RUNNING TITLE: CROSS-CULTURAL EMOTIONAL PROSODY RECOGNITION

Cross-cultural emotional prosody recognition:

Evidence from Chinese and British listeners

Silke Paulmann* & Ayse K. Uskul[1, **]

University of Essex

In Press, *Cognition & Emotion*

Key words: emotion, culture, tone of voice, in-group hypothesis, bilingualism

*Corresponding Author:
Silke Paulmann
University of Essex
Department of Psychology & Centre for Brain Science
& Centre for Brain Science
Wivenhoe Park
Colchester
CO4 3SQ
United Kingdom
Email: paulmann@essex.ac.uk; Phone: 01206-873422; Fax: 01206-873801

** Authors share first authorship.

[1] Uskul is now at the School of Psychology, University of Kent.

Authors' Note

We would like to thank Sarah Harris, Laura King, Constance Lau, Yuchen Lao, and Hoi Lee for their help with stimulus preparation and data collection.

Abstract

This cross-cultural study of emotional tone of voice recognition tests the in-group advantage hypothesis (Elfenbein & Ambady, 2002) employing a quasi-balanced design. Individuals of Chinese and British background were asked to recognize pseudo-sentences produced by Chinese and British native speakers, displaying one of seven emotions (anger, disgust, fear, happy, neutral tone of voice, sad, and surprise). Findings revealed that emotional displays were recognized at rates higher than predicted by chance; however, members of each cultural group were more accurate in recognizing the displays communicated by a member of their own cultural group than a member of the other cultural group. Moreover, the evaluation of error matrices indicates that both culture groups relied on similar mechanism when recognizing emotional displays from the voice. Overall, the study reveals evidence for both universal and culture-specific principles in vocal emotion recognition.

Word count: 137

You are sitting in a Chinese restaurant in Soho, London: the customer next to you is talking to the restaurant manager in Chinese and although you cannot understand what they are saying (you do not speak Chinese), you are certain that he is very antagonized simply because of *the way in which* he says things – he is using a subtly pressed, harsh tone of voice. Your 'emotion evaluation' is confirmed later when the manager brings in a newly prepared dish and the customer turns to you mumbling in English that he cannot believe it took three attempts to finally get what he ordered. This example suggests that listeners can make inferences about others' feelings in the absence of meaningful words by paying attention to the melodic and rhythmic attributes of the spoken utterance. In fact, so-called prosodic features of language are long known to play an important role in spoken interactions. By varying characteristics such as pitch, loudness, voice timbre, speed, and rhythm of speech, a range of linguistic (e.g. stating/questioning) and non-linguistic (e.g. emotion/attitudes) functions can be communicated. In the current study, we focus on the non-linguistic function of emotion communication and investigate how emotional displays are de- and encoded cross-culturally.

The situation described above suggests that emotions are conveyed similarly across different languages and cultures. Indeed, based on the observation that cross-cultural emotional tone of voice recognition is typically better than expected by chance, researchers have argued that emotion relevant prosodic cues are decoded in a *universal* manner (e.g. Scherer, Banse & Wallbott, 2001). However, consider someone using a level (i.e. even pitch) tone when speaking in English. Based on the speaker's non-modulating pitch use, this person is likely going to be interpreted as being bored or potentially unfriendly by native speakers of English. Native speakers of Russian, however, would be unlikely to make the same inference as using a mono-tone pitch pattern is rather common in Russian. Thus, (emotional) prosodic cue decoding may be

shaped by language and culture. This argument is supported by evidence demonstrating an *in-group advantage* with native listeners outperforming non-native listeners in emotional prosody recognition (e.g., Beier & Zautra, 1972; Scherer et al., 2001; Van Bezooijen, Otto, & Heenan, 1983).

Given the findings pointing to cross-cultural similarities *and* differences in decoding emotions from tone of voice, researchers have developed accounts of emotion recognition to explain both universality and linguistic and cultural variation (e.g. Elfenbein & Ambady, 2002; Matsumoto, 2006). These theories are based primarily on data originating from studies that investigated *facial* emotion recognition. In comparison, much fewer studies have explored cross-cultural *vocal* emotion recognition. The limited amount of research on the vocal channel is surprising given the increasing importance of spoken communication between members of different cultures in organizational (e.g., employers of multinational companies working together), educational (e.g., teachers of one cultural group interacting with students from other cultural groups), health (e.g., counseling services provided by one member of a cultural groups to a member of another cultural group), and interpersonal (e.g., interethnic marriages) settings. Moreover, as evident from a meta-analysis comparing vocal and facial emotion recognition, cross-cultural emotion accuracy is generally lower in studies using voice materials (c.f. Elfenbein & Ambady, 2002). Thus, further research exploring cross-cultural emotional prosody recognition is needed to help develop present accounts on how emotions are communicated (vocally) across cultures.

