Expresser Culture interaction. These were statistically significant for 13 of the 16 cues, including at least one cue per type, that is, frequency-related cues, energy-related cues, spectral balance cues, and temporal cues. After using a Bonferroni correction to account for the large number of tests conducted, values remained significant for 9 of the 16 cues, again including at least one cue per type. Given these significant omnibus results, we conducted a series of post hoc analyses to understand how these cues varied across expresser culture. Separately for each emotion and for each acoustic cue, we conducted one-way ANOVAs (5 expresser cultures) predicting the speaker's use of the cue. Table 3 lists acoustic cues that differ across cultures. In the interest of space and exploratory analysis, only statistically significant differences are displayed in Table 3. Descriptive statistics for each cue as a function of emotion and culture are available in the supplementary Table S3. Significant effects of culture on acoustic cues were observed for each emotion, with the largest number of effects emerging for anger and lust. For anger, portrayals from Singapore seem to have been portrayed with stronger vocal effort compared with other cultures, as suggested by, for example, relatively higher values of F0M and IntM, and lower values of Hammarberg and H1-A3. For lust, Indian portrayals stood out with lower values of HNR, F1Amplitude, Hammarberg, H1-A3, and VoicedSegM, which in turn suggests that they were portrayed more forcefully than in other cultures.

Discussion

These data provide evidence for both cultural similarities and cultural differences in the use of acoustic cues to convey emotion through the human voice. Alongside universals in the pattern of cues, there were subtle differences in the acoustic cues in these nonverbal expressions of emotion. We used a corpus of stimuli that were collected with the intention of being as similar as possible in every manner other than the cultural background of speakers. A total of 100 professional actors portrayed 10 different affective states. Based on pilot testing, a total of 550 vocal tones were used from five cultures, which is a larger stimulus set than that used in research typically conducted on vocal tone. Results suggest that,

Table 3
Analysis of Variance of Acoustic Cues Used in Vocal Emotion Stimuli Across Five Cultures (Study 1)

		Culture effect			
Emotion	Acoustic cue	$F(4, 45)$	p	η_p^2	Main trend for culture effect
Anger	F0M	3.88	.009	.26	SIN > KEN, AUS, USA
	F0SlopeRise	3.63	.012	.24	KEN > USA, AUS, SIN
	F0FracRise	4.05	.007	.26	SIN, IND > USA
	IntM	6.44	.001	.36	SIN > USA, AUS, KEN
	IntFracFall	3.14	.023	.22	SIN > USA, IND
	Hammarberg	6.18	.001	.35	IND > AUS, SIN
	H1-A3	3.75	.010	.25	IND, USA, AUS > SIN
Contempt	IntFracFall	2.94	.031	.21	IND > SIN
	UnvoicedSegM	4.03	.007	.26	AUS > IND, KEN
Fear	VoicedSegM	3.82	.009	.25	SIN, USA > IND
Happiness	F0M	3.26	.020	.22	KEN > USA
	IntM	6.80	.001	.38	KEN > AUS, SIN, USA
	Hammarberg	3.69	.011	.25	USA, AUS > KEN
Interest	F0PercRange	6.89	.001	.38	AUS > USA, SIN, IND, KEN
	HNR	3.31	.018	.23	IND > KEN
	Hammarberg	3.19	.022	.22	USA, IND > KEN
Lust	HNR	4.68	.003	.29	SIN, KEN, AUS, USA > IND
	F1Amplitude	5.18	.002	.32	AUS, KEN, SIN > IND
	Hammarberg	9.55	.001	.46	AUS, KEN, USA, SIN > IND
	H1-A3	6.24	.001	.36	KEN, AUS, SIN > IND
	VoicedSegM	4.91	.002	.30	KEN, AUS, SIN > IND
Neutral	F0SlopeRise	4.51	.004	.29	AUS > SIN, IND, USA
	F0FracRise	11.97	.001	.52	IND > KEN, SIN, USA, AUS; KEN > AUS
	IntFracFall	8.65	.001	.43	IND > AUS, USA, KEN, SIN
	VoicedSegM	2.69	.043	.19	na
Pride	F1Amplitude	3.18	.022	.22	USA > KEN
Relief	F1FreqM	3.19	.022	.22	KEN > IND
	IntM	3.22	.021	.22	USA > KEN
	UnvoicedSegM	2.90	.032	.20	USA > AUS
Sadness	F1FreqM	2.99	.029	.21	USA > KEN, SIN
	F1Amplitude	4.33	.005	.28	KEN > USA, SIN
	VoicedSegM	5.62	.001	.33	KEN > IND, AUS, USA, SIN
Shame	F0M	2.83	.035	.20	na
	HNR	2.80	.037	.20	na
	Hammarberg	3.18	.022	.22	KEN > IND

Note. Multiple comparisons for assessing main trends for culture were conducted using Tukey honest significant difference tests ($ps < .05$). See Table 1 for an explanation of cue abbreviations.

consistent with dialect theory, speakers tend to have subtly differing schemas across cultures for the cues they use to convey their emotional states. This new evidence provides detailed information about the specific acoustic cues involved in cross-cultural similarities and differences. Notably, effects of culture were observed for all types of acoustic cues (frequency, intensity, spectral balance and temporal features), which suggests that many aspects of vocal tone can be influenced by culture.

Limitations to this study, as well as implications, are discussed below in the General Discussion.

Study 2: Effect of Culture on the Perception of Emotion in the Voice

Study 1 documented differences in the style of expressing emotion through vocal tones across five distinct cultural groups. In Study 2, we examine whether these expressive differences correspond to cultural differences in emotion recognition accuracy. Dialect theory distinguishes between nonverbal dialects and what are called *nonverbal accents* (Marsh, Elfenbein, & Ambady, 2003). In the linguistic metaphor, typically an accent is noticeable yet unchallenging, whereas a difference in dialect can create difficulty in understanding another person's speech. As such, an accent may be considered a weaker instantiation of a dialect.[1] Taken together, nonverbal dialects are those differences in expressive style that impede accurate emotion recognition, whereas nonverbal accents consist of any differences that do not affect recognition accuracy.

We employ a large scale balanced design where the vocal expressions from Study 1 are judged by individuals from the same five cultural groups in a forced-choice task. Consistent with previous findings of in-group advantage, we expect participants to show higher recognition rates when judging expressions from their own cultural group versus expressions from another cultural group. If so, this would provide evidence that the cultural differences in vocal expression style found in Study 1 are indeed dialects rather than accents. Further, we investigate if relative cultural distance is associated with the accuracy of cross-cultural emotion recognition.

[1] We thank an anonymous reviewer for this point.

Method

Participants. A total of 320 participants (143 women) took part from Australia ($N = 59$; 29 women; age $M = 21.7$), India ($N = 60$; 30 women; age $M = 20.6$), Kenya ($N = 60$; 27 women; age $M = 21.9$), Singapore ($N = 62$; 34 women; age $M = 22.8$), and the United States ($N = 79$; 23 women; age $M = 19.0$). All were born and raised in their respective country. All participants were members of their regional majority ethnic group in Australia (Caucasian), India (Marathi), and Singapore (Chinese). In Kenya, the sample included members of the Kikuyu (33%), Luo (20%), Kalenjin (10%), Kamba (8%), and other (28%) African ethnic groups. In the United States, the participants were Caucasian (91%), African American (6%), and Latino/Latina (3%), with individuals excluded from Asian heritage due to the greater possibility of cultural learning with respect to Asian groups in this study. All experiments were conducted on location in each country (Brisbane, Australia; Pune, India; Nairobi, Kenya; Singapore, Singapore; and St. Louis, United States).

Procedure. Participants judged the same 550 vocal stimuli that were used in Study 1. To reduce fatigue from responding to large numbers of trials, stimuli were split arbitrarily into two sets. Each set contained five items from each of the five cultures (Australia, India, Kenya, Singapore, United States) and 11 expressions (anger, contempt, fear, happiness, interest, lust, neutral, pride, relief, sadness, and shame), which resulted in 275 trials per set. Participants were assigned at random to judge one of these two sets. The number of participants judging each set in each culture was nearly identical, which meant that each stimulus was judged by 28–40 listeners from each culture, in a fully balanced 5 (expresser culture, within-subject) × 5 (perceiver culture, between-subjects) design.

Judgments were collected using a forced-choice response method. Participants were instructed to choose one label that best represented the expression conveyed by each speech stimulus. Their alternatives were the same as the 11 intended expressions above. The response alternatives were explained by giving the participants the same scenarios as presented to the actors (see online supplemental material), to control for possible idiosyncratic or culturally specific nuances in the meaning of emotion labels. Responses were scored as correct if the response matched the intended emotion portrayal. Data were collected individually using MediaLab software (Jarvis, 2008), which presented stimuli one at a time. The presentation order of the vocal tones was randomized, and the participants could listen to each one as many times as needed to make a judgment. The participants listened to stimuli through headphones with constant sound levels. The length of each experimental session was approximately 1 h.

Results

Emotion recognition accuracy. We used Wagner's (1993) "unbiased hit rate" (Hu) as the dependent variable in our analyses of better-than-chance accuracy and in-group advantage. This measure accounts for the possibility that systematic biases in participants' use of response categories could artificially inflate their apparent accuracy. For example, a participant could at an extreme receive an apparent recognition rate of 100% for anger by simply responding to all stimuli as angry, even though this participant had no apparent ability to distinguish anger stimuli from others. Hu is an estimate of "the joint probability both that a stimulus is correctly identified (given that it is presented) and that a response is correctly used (given that it is used)" (Wagner, 1993, p. 16). It is calculated as the hit rate multiplied by one minus the rate of false alarms. Hu ranges from zero to one. A score of one indicates that all stimuli of an emotion have been correctly classified and that the respective emotion has never been misclassified as a different emotion. This correction is similar to signal detection methods except, unlike signal detection terms, it allows separate analyses for each stimulus category.

Better-than-chance accuracy. Table 4 displays emotion recognition accuracy for each combination of expresser culture, perceiver culture, and intended emotion. Using Hu, we can conduct a rigorous test of the extent to which perceiver judgments were accurate at rates greater than chance. As a null hypothesis, it is possible to calculate the Hu that would be expected merely by chance guessing. This is provided by Wagner (1993) as the joint probability that a stimulus and a response of the corresponding category would occur by chance, and can be calculated for each judge by multiplying together the independent probabilities of each event occurring alone. We compared these chance scores with the observed Hu scores using paired t tests. The vast majority of Hu scores were significantly higher than the chance level, $p < .05$, with Bonferroni corrections to reflect the large number of statistical tests. The few exceptions are labeled as nonsignificant in Table 4. Notably, accuracy was significant for all combinations of expresser and perceiver groups and for all emotions except pride and shame. For these two emotions, performance for some combinations was not significantly above chance-level performance following the Bonferroni correction. Given that each participant judged only half of the stimuli, we also report Hu scores separately for each stimulus set in Table S4 in the online supplementary materials while noting that there were no systematic differences.

Analysis of variance. Accuracy scores (Hu) were analyzed using a 5 (perceiver culture: Australia, India, Kenya, Singapore, and United States) × 5 (expresser culture: Australia, India, Kenya, Singapore, and United States) × 11 (emotion: anger, contempt, fear, happiness, interest, lust, neutral, pride, relief, sadness, and shame) mixed measures ANOVA, with perceiver culture between-subjects and expresser culture and emotion within-subject.[2]

In-group advantage. The key effect of interest is the interaction of Expresser Culture × Perceiver Culture. Central to a test of dialect theory is the notion that this term is significant, and in the direction of greater accuracy for matched versus mismatched cultural group membership. This interaction term was significant, $F(16, 1260) = 16.71$, $p < .001$, $\eta_p^2 = .18$, and is illustrated in Figure 2. This omnibus interaction term of a 5 × 5 balanced design contains 16 degrees of freedom in the numerator, as an unfocused test. In particular, we examined evidence for in-group advantage,

[2] In a preliminary analysis, we also included a factor for stimulus set. This preliminary analysis revealed neither a significant main effect of stimulus set, $F(1, 310) = 0.58$, $p = .45$, $\eta_p^2 = .002$—with overall accuracy rates essentially identical across both sets, at Hu of .22 and .21—nor a significant Stimulus Set × Expresser Culture × Perceiver Culture interaction, $F(16, 1240) = 1.26$, $p = .21$, $\eta_p^2 = .016$. For this reason, data were merged across the two sets of stimuli for the analyses. However, for the sake of completeness, we also conducted similar ANOVA analyses separately for each stimulus set. Results showed great consistency across data sets, and are reported in full in the online supplemental material.

Table 4
Emotion Recognition Accuracy (Hu) by Emotion, Expresser Culture, and Perceiver Culture (Study 2)

Expresser and perceiver culture	Intended emotion											
	An	Co	Fe	Ha	In	Lu	Ne	Pr	Re	Sa	Sh	Total
AUS expr												
AUS perc	**.29**	**.12**	**.38**	**.30**	**.19**	**.34**	**.33**	**.07** ns	**.09**	**.32**	**.09**	**.22**
IND perc	.23	.04	.27	.24	.13	.24	.21	.10	.09	.26	.06	.17
KEN perc	.22	.07	.34	.27	.14	.31	.18	.09	.10	.28	.05	.18
SIN perc	.28	.07	.33	.30	.19	.30	.26	.09	.13	.30	.07	.21
USA perc	.33	.12	.31	.30	.18	.34	.31	.11	.14	.30	.11	.23
Overall	.27	.09	.32	.28	.17	.31	.26	.09	.11	.29	.08	.21
IND expr												
AUS perc	.46	.11	.24	.33	.14	.18	.27	.05	.24	.19	.09	.21
IND perc	**.41**	**.10**	**.34**	**.51**	**.27**	**.29**	**.36**	**.08**	**.27**	**.28**	**.18**	**.28**
KEN perc	.34	.06	.23	.38	.18	.14	.21	.06	.28	.19	.11	.20
SIN perc	.40	.07	.32	.39	.25	.11	.30	.09	.26	.24	.15	.24
USA perc	.47	.08	.28	.39	.17	.18	.30	.09	.25	.24	.14	.24
Overall	.42	.08	.28	.40	.20	.18	.29	.08	.26	.23	.14	.23
KEN expr												
AUS perc	.20	.05	.15	.19	.07	.33	.15	.02 ns	.08	.12	.05 ns	.13
IND perc	.18	.05	.21	.29	.07	.23	.17	.05	.10	.20	.06	.15
KEN perc	**.13**	**.05**	**.25**	**.33**	**.08**	**.34**	**.17**	**.03** ns	**.10**	**.14**	**.09**	**.16**
SIN perc	.19	.06	.22	.30	.07	.24	.19	.05	.10	.15	.07	.15
USA perc	.18	.05	.15	.25	.05	.36	.18	.05	.09	.13	.05	.14
Overall	.18	.05	.19	.27	.07	.30	.17	.04	.09	.15	.06	.14
SIN expr												
AUS perc	.49	.09	.24	.16	.12	.27	.18	.04 ns	.13	.24	.10	.19
IND perc	.52	.11	.31	.24	.11	.22	.21	.05	.11	.23	.11	.20
KEN perc	.50	.06	.23	.23	.10	.28	.16	.08	.13	.22	.08	.19
SIN perc	**.62**	**.15**	**.34**	**.37**	**.18**	**.35**	**.27**	**.10**	**.15**	**.28**	**.13**	**.27**
USA perc	.56	.13	.30	.25	.11	.27	.21	.10	.11	.25	.12	.22
Overall	.54	.11	.29	.25	.12	.28	.21	.08	.13	.24	.11	.21
USA expr												
AUS perc	.53	.17	.45	.29	.17	.46	.28	.07	.38	.31	.13	.29
IND perc	.32	.08	.35	.37	.16	.33	.22	.05	.34	.26	.12	.24
KEN perc	.35	.09	.26	.31	.18	.44	.17	.04 ns	.44	.21	.14	.24
SIN perc	.45	.12	.35	.35	.18	.36	.28	.06	.39	.31	.16	.27
USA perc	**.48**	**.16**	**.42**	**.36**	**.24**	**.49**	**.29**	**.11**	**.35**	**.32**	**.21**	**.31**
Overall	.43	.13	.37	.34	.19	.42	.25	.07	.38	.28	.16	.27
Total Hu	.37	.09	.29	.31	.15	.30	.23	.07	.19	.24	.11	.21
In-group advantage	.03	.04	.07	.08	.06	.09	.06	.01	.01	.04	.04	.05

Note. Perceivers from Australia ($N = 59$), India ($N = 60$), Kenya ($N = 60$), Singapore ($N = 62$), and the United States ($N = 79$). The in-group advantage was calculated as the difference between accuracy judging same-culture stimuli and accuracy judging other-culture stimuli, and was averaged across all perceivers. Unless otherwise indicated, Hu scores were significantly higher than chance-level performance (paired *t*-tests, $p < .05$, Bonferroni corrected). Values for in-group accuracy are marked in bold. An = anger; Co = contempt; Fe = fear; Ha = happiness; In = interest; Lu = lust; Ne = neutral; Pr = pride; Re = relief; Sa = sadness; Sh = shame.

calculated in terms of the mean difference between in-group and out-group Hu across conditions, which is summarized in Table 4. Overall Hu was higher in the within-cultural (.25) versus cross-cultural (.20) conditions. A separate *t* test comparing each listener's overall in-group Hu to their out-group Hu indicated that this difference was statistically significant, $t_{319} = 10.61$, $p < .001$, $d = .60$.[3]

Figure 2 shows the average accuracy levels that perceivers from the five cultures achieved when judging stimuli from all five cultures. These values were standardized using the Fisher *z* transformation at the level of each individual participant, in order to control for group and individual differences in overall accuracy. The figure illustrates that relative accuracy for expressions from the perceivers' own culture was generally higher than for expressions from the other cultures. In most cases, there is a nonoverlapping 95% confidence interval between the in-group judgments—which are indicated with asterisks—and any of the four cross-cultural judgments made by the perceivers. One exception was that American and Australian perceivers did not differ with regard to their recognition of expressions from the United States (although perceivers from Australia performed better than perceivers from the United States for Australian expressions, $t_{136} = 2.31$, $p = .023$, $d = .40$). We speculate that this one exception may be related to the relatively lower cultural difference between the

[3] A similar *t* test conducted on the uncorrected hit rates (proportion of occasions that participants labeled stimuli with the intended category) also showed evidence for in-group advantage, with significantly higher accuracy for within-cultural (43.6%) versus cross-cultural (38.1%) judgments, $t_{319} = 10.60$, $p < .001$, $d = .54$.

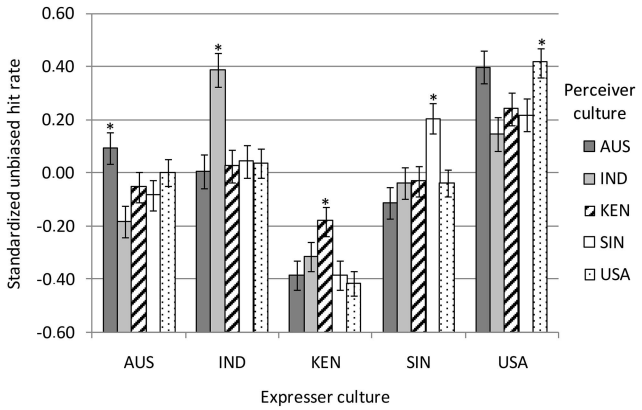

Figure 2. Standardized overall unbiased hit rates (Hu) as a function of expresser and perceiver culture. Error bars represent 95% confidence intervals and asterisks indicate in-group conditions (i.e., match between expresser and perceiver culture). AUS = Australia; IND = India; KEN = Kenya; SIN = Singapore; USA = United States.

United States and Australia than the other country-pairs in this study (Hofstede, 2001).

A small but significant three-way interaction of perceiver culture, expresser culture, and emotion, $F(160, 12600) = 2.25$, $p < .001$, $\eta_p^2 = .03$, indicated that the Perceiver × Expresser interaction also varied across emotions. To examine this interaction further, a series of pairwise *t* tests examined in-group advantage by comparing each listener's in-group versus out-group Hu scores separately for each emotion. These tests showed that the in-group advantage was statistically significant for all emotions except for relief, $t_{319} = 0.92$, $p = .36$, $d = .07$. The values for the other emotions were, in order from smallest to largest effect size: anger, $t_{319} = 2.20$, $p = .029$, $d = .17$; pride, $t_{319} = 2.21$, $p = .028$, $d = .17$; sadness, $t_{319} = 3.69$, $p < .001$, $d = .27$; contempt, $t_{319} = 4.69$, $p < .001$; $d = .34$; shame, $t_{319} = 5.18$, $p < .001$, $d = .39$; neutral, $t_{319} = 6.39$, $p < .001$, $d = .40$; fear, $t_{319} = 5.89$, $p < .001$, $d = .41$; interest, $t_{319} = 5.61$, $p < .001$, $d = .44$; happiness, $t_{319} = 6.80$, $p < .001$, $d = .47$; and lust, $t_{319} = 7.45$, $p < .001$, $d = .47$. In a post hoc analysis that examined individual combinations of Emotion × Expresser Group × Perceiver Group, the largest in-group advantage was observed for Indian expressions of happiness and lust, and for Singaporean expressions of happiness. To further illustrate how in-group advantage varies across conditions, the online supplemental Figure S1 shows the same material that appears in Figure 2, separately for each culture pair for each of the 11 emotions tested.

Other effects. We also observed significant main effects of perceiver culture, expresser culture, and emotion, as well as significant Perceiver Culture × Emotion and Expresser Culture × Emotion interactions. These effects do not qualify the above conclusions about in-group advantage, and are presented in full in the online supplemental material.

Cultural distance analyses. The above results and prior research (e.g., Elfenbein & Ambady, 2003a) suggest that cultures with greater cultural similarity may be more accurate when judging each other's emotions. To further investigate this possibility, we conducted an additional set of analyses which could only be accomplished with a relatively large design such as a balanced set of 5 × 5 cultures. This involved examining the 5 × 5 matrix consisting of residual accuracy after controlling for main effects of expression and perception culture. This matrix of residuals reveals the degree of relative accuracy versus inaccuracy for each country pair. It was compared vis-à-vis an additional 5 × 5 matrix consisting of the cultural distance between each country pair based on Hofstede's (2001) survey research.[4] A test of statistical contrast was used to examine whether the correlation between these two matrices predicted emotion recognition accuracy between country pairs (Rosnow, Rosenthal, & Rubin, 2000). We combined the cultural dimensions from Hofstede's (2001) model together into overall cultural distance by calculating the euclidean distance between country pairs in a 5-dimensional space for the five cultural factors. As predicted, greater overall cultural distance predicted lower recognition accuracy, $F(1, 1260) = 10.26$, $p < .001$, $r_{contrast} = .20$.

In this analysis, anger appeared to be an outlier, which might result from anger showing relatively low in-group advantage. A contrast analysis that attempts to explain the correlates of accuracy in terms of cultural distance is closely related with in-group advantage, in that in-group judgments have zero cultural distance. Along these lines, an additional analysis excluded the three emotions that had the lowest magnitude of in-group advantage, namely relief, pride, and anger. The association became stronger between greater cultural distance and lower relative emotion recognition accuracy, $F(1, 1260) = 22.01$, $p < .001$, $r_{contrast} = .29$. We conducted additional analysis to examine another potential outlier in these data. Australia and the United States have very low cultural distance—less than one-quarter the distance as the next most culturally similar nation-pairs, that is, India and Kenya—and there is also relatively high mutual emotion recognition accuracy between Australia and the United States. To test whether this might influence the results, we removed the United States from analysis.[5] The association increased in effect size, $F(1, 711) = 12.10$, $p = .001$, $r_{contrast} = .25$, and more so when excluding anger, pride, and relief, $F(1, 711) = 22.74$, $p < .001$, $r_{contrast} = .31$. Overall, these analyses show that there can be greater nuances than simply classifying country pairs as in-group versus out-group.

Confusion patterns. Although the use of unbiased hit rates (Hu) controls for many biases in recognition accuracy, the most complete picture of perceiver judgments is provided by a confusion matrix (e.g., Bänziger, Mortillaro, & Scherer, 2012; Elfenbein, Mandal, Ambady, Harizuka, & Kumar, 2002). Table 5 summarizes the results of participant ratings using such a matrix, which plots the intended emotions in the columns and the participant judgments in the rows. In each cell, the value on the left of the slash indicates the average proportion of judgments from

[4] Data are provided in the appendix of Hofstede's (2001) book for Australia, India, Singapore, and the United States. The book presents data for East Africa aggregated across nations, and data for Kenya in particular are provided online (www.geert-hofstede.com). Note that these online materials regarding the cultural dimension of long-term orientation do not include a value for Kenya. So as not to lose the nation from analysis, in this one case the value was used instead from the book's appendix for East African nations.

[5] The United States was removed instead of Australia due to potential asymmetry in that the United States has a highly developed industry that distributes emotional expression stimuli within entertainment as a cultural export product.

Table 5
Confusion Matrix Illustrating the Proportion of Participant Judgments Made Across Intended Emotions as a Function of In-Group vs. Out-Group Stimuli (Study 2)

Judgment	Intended emotion										
(In-group/out-group)	Anger	Contempt	Fear	Happiness	Interest	Lust	Neutral	Pride	Relief	Sadness	Shame
Anger	**56/55**	14/12	1/4	5/7	2/3	1/1	2/2	8/8	3/3	1/1	1/1
Contempt	16/15	**25/19**	2/2	2/4	5/6	5/3	4/6	15/13	5/5	1/1	3/3
Fear	2/2	1/2	**55/46**	2/4	5/5	6/7	2/2	0/1	4/5	13/12	11/9
Happiness	2/3	4/5	3/3	**56/46**	9/10	5/5	1/2	9/9	4/4	2/2	1/1
Interest	5/6	14/16	7/8	12/13	**43/35**	7/8	9/10	17/19	8/6	3/3	6/6
Lust	1/1	4/5	1/2	2/3	3/4	**48/40**	1/1	3/3	5/4	2/3	3/5
Neutral	5/7	16/18	5/6	5/6	19/21	6/7	**64/57**	20/21	12/14	10/9	18/21
Pride	7/8	16/16	2/1	7/7	5/7	4/3	4/4	**18/16**	4/4	1/1	2/2
Relief	3/3	4/4	6/6	5/4	4/5	7/7	4/5	6/6	**35/37**	5/4	8/10
Sadness	2/1	2/2	10/15	3/5	1/2	6/10	5/7	2/3	10/10	**49/49**	18/22
Shame	1/1	1/1	8/7	1/1	2/2	5/7	3/4	1/1	9/8	15/14	**27/21**

Note. Recognition rates for which the expression portrayed is the same as the expression judged are shown in the diagonal cells (marked in bold). Observations/emotion: in-group conditions (N = 3,180), out-group conditions (N = 12,720). All recognition rates and underlined misclassification rates were higher than what would be expected by chance guessing (9.09%), as indicated by z tests ($ps < .05$, Bonferroni corrected).

members of the one in-group culture, and the value on the right indicates the average proportion of judgments from members of the four out-group cultures. Inspection of Table 5 reveals that confusion patterns were very similar in in-group and out-group conditions. Diagonal entries (printed in bold) indicate the proportion of occasions that participants labeled stimuli with the intended category. Underlined font indicates the most frequent confusions, for which a particular intended state was judged as another particular state with frequency higher than that expected by chance, as indicated by z tests ($zs \geq 5.05$, $ps < .05$, Bonferroni corrected for multiple testing). Anger, fear, happiness, interest, lust, neutral, relief and sadness were rarely misclassified as other emotions. Contempt, pride and shame received the lowest recognition rates, and for this reason they had more frequent misclassifications. Contempt was mainly confused with neutral and pride, pride with neutral and interest, and shame with sadness and neutral. Indeed, for both pride and shame, some misclassifications were more frequent than the correct classification. For pride this occurred in both in- and out-group conditions, but for shame it occurred in out-group conditions only. For the sake of completeness, 25 separate confusion matrices appear in the online supplemental Table S5—one matrix for each combination of five expresser and five perceiver cultures.

Discussion

Results showed that emotion recognition accuracy was greater for same-culture versus other-culture judgments, which documented the presence of in-group advantage. This in turn suggests that the cultural differences in expression patterns observed in Study 1 may be considered as dialects rather than accents (see Marsh et al., 2003). There was in-group advantage across positive emotions, negative emotions, and even neutral expressions, with the magnitude varying across emotions.

We further explored differences in accuracy beyond the dichotomous status of in-group versus out-group, and examined to what extent cultural similarity led to relatively greater accuracy for each set of country-pairs in the five groups. This test was significant for an overall measure of cultural distance which combined the five cultural dimensions from Hofstede's (2001) landmark research. We do note a limitation to these analyses, in that they do include in-group judgments and would represent a more precise test if analyses could include only out-group judgments.

Limitations and implications are discussed below in the General Discussion.

Combining Expression and Perception in a Lens Model Analysis

Completing the lens model provides the primary analysis to test dialect theory, which is the primary purpose of this paper. Doing so involves comparing the results of the acoustic analyses from Study 1—which document cross-cultural differences in expressive style—with the results from Study 2—which document in-group advantage in emotion recognition. Together, these two studies provide the data necessary to test directly the dialect theory's proposition that individuals are more accurate when judging emotion expressions conveyed in their own cultural style.

To complete the lens model, we examine the acoustic cues that correspond with the speakers' intended expressions from Study 1. This is called the cue validity stage of the lens model (see Figure 2), and can be described as the schemas speakers use to express vocal tones. We also examine the acoustic cues that correspond with the listener judgments from Study 2. This is called the cue utilization stage of the lens model, and can be described as the schemas listeners use to perceive vocal tones. Consistent with Brunswik's (1956) lens model, judges achieve greater accuracy the greater the match between schemas used for expression (cue validity) and schemas used for perception (cue utilization). Our analyses allow us to examine quantitatively this degree of match versus mismatch for the communication of vocal emotion. Consistent with dialect theory (Elfenbein, 2013), our prediction is that in-group versus out-group judgments will show a greater match between cue validity and cue utilization. If so, this would provide direct evidence for dialect theory.

Results

Patterns of valid and utilized cues. First, using the data from Study 1, separately for each acoustic cue we calculated the point-biserial correlations between the expressers' intended emotions (dichotomously dummy-coded as 1 or 0) and the acoustic cue values for each vocal stimulus. These correlations represent the *cue validity* and provide an index of the degree to which expressers used each cue to convey each of the 11 emotions. Second, separately for each acoustic cue we also calculated the correlations between the perceivers' average emotion judgments collected in Study 2 and the acoustic cue values for each vocal stimulus established in Study 1. Emotion judgments were coded in terms of the continuous proportion of correct responses rather than a dichotomous variable. The resulting correlations represent the *cue utilization*, and provide an index of the degree to which each cue was used by perceivers on a consistent basis to make inferences about each of the conveyed emotions. We used the same selection of 16 acoustic cues as in Study 1, and refer to Table 1 for a description of the parameters.

Table 6 presents cue validity and cue utilization correlations for each acoustic cue, which are displayed as a function of emotion across all expresser and perceiver groups. To give an example of how to read Table 6, expressers tended to portray anger using a loud voice, which is exemplified by a high cue validity correlation for mean voice intensity (IntM) for anger, $r = .48$, $p < .001$. Perceivers' anger judgments were also associated with loud voice intensity, as exemplified by a high cue utilization correlation for IntM for anger, $r = .55$, $p < .001$. Focusing on the cues that were both valid and utilized, Table 6 paints a similar picture as Table 2 in terms of the culturally invariant acoustic profiles for each emotion, as described above in the results section of Study 1.

We also calculated cue validity and cue utilization correlations separately for each pairing of expresser and perceiver culture, see Table S6 in the online supplemental material. Due to the oversized nature of this table (which contains *r*-values for all combinations of 16 acoustic cues × 11 emotions × 5 expresser cultures × 5 perceiver cultures), it appears in the online supplemental material, as Table S6. Inspection of this table reveals that there were few instances where valid cues were limited in their utilization exclusively to in-group conditions (e.g., IntFracFall for Indian expressions of lust and UnvoicedSegM for Kenyan expressions of pride). Most of the valid cues were instead utilized in both in-group and out-group conditions—however, they were used to varying degrees. This indicates that cultural differences in cue utilization are not about "all or nothing" differences, where some cues are utilized in entirely different ways in different cultures. Rather, the cross-cultural differences in expression and perception style are more subtle and yet systematic. Indeed, this makes sense in that it may often be difficult to change only one acoustic cue at a time, because these properties of the voice are intercorrelated due to the physiology associated with producing sounds.

Correspondence between cue validity and cue utilization. Finally, in order to quantify the degree of match between expressers' and perceivers' uses of cues, we calculated the correlations between the cue validity and cue utilization patterns, that is, correspondence scores that represent how well perceivers make use of the true schema for emotional expression. Following Laukka et al. (2013), these correlations were calculated across the 16 selected acoustic cues for each combination of intended emotion, expresser culture, and perceiver culture, using the values reported in online supplemental Table S6. The full set of correspondence scores is shown in Table S7 in the online supplemental material, and a high correlation suggests a good match between cue validity and cue utilization patterns.

For each emotion and pairing of expresser and perceiver cultures, we compared the in-group correspondence score with the average of the correspondence scores across the four out-group conditions (also shown in online supplemental Table S7). A dependent measures *t* test indicated that the correlation between cue validity and cue utilization was significantly higher in in-group conditions (mean Fisher $Z = 1.54$) than in out-group conditions (mean Fisher $Z = 1.41$, $t_{54} = 3.97$, $p < .001$, $d = .22$). This provides evidence for the central proposition of dialect theory, namely that there is a better match between expression and perception styles when expressers and perceivers share the same cultural background.[6] Inspection of online supplemental Table S7 revealed that the overall differences between in-group and out-group correspondence scores were largest for contempt, interest and pride. Note that these results vary slightly from the previous analyses, which examined the effects of individual acoustic cues versus the gestalt across all acoustic cues. The two methods thus capture different aspects for interpreting the influence of culture on vocal expression and perception, and results are consistent across them.

Discussion

Results from the lens model analysis provided an in-depth look at which acoustic cues were correlated significantly with both the expressers' intended emotion (cue validity) as well as with the perceivers' judgments (cue utilization). It is important to note that findings support dialect theory's key proposition that systematic cultural differences in expression style are responsible for systematic cultural differences in recognition accuracy. That is, for in-group judgments, there was a higher correlation between what Brunswik's (1956) theory referred to as cue validity and cue utilization. This indicates a tighter mapping for in-group judgments between the schemas used to express and the schemas used to perceive emotional expressions, and this tighter mapping improves recognition accuracy.

Limitations and implications are discussed below in the General Discussion.

General Discussion

Taking a large-scale approach to examining the expression and recognition of vocal emotion, the current investigation expanded beyond longstanding tradition on cross-cultural research that has tended to focus primarily on still photographs of the face. Emotional expressions were collected from 100 professional actors from five English-speaking nations that span 4 continents. These

[6] To ensure that this effect was not limited to the particular 16 acoustic cues selected to limit the number of analyses, we also calculated the correspondence scores using the entire set of 65 acoustic cues. The in-group mean ($Z = 1.53$) was again significantly higher than the out-group mean ($Z = 1.40$, $t_{54} = 4.53$, $p < .001$, $d = .23$).

Table 6
A Lens-Model Analysis of the Correlations (Pearson r) Between Acoustic Cues and (a) the Expressers' Intended Expression (Cue Validity) and (b) the Perceivers' Emotion Judgments (Cue Utilization), Across Cultural Conditions

	Anger	Contempt	Fear	Happiness	Interest	Lust	Neutral	Pride	Relief	Sadness	Shame
Frequency-related cues											
F0M											
Validity	**.16*****	**−.10***	**.31*****	**.23*****	**.20*****	**−.34*****	**−.26*****	−.03	−.08	.07	**−.15*****
Utilization	**.19*****	**−.09***	**.29*****	**.23*****	**.15****	**−.32*****	**−.30*****	−.05	−.10*	.05	**−.22*****
F0PercRange											
Validity	.02	.07	−.12**	.05	**.17*****	.09*	**−.14****	.05	−.08	−.02	−.09*
Utilization	.01	.07	−.08	.06	**.22*****	.04	**−.13****	.12**	−.05	−.08	−.07
F0SlopeRise											
Validity	−.08	−.01	**.14****	**−.10***	**−.11***	.00	.03	.00	−.04	**.13***	.03
Utilization	−.09*	−.03	**.17*****	**−.12****	**−.14****	−.02	−.04	−.03	−.05	**.22*****	.13**
F0SlopeFall											
Validity	−.01	.03	−.03	−.05	.04	.10*	−.08	.01	**−.10***	.06	.03
Utilization	−.06	−.01	−.00	.01	.05	.06	−.09*	.00	**−.10***	.11**	.08
F0FracRise											
Validity	.04	**−.14****	.05	**.14****	**.21*****	.03	**−.20*****	.09*	**−.14****	.04	**−.12****
Utilization	.06	**−.10***	.07	**.17*****	**.21*****	.02	**−.20*****	.02	**−.10***	−.06	**−.16*****
F1FreqM											
Validity	**.14****	−.07	**.15*****	.05	.04	.02	**−.19*****	**−.09***	.01	.01	−.08
Utilization	**.17*****	−.13**	**.16*****	.06	.00	.05	**−.24*****	**−.14****	−.03	.02	−.08
F1FreqSD											
Validity	**.11****	.00	.04	.09*	−.04	−.05	**−.17*****	.03	.00	.09*	−.10*
Utilization	**.14****	.06	.03	.07	−.06	−.06	**−.21*****	.06	.01	.03	−.03
Energy-related cues											
IntM											
Validity	**.48*****	−.01	**.19*****	**.26*****	.01	**−.29*****	**−.17*****	−.01	−.05	**−.16*****	**−.26*****
Utilization	**.55*****	.11*	**.12****	**.26*****	.01	**−.32*****	**−.24*****	.06	−.08	**−.26*****	**−.40*****
HNR											
Validity	−.08	**−.12****	**.20*****	**.11***	**.20*****	**−.31*****	**−.11***	−.07	.02	**.13****	.02
Utilization	−.09*	**−.13****	**.19*****	**.12****	**.13****	**−.25*****	**−.09***	−.08	.02	**.14****	−.02
IntFracRise											
Validity	**.14****	.05	.07	.03	.04	**−.20*****	.00	.03	**−.15*****	.06	−.05
Utilization	**.21*****	.05	.07	.00	.00	**−.23*****	.02	−.03	**−.18*****	.02	−.05
IntFracFall											
Validity	**.16*****	−.02	.11*	**.14****	.09*	**−.20*****	.01	.00	**−.18*****	.02	**−.12****
Utilization	**.20*****	.01	.04	**.10***	.06	**−.24*****	.01	.01	**−.19*****	.02	**−.16*****
Spectral-balance cues											
F1Amplitude											
Validity	**.11****	.03	−.02	**.23*****	**.11***	**−.30*****	**.19*****	.04	**−.23*****	−.02	**−.14****
Utilization	**.13****	.05	−.08	**.27*****	**.17*****	**−.25*****	**.20*****	.12**	**−.27*****	−.17***	**−.31*****
Hammarberg											
Validity	**−.32*****	−.06	−.08	−.10*	**.12****	.02	**.25*****	−.05	**−.18*****	**.25*****	**.14****
Utilization	**−.35*****	−.16***	−.07	−.08	**.09***	.04	**.32*****	−.10*	**−.18*****	**.24*****	**.18*****
H1-A3											
Validity	**−.24*****	.01	**−.19*****	**−.14****	.02	.00	**.22*****	.00	−.01	**.15*****	**.18*****
Utilization	**−.26*****	−.02	**−.18*****	**−.14****	.00	.03	**.32*****	−.04	.03	**.12****	**.21*****
Temporal cues											
VoicedSegM											
Validity	.05	.08	−.03	**.18*****	−.04	−.05	−.06	.05	−.06	−.04	**−.09***
Utilization	.09*	.10*	−.07	**.24*****	.03	.02	−.09*	.15***	−.07	−.16***	**−.19*****
UnvoicedSegM											
Validity	−.06	.03	−.08	**−.13****	**−.18*****	**.30*****	**−.22*****	.00	**.21*****	−.02	**.14****
Utilization	−.05	.04	−.04	**−.18*****	**−.18*****	**.29*****	**−.24*****	−.02	**.27*****	.08	**.23*****

Note. $N = 550$ except for IntFracRise and IntFracFall ($N = 549$) and F0FracRise ($N = 548$). A significant cue-validity correlation (i.e., the correlation between a cue and the expressers' intended emotion) suggests that the cue is used in a consistent fashion by the expressers to convey a certain emotion. A significant cue-utilization correlation (i.e., the correlation between a cue and the perceivers' mean recognition accuracy) suggests that the cue is used in a consistent fashion to make inferences about the conveyed emotion. Bold typeface indicates which cues were both valid and utilized for each emotion. See Table 1 for an explanation of cue abbreviations.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

actors portrayed a large number of distinct emotional categories, including more positive states than typically sampled in cross-cultural research. In addition to neutral tones, the list included anger, contempt, fear, happiness, interest, lust, pride, relief, sadness, and shame. In conducting a project of this unusual size, the study increased by an order of magnitude the body of available data to fill in the basic science of understanding the human voice in expressing emotion. The five cultures served as replications of each other, with portrayals that were collected to be as similar as possible in every manner other than the actors' cultural background.

The resulting data set enabled us to conduct the first direct test of dialect theory (Elfenbein, 2013). In the first step (Study 1), stimuli were analyzed in terms of their fundamental acoustic properties. As predicted, both cultural similarities and systematic differences emerged. In the next step (Study 2), human listeners made forced-choice judgments of these stimuli. There was substantial accuracy across cultures, which indicated a degree of basic universality. There were also cultural differences in the form of an in-group advantage. In the last step, Brunswik's (1956) lens model was used for the first time to study human vocal emotion across cultures, and provided the most precise test to date of dialect theory by examining emotional expression style down the microlevel features of acoustic cues. Results demonstrated that in-group versus out-group judgments showed a greater match versus mismatch between the pattern of emitting acoustic cues and the pattern of judging them. Consistent with dialect theory, this tighter mapping in schemas was responsible for the presence of in-group advantage. Taken together, findings are consistent with interactionist perspectives that attempt to incorporate both universals and cultural differences in emotional expression. This topic has been controversial over the decades (e.g., Russell, 1994), and hard data are worthwhile alongside theoretical arguments.

Considering Origins: Why Paralinguistic Dialects?

This paper provides evidence to document paralinguistic dialects, but does not provide evidence for how or why they develop over time. Why should cultures have dialects in their nonverbal communication of emotion, and why should there be dialects for acoustical properties in particular? We argue that attempts to answer this question benefit from taking seriously the linguistic metaphor of dialects, which allows theories to be grounded in related theory from the field of linguistics. Linguists argue that language is constantly evolving, and that it tends to diverge across groups of people who are separated by geographic or social boundaries (O'Grady, Archibald, Aronoff, & Rees-Miller, 2001). As such, greater social stratification tends to increase the degree of differentiation between speakers of the same language, even if these dialects are mutually intelligible. In the case of verbal language, ultimately a large enough degree of stratification renders languages unable to be mutually understood.

In the case of nonverbal dialects, we argue that the at least partial biological nature of emotional expression prevents drift of nonverbal styles past the point of mutual unintelligibility. Bühler's (1934/1990) Organon model argues that there are three distinct functions for emotional expression (Scherer, 1988). First, expressions can be symptoms of internal states. As such, our basic biology can determine how emotion-related physiological changes can influence the voice production apparatus (Scherer, 1986)—such as throat tightening versus loosening—and this should be similar across cultures. Second, expressions are used as signals to produce a reaction in others (e.g., Fridlund, 1994; Owren & Rendall, 2001). It is likely that deliberate expressions tend to imitate changes in vocal quality that are found in spontaneous expressions, because these cues would be better recognized for their intended meaning (Scherer & Bänziger, 2010). This would serve as another constraint on the magnitude of cultural differences in nonverbal expression—even if it is a looser constraint than the constraint on cues resulting from strictly biological processes. Third, there is a "symbolic" function, in that expressions represent objects or events, similar to linguistic expressions (e.g., Laukka & Elfenbein, 2012; Russell, Bachorowski, & Fernandez-Dols, 2003). Inasmuch as the symbolic function maps most directly onto verbal language—which drifts substantially across cultures—one might expect greater drift across cultures in the evolution of cues that draw from the symbolic function. Taking these three functions together, with their varying potential for cultural difference, nonverbal dialects should be relatively subtle, and mutually intelligible. This is consistent with the base of empirical findings.

The concept of social stratification leads to two distinct mechanisms for the development of dialects, which can occur separately or in tandem. First, some changes in language occur merely through random drift. Particularly in the absence of formal records for vocal characteristics—unlike written text—passing down language from one generation to the next involves evolution through no deliberate effort. Cultural groups that are more physically distant have less opportunity to maintain consistency in their styles that evolve through random drift. As such, dialects may emerge inadvertently, when drifts become shared among some speakers but not others (O'Grady et al., 2001). In the second psychological mechanism, some changes occur through motivated processes of asserting a distinct social identity. For example, jargon and slang can create a marker or even deliberate barrier that defines group membership. Both of these explanations can be consistent with the above results. Notably, we found relatively greater emotion recognition accuracy among culture-pairs that were similar along Hofstede's (2001) dimensions. Cultural similarity between groups may provide less opportunity for random drift. This could help to maintain expressive style similarity and mutual understanding even as these styles potentially change over time. A potential mechanism for this is that cultural similarity might open up greater channels for cross-cultural contact, and verbal communication could help to facilitate convergence in nonverbal expression style.

It is worth noting that the exact form of a dialect does not necessarily need a functional goal. For example, there may be no reason why Bostonians drop the retroflex *r* at the end of a word instead of the dental *t*. This may or may not be the case for nonverbal accents. Functional goals are possible, and some explanations for in-group advantage do not necessarily follow linguistic principles. For example, an appraisal view of nonverbal dialects has the potential to preserve the notion that people across cultures have a universal mapping from their internal feeling states to their outward displays (Hess & Thibault, 2009). The idea here is that emotions exist within broader families—such as irritation, rage, and anger—and cultures may differ in their modal experience within these emotion families (Fontaine, Scherer, & Soriano, 2013). If there is a one-to-one mapping from emotional experi-

ences to the appearance of expressions, then this difference in modal experience could lead to dialects that are better recognized by in-group members. Empirical support would be necessary to support this account, but it is worth mentioning even so as a potential alternative. However, we note that the potential for this explanation in the current study is reduced by design, because we used a protocol in which participants received explicit instructions based on appraisal theory that limited room for idiosyncratic or culturally specific interpretations of the emotion categories.

Limitations and Further Work

A number of important limitations qualify the findings reported above.

One concern is a potential limit to generality in that all five cultures represented in the study are English-speaking. This was intended to control for differences in the words that were used in standard-content sentences spoken by the actors, so that linguistic properties could be kept constant while paralinguistic properties varied. This also keeps relatively consistent across cultures the meaning of the particular emotional categories. As mentioned directly above, there can be subtle differences in the modal experience within broader emotion families (Hess & Thibault, 2009), but we attempted to reduce this potential influence by providing clear guidance about how to interpret emotion labels. Further, in the case of expression, we controlled for group differences in emotion intensity by instructing actors to portray emotions with well-defined levels of intensity. The use of English speakers had the additional benefit of allowing us to sample distinct cultures from four different continents, given the pervasiveness of English around the globe. Nonetheless, it would be valuable to conduct research on the cross-cultural understanding of vocal emotion in multiple languages.

Another limitation is that the judgment data in Study 2 were collected using forced-choice responses, rather than the more naturalistic free labeling of stimuli (Russell, 1993). This was done because the lens model analysis requires the same response categories for emotional expression and perception. Using forced-choice also provided greater methodological convenience, given that there were 275 trials tested for each participant. It is possible that a design using free labeling would provide different insights, in that linguistic categories can be woven into interpersonal judgments (e.g., Gendron, Roberson, van der Vyver, & Barrett, 2014). Future studies using open-ended responses would therefore be valuable.

A limitation that will be important to address in future work is the use of posed expressions instead of spontaneous speech. This is consistent with the vast majority of research that has been conducted about recognizing emotional expression across cultures. We followed this dominant convention without endorsing it as complete. The use of deliberate expression made it possible to elicit each of 11 distinct emotional states each from a large number of actors. However, one potential limitation with acted portrayals is that actors may sometimes develop norms to exaggerate their expressions in order to make them more recognizable, or otherwise produce speech samples that differ from what would be produced spontaneously. As such, deliberate portrayals could come at the tradeoff of authenticity. To address this, the protocol instructed actors not to use overtly stereotypical expressions, and to use moderate levels of intensity. Further, the protocol involved the Stanislavski (1936) technique of method acting, in which speakers attempted to reactivate past authentic emotional experience. The Stanislavski technique has the potential to increase authenticity and help mitigate the challenge of producing vocal cues by voluntary intention. The use of a single professional technique across all 100 actors also attempted to reduce potential cultural differences in the training of actors, as well as idiosyncratic differences in acting styles. Although we believe that professional actors may provide the best case scenario to enact voluntary nonverbal cues (see Scherer & Bänziger, 2010), it could also be a worthwhile extension to examine portrayals from speakers who are not professional actors.

In discussing this concern, we note empirical evidence that speak against the possibility that it threatens the validity of the present work. Posed and spontaneous expressions tend to produce highly similar patterns of acoustic cues, and the differences reported tend to be relatively small. In a recent investigation, Scherer (2013) compared speech samples obtained from mood induction and acting procedures, and reported that speech samples from both techniques tended to be comparable at the acoustic level. In addition, evidence from human listeners suggests the similarity of acted and spontaneous portrayals. Participants tend to perform poorly when asked to distinguish whether stimuli are posed versus portrayed (Jürgens et al., 2013). Reviewing these and other findings, Scherer (2013) challenged the notion that using acted samples in empirical research is a concern. Taken together, we believe that the limitation of using posed expressions is an important one, but that it does not invalidate the research presented here. Nevertheless, we would welcome efforts that aim to replicate the current results using instead spontaneous expressions. It will be a worthwhile challenge to develop an experimental manipulation for each of these 11 categories that would involve the production of spontaneous same-content verbal utterances. Increased efforts have recently been directed at developing databases of spontaneous affective speech in several research communities (e.g., Schuller, Batliner, Steidl, & Seppi, 2011). We speculate that, when suitable databases of spontaneous expressions become available, it may also be possible to conduct comparisons of the acoustic properties of our stimuli and spontaneous vocal tones using data from speech collections not originally designed for cross-cultural comparisons.[7]

A final limitation we note is the potential for linguistic accents to be confounded with paralinguistic accents. In past work, in-group advantage has been documented even when linguistic dialects were not a plausible explanation. In particular, some studies used expressions standardized to be identical across cultural groups, and in others the apparent cultural origin of stimuli was experimentally manipulated (for reviews, see Elfenbein & Ambady, 2002; Sauter, 2013; Thibault, Bourgeois, & Hess, 2006). As such, the sound of a clearly foreign accent could lead to lesser motivation and accuracy, but verbal accents cannot explain away the lens model analysis presented above. This is because the lens analysis was conducted after standardizing all acoustic cues within-speaker. Because the stimuli consisted of standard-content sentences, the influence on speech of a verbal accent should be the same across each emotion category. Analyses examined the rela-

[7] We thank an anonymous reviewer for this suggestion.

tive change in the use of cues from one emotion to the next. As such, any systematic pattern in the use of cues across particular emotions is part of the phenomenon itself—that is, showing that there can be cultural differences in paralinguistic style.

Practical Implications: Practice Makes Perfect

We argue that the findings above present a source of optimism. To the extent that cultural differences in emotional expression style can create barriers for individuals who interact across cultural group boundaries, these barriers can be overcome. Dialect theory focuses on the role of information in explaining cultural differences in emotion recognition accuracy. If someone lacks the necessary familiarity with another group's dialects, this is remediable versus destiny. By contrast, it would be harder to overcome a deficit resulting from biological preprogramming, or even from prejudice against members of foreign groups. Past research based on facial expressions has shown that individuals can learn to bridge the gap through study abroad and even experimental feedback (Elfenbein, 2006; Elfenbein & Ambady, 2003b). In a recent study examining vocal tones, Altrov (2013) found likewise that Russians who lived in Estonia were more accurate with Estonian expressions than Russians who lived in Russia. Integrating the available evidence emphasizes the value of expanding our horizons. Cross-group contact and even deliberate training can help maximize our ability to communicate with individuals from all around the world.

References

Albas, D. C., McCluskey, K. W., & Albas, C. A. (1976). Perception of the emotional content of speech: A comparison of two Canadian groups. *Journal of Cross-Cultural Psychology, 7,* 481–490. http://dx.doi.org/10.1177/002202217674009

Altrov, R. (2013). Aspects of cultural communication in recognizing emotions. *Trames, 17,* 159–174. http://dx.doi.org/10.3176/tr.2013.2.04

Altrov, R., & Pajupuu, H. (2015). The influence of language and culture on the understanding of vocal emotions. *Journal of Estonian and Finno-Ugric Linguistics, 6,* 11–48. http://dx.doi.org/10.12697/jeful.2015.6.3.01

Anolli, L., Wang, L., Mantovani, F., & de Toni, A. (2008). The voice of emotions in Chinese and Italian young adults. *Journal of Cross-Cultural Psychology, 39,* 565–598. http://dx.doi.org/10.1177/0022022108321178

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70,* 614–636. http://dx.doi.org/10.1037/0022-3514.70.3.614

Bänziger, T., Hosoya, G., & Scherer, K. R. (2015). Path models of vocal emotion communication. *PLoS ONE, 10,* e0136675. http://dx.doi.org/10.1371/journal.pone.0136675

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion, 12,* 1161–1179. http://dx.doi.org/10.1037/a0025827

Beaupré, M. G., & Hess, U. (2005). Cross-cultural emotion recognition among Canadian ethnic groups. *Journal of Cross-Cultural Psychology, 36,* 355–370. http://dx.doi.org/10.1177/0022022104273656

Beier, E. G., & Zautra, A. J. (1972). Identification of vocal communication of emotions across cultures. *Journal of Consulting and Clinical Psychology, 39,* 166. http://dx.doi.org/10.1037/h0033170

Bernieri, F. J., Gillis, J. S., Davis, J. M., & Grahe, J. E. (1996). Dyad rapport and the accuracy of its judgment across situations: A lens model analysis. *Journal of Personality and Social Psychology, 71,* 110–129. http://dx.doi.org/10.1037/0022-3514.71.1.110

Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., Kudoh, T., & Ton, V. (1997). Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability data and cross-national differences. *Journal of Nonverbal Behavior, 21,* 3–21. http://dx.doi.org/10.1023/A:1024902500935

Boersma, P., & Weenink, D. (2008). Praat: Doing phonetics by computer [Computer software]. Retrieved from http://www.praat.org

Braun, A., & Heilmann, C. M. (2012). *SynchronEmotion*. Bern, Switzerland: Peter Lang. http://dx.doi.org/10.3726/978-3-653-01595-9

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

Bryant, G. A., & Barrett, H. C. (2008). Vocal emotion recognition across disparate cultures. *Journal of Cognition and Culture, 8,* 135–148. http://dx.doi.org/10.1163/156770908X289242

Bühler, K. (1990). *Theory of language. The representational function of language* (D. F. Goodwin, Trans.). Amsterdam, the Netherlands: John Benjamins. http://dx.doi.org/10.1075/fos.25 (Original work published 1934)

Cordaro, D. T., Keltner, D., Tshering, S., Wangchuk, D., & Flynn, L. M. (2016). The voice conveys emotion in ten globalized cultures and one remote village in Bhutan. *Emotion, 16,* 117–128. http://dx.doi.org/10.1037/emo0000100

Davitz, J. R. (1964). Minor studies and some hypotheses. In J. R. Davitz (Ed.), *The communication of emotional meaning* (pp. 143–156). New York, NY: McGraw-Hill.

Dinno, A. (2009). Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data. *Multivariate Behavioral Research, 44,* 362–388. http://dx.doi.org/10.1080/00273170902938969

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6,* 169–200. http://dx.doi.org/10.1080/02699939208411068

Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry, 32,* 88–106.

Elfenbein, H. A. (2006). Learning in emotion judgments: Training and the cross-cultural understanding of facial expressions. *Journal of Nonverbal Behavior, 30,* 21–36. http://dx.doi.org/10.1007/s10919-005-0002-y

Elfenbein, H. A. (2013). Nonverbal dialects and accents in facial expressions of emotion. *Emotion Review, 5,* 90–96. http://dx.doi.org/10.1177/1754073912451332

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128,* 203–235. http://dx.doi.org/10.1037/0033-2909.128.2.203

Elfenbein, H. A., & Ambady, N. (2003a). Cultural similarity's consequences: A distance perspective on cross-cultural differences in emotion recognition. *Journal of Cross-Cultural Psychology, 34,* 92–110. http://dx.doi.org/10.1177/0022022102239157

Elfenbein, H. A., & Ambady, N. (2003b). When familiarity breeds accuracy: Cultural exposure and facial emotion recognition. *Journal of Personality and Social Psychology, 85,* 276–290. http://dx.doi.org/10.1037/0022-3514.85.2.276

Elfenbein, H. A., Beaupré, M., Lévesque, M., & Hess, U. (2007). Toward a dialect theory: Cultural differences in the expression and recognition of posed facial expressions. *Emotion, 7,* 131–146. http://dx.doi.org/10.1037/1528-3542.7.1.131

Elfenbein, H. A., Mandal, M. K., Ambady, N., Harizuka, S., & Kumar, S. (2002). Cross-cultural patterns in emotion recognition: Highlighting design and analytical techniques. *Emotion, 2,* 75–84. http://dx.doi.org/10.1037/1528-3542.2.1.75

Elfenbein, H. A., Mandal, M. K., Ambady, N., Harizuka, S., & Kumar, S. (2004). Hemifacial differences in the in-group advantage in emotion recognition. *Cognition and Emotion, 18,* 613–629. http://dx.doi.org/10.1080/02699930341000257

Ellsworth, P. C., & Scherer, K. R. (2003). Appraisal processes in emotion.

In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 572–595). New York, NY: Oxford University Press.

Eyben, F., Scherer, K. R., Schuller, B., Sundberg, J., André, E., Busso, C., . . . Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing, 7,* 190–202. http://dx.doi.org/10.1109/TAFFC.2015.2457417

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In A. Jaimes, N. Sebe, N. Boujemaa, D. Gatica-Perez, D. A. Shamma, M. Worring, & R. Zimmermann (Eds.), *Proceedings of the 21st Association for Computing Machinery International Conference on Multimedia* (pp. 835–838). New York, NY: Association for Computing Machinery. http://dx.doi.org/10.1145/2502081.2502224

Fiske, A., Kitayama, S., Markus, H. R., & Nisbett, R. E. (1998). The cultural matrix of social psychology. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 915–981). San Francisco, CA: McGraw-Hill.

Fonagy, I., & Magdics, K. (1963). Emotional patterns in intonation and music. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung, 16,* 293–326.

Fontaine, J. J. R., Scherer, K. R., & Soriano, C. (Eds.). (2013). *Components of emotional meaning: A sourcebook*. New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199592746.001.0001

Fridlund, A. J. (1994). *Human facial expression: An evolutionary view*. San Diego, CA: Academic Press.

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014). Cultural relativity in perceiving emotion from vocalizations. *Psychological Science, 25,* 911–920. http://dx.doi.org/10.1177/0956797613517239

Gifford, R. (1994). A lens-mapping framework for understanding the encoding and decoding of interpersonal dispositions in nonverbal behavior. *Journal of Personality and Social Psychology, 66,* 398–412. http://dx.doi.org/10.1037/0022-3514.66.2.398

Graham, C. R., Hamblin, A., & Feldstein, S. (2001). Recognition of emotion in English voices by speakers of Japanese, Spanish, and English. *International Review of Applied Linguistics in Language Teaching, 39,* 19–37. http://dx.doi.org/10.1515/iral.39.1.19

Hall, J. A., Andrzejewski, S. A., & Yopchick, J. E. (2009). Psychosocial correlates of interpersonal sensitivity: A meta-analysis. *Journal of Nonverbal Behavior, 33,* 149–180. http://dx.doi.org/10.1007/s10919-009-0070-5

Hammerschmidt, K., & Jürgens, U. (2007). Acoustical correlates of affective prosody. *Journal of Voice, 21,* 531–540. http://dx.doi.org/10.1016/j.jvoice.2006.03.002

Hess, U., & Thibault, P. (2009). Darwin and emotion expression. *American Psychologist, 64,* 120–128. http://dx.doi.org/10.1037/a0013386

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage.

Izard, C. E. (1971). *The face of emotions*. New York, NY: Appleton-Century-Crofts.

Jack, R. E., Caldara, R., & Schyns, P. G. (2012). Internal representations reveal cultural diversity in expectations of facial expressions of emotion. *Journal of Experimental Psychology: General, 141,* 19–25. http://dx.doi.org/10.1037/a0023463

Jarvis, B. G. (2008). MediaLab (Version 2008) [Computer software]. New York, NY: Empirisoft Corporation.

Jürgens, R., Drolet, M., Pirow, R., Scheiner, E., & Fischer, J. (2013). Encoding conditions affect recognition of vocally expressed emotions across cultures. *Frontiers in Psychology, 4,* 111. http://dx.doi.org/10.3389/fpsyg.2013.00111

Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion, 1,* 381–412. http://dx.doi.org/10.1037/1528-3542.1.4.381

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin, 129,* 770–814. http://dx.doi.org/10.1037/0033-2909.129.5.770

Kramer, E. (1964). Elimination of verbal cues in judgments of emotion from voice. *Journal of Abnormal Psychology, 68,* 390–396. http://dx.doi.org/10.1037/h0042473

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. Chichester, UK: Wiley-Blackwell. http://dx.doi.org/10.1002/9781444395068

Laukka, P., Eerola, T., Thingujam, N. S., Yamasaki, T., & Beller, G. (2013). Universal and culture-specific factors in the recognition and performance of musical affect expressions. *Emotion, 13,* 434–449. http://dx.doi.org/10.1037/a0031388

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

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. In L. Devillers, B. Schuller, R. Cowie, E. Douglas-Cowie, & A. Batliner (Eds.), *Proceedings of the LREC 2010 Workshop on Corpora for Research on Emotion and Affect* (pp. 53–57). Paris, France: European Language Resources Association.

Laukka, P., Neiberg, D., & Elfenbein, H. A. (2014). Evidence for cultural dialects in vocal emotion expression: Acoustic classification within and across five nations. *Emotion, 14,* 445–449. http://dx.doi.org/10.1037/a0036048

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,* 84–104. http://dx.doi.org/10.1016/j.csl.2010.03.004

Lazarus, R. S. (1991). *Emotion and adaptation*. New York, NY: Oxford University Press.

Marsh, A. A., Elfenbein, H. A., & Ambady, N. (2003). Nonverbal "accents": Cultural differences in facial expressions of emotion. *Psychological Science, 14,* 373–376. http://dx.doi.org/10.1111/1467-9280.24461

McCluskey, K. W., Albas, D. C., Niemi, R. R., Cuevas, C., & Ferrer, C. A. (1975). Cross-cultural differences in the perception of emotional content of speech: A study of the development of sensitivity in Canadian and Mexican children. *Developmental Psychology, 11,* 551–555. http://dx.doi.org/10.1037/0012-1649.11.5.551

Mesquita, B., & Frijda, N. H. (1992). Cultural variations in emotions: A review. *Psychological Bulletin, 112,* 179–204. http://dx.doi.org/10.1037/0033-2909.112.2.179

O'Grady, W., Archibald, J., Aronoff, M., & Rees-Miller, J. (2001). *Contemporary linguistics* (4th ed.). Boston, MA: Bedford/St. Martin's Press.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511571299

Owren, M. J., & Rendall, D. (2001). Sound on the rebound: Bringing form and function back to the forefront in understanding nonhuman primate vocal signaling. *Evolutionary Anthropology, 10,* 58–71. http://dx.doi.org/10.1002/evan.1014

Patel, S., Scherer, K. R., Björkner, E., & Sundberg, J. (2011). Mapping emotions into acoustic space: The role of voice production. *Biological Psychology, 87,* 93–98. http://dx.doi.org/10.1016/j.biopsycho.2011.02.010

Paulmann, S., & Uskul, A. K. (2014). Cross-cultural emotional prosody recognition: Evidence from Chinese and British listeners. *Cognition and Emotion, 28,* 230–244. http://dx.doi.org/10.1080/02699931.2013.812033

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior, 33,* 107–120. http://dx.doi.org/10.1007/s10919-008-0065-7

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. http://dx.doi.org/10.1016/j.wocn.2009.07.005

Raphael, L. J., Borden, G. J., & Harris, K. S. (2011). *Speech science primer: Physiology, acoustics, and perception of speech* (6th rev. ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test*. Baltimore, MD: Johns Hopkins University Press.

Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science, 11,* 446–453. http://dx.doi.org/10.1111/1467-9280.00287

Ross, E. D., Edmondson, J. A., & Seibert, G. B. (1986). The effect of affect on various acoustic measures of prosody in tone and non-tone languages: A comparison based on computer analysis of voice. *Journal of Phonetics, 14,* 215–223.

Russell, J. A. (1993). Forced-choice response format in the study of facial expression. *Motivation and Emotion, 17,* 41–51. http://dx.doi.org/10.1007/BF00995206

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin, 115,* 102–141. http://dx.doi.org/10.1037/0033-2909.115.1.102

Russell, J. A., Bachorowski, J.-A., & Fernandez-Dols, J.-M. (2003). Facial and vocal expressions of emotion. *Annual Review of Psychology, 54,* 329–349. http://dx.doi.org/10.1146/annurev.psych.54.101601.145102

Sauter, D. A. (2010). More than happy: The need for disentangling positive emotions. *Current Directions in Psychological Science, 19,* 36–40. http://dx.doi.org/10.1177/0963721409359290

Sauter, D. A. (2013). The role of motivation and cultural dialects in the in-group advantage for emotional vocalizations. *Frontiers in Psychology, 4,* 914. http://dx.doi.org/10.3389/fpsyg.2013.00814

Sauter, D. A., Eisner, F., Calder, A. J., & Scott, S. K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *Quarterly Journal of Experimental Psychology, 63,* 2251–2272. http://dx.doi.org/10.1080/17470211003721642

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. http://dx.doi.org/10.1073/pnas.0908239106

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99,* 143–165. http://dx.doi.org/10.1037/0033-2909.99.2.143

Scherer, K. R. (1988). On the symbolic functions of vocal affect expression. *Journal of Language and Social Psychology, 7,* 79–100. http://dx.doi.org/10.1177/0261927X8800700201

Scherer, K. R. (1997). The role of culture in emotion-antecedent appraisal. *Journal of Personality and Social Psychology, 73,* 902–922. http://dx.doi.org/10.1037/0022-3514.73.5.902

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

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

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. http://dx.doi.org/10.1177/0022022101032001009

Scherer, K. R., & Bänziger, T. (2010). On the use of actor portrayals in research on emotional expression. In K. R. Scherer, T. Bänziger, & E. B. Roesch (Eds.), *Blueprint for affective computing: A sourcebook* (pp. 271–294). New York, NY: Oxford University Press.

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,* 401–435. http://dx.doi.org/10.1080/00207594.2011.626049

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognizing realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication, 53,* 1062–1087. http://dx.doi.org/10.1016/j.specom.2011.01.011

Shochi, T., Rilliard, A., Aubergé, V., & Erickson, D. (2009). Intercultural perception of English, French and Japanese social affective prosody. In S. Hancil (Ed.), *The role of prosody in affective speech* (pp. 31–60). Bern, Switzerland: Peter Lang.

Stanislavski, C. (1936). *An actor prepares*. New York, NY: Theatre Arts Books/Methuen.

Sundberg, J., Patel, S., Björkner, E., & Scherer, K. R. (2011). Interdependencies among voice source parameters in emotional speech. *IEEE Transactions on Affective Computing, 2,* 162–174. http://dx.doi.org/10.1109/T-AFFC.2011.14

Tangney, J. P., & Tracy, J. L. (2012). Self-conscious emotions. In M. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (2nd ed., pp. 446–478). New York, NY: Guilford Press.

Thibault, P., Bourgeois, P., & Hess, U. (2006). The effect of group-identification on emotion recognition: The case of cats and basketball players. *Journal of Experimental Social Psychology, 42,* 676–683. http://dx.doi.org/10.1016/j.jesp.2005.10.006

Thompson, W. F., & Balkwill, L.-L. (2006). Decoding speech prosody in five languages. *Semiotica, 158,* 407–424. http://dx.doi.org/10.1515/SEM.2006.017

Tomkins, S. S., & McCarter, R. (1964). What and where are the primary affects? Some evidence for a theory. *Perceptual and Motor Skills, 18,* 119–158. http://dx.doi.org/10.2466/pms.1964.18.1.119

Tracy, J. L., & Randles, D. (2011). Four models of basic emotions: A review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emotion Review, 3,* 397–405. http://dx.doi.org/10.1177/1754073911410747

van Bezooijen, R., Otto, S. A., & Heenan, T. A. (1983). Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics. *Journal of Cross-Cultural Psychology, 14,* 387–406. http://dx.doi.org/10.1177/0022002183014004001

Waaramaa, T. (2015). Perception of emotional nonsense sentences in China, Egypt, Estonia, Finland, Russia, Sweden, and the USA. *Logopedics, Phoniatrics, Vocology, 40,* 129–135. http://dx.doi.org/10.3109/14015439.2014.915982

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

Zhu, Y. (2013). Which is the best listener group? Perception of Chinese emotional prosody by Chinese natives, naive Dutch listeners, and Dutch L2 learners of Chinese. *Dutch Journal of Applied Linguistics, 2,* 170–183. http://dx.doi.org/10.1075/dujal.2.2.03zhu

Received December 23, 2014
Revision received May 11, 2016
Accepted May 22, 2016 ■