Empirical Evidence on Cross-Linguistic Emotional Prosody Recognition

Studies investigating emotional recognition in the voice by members of different cultural or linguistic groups can be grouped into three categories. *One against all*:

listeners in different cultural groups are asked to identify emotions expressed by a speaker in one cultural group; *all against one*: listeners in one cultural group are asked to identify emotions expressed by speakers in different cultural groups; and *all against all* (fully-crossed design): listeners in different cultural groups are asked to identify emotions expressed by speakers in these same cultural groups. The majority of studies fall in one of the first two categories, with a limited number falling in the third category. Regardless of the adopted design, these studies aim to determine whether the ability to recognize emotions from speech is independent of language and cultural background (universalist approach) or whether successful emotional prosody decoding is shaped by cultural or linguistic variables (e.g., in-group advantage). Below we briefly summarize studies that investigated cross-cultural *vocal* emotion recognition.

One against all

Beier and Zautra (1972) asked American, Polish, and Japanese listeners to judge sentences of different length (e.g. hello, how are you?) intoned in different emotions by American English speakers. They reported an in-group advantage for American English listeners when recognizing emotions from short speech samples as American English listeners outperformed Polish and Japanese listeners in emotion recognition from American English prosody. However, this in-group advantage disappeared when recognizing emotions from *longer* speech samples, suggesting that exposure duration influences cross-cultural emotion recognition. In a similar vein, Van Bezooijen and colleagues (1983) studied vocal emotion recognition using a short phrase produced by several Dutch speakers. Groups of about 40 young adults each from the Netherlands, Taiwan, and Japan were able to recognize the intended emotions with better than chance accuracy. Again, an in-group advantage was observed: Dutch participants performed

significantly better in identifying Dutch vocal emotion expressions than Taiwanese or Japanese participants.

Similar results were found in one of the most comprehensive studies on cross-cultural emotional prosody recognition conducted by Scherer, Banse, and Wallbott (2001). Here, listeners from nine different countries across three continents were presented with 30 semantically-anomalous pseudo-utterances spoken in five emotional tones by four native German actors. While all listener groups recognized fearful, joyful, sad, angry and neutral utterances at above chance accuracy levels (66% accuracy overall in a five-choice task), Scherer et al. (2001) also found evidence for an in-group advantage. German (i.e., native) listeners performed significantly better on the emotion recognition task than any other language group.

All against one

Kramer (1964) reported that American judges could identify vocal expressions of emotions by American (content-filtered) and Japanese speakers (unknown language) with better than chance accuracy. Thompson and Balkwill (2006) explored how well native speakers of English recognized semantically neutral but emotionally-inflected sentences spoken by native speakers of German, English, Chinese, Japanese, and Tagalog. While listeners were generally successful at recognizing emotions from non-native speech material, significantly better recognition rates were observed for stimuli spoken in English than in any other language; stimuli spoken by Japanese and Chinese speakers were the most difficult to recognize.

Similarly, Pell, Monetta, Paulmann and Kotz (2009) presented native speakers of Argentine Spanish with emotionally-inflected pseudo-sentences spoken by native speakers of Arabic, English, German and Spanish. They found evidence for an in-group

advantage as Argentine Spanish speakers were better at recognizing emotions from materials spoken in their native language.

Taken together, findings from the studies adopting a *one against all* or *all against one* approach to investigate cross-cultural emotion recognition suggest that listeners rely on universal inference rules (i.e. similar decoding mechanisms underlying vocal emotion recognition across cultures) when recognizing emotions from speech (Scherer et al., 2001), but the majority of these studies also reveal evidence for an in-group advantage suggesting culture-specific prosody cue use.

Although informative, *one against all* or *all against one* designs have been criticized for not being able to eliminate alternative explanations such as some cultures being more or less expressive (and hence more/less easily recognizable) than others or listeners of one group being superior in their decoding skills than those in other groups. Thus, it has been argued that to adequately test any in-group advantage and provide strongest source of evidence, one must consider the cultural match between decoders and encoders of emotional displays, rather than considering either group independently (see Elfenbein & Ambady, 2002; 2003). So far, only a handful of studies have employed such a fully-balanced design which we review below.

All against all

McCluskey and Albas (1981) explored how Canadian and Mexican listeners judged vocal expressions produced by Canadian and Mexican speakers after the semantic content of the speech was removed by means of low-pass filtering. They found that overall Mexican listeners were more accurate than English listeners in identifying stimuli spoken in Spanish and English. Interestingly, however, both Canadian and Mexican participants judged speech samples from Mexican speakers more accurately

than those from Canadian speakers. Similar findings were observed in a study with Mexican and Canadian children (McCluskey, Albas, Niemi, Cuevas, & Ferrer, 1975). These findings suggest that a mismatch between the speaker and the listener does not have to be associated with a decreased performance in emotion recognition.

In another study, Albas, McCluskey, and Albas (1976) asked male Caucasian and Native American Cree speakers to express happiness, sadness, anger, and love, in a free word choice task. Their answers were recorded and speech samples were again low-pass filtered to eliminate semantic information. These stimuli were presented to Caucasian and Cree listeners. Here, results showed that each group of listeners showed superior performance when inferring emotions expressed by a member of their own group, providing evidence for an in-group advantage. However, Albas and colleagues admitted that their data are difficult to interpret as the content of the speech material used for encoding was not controlled: speakers may have used culture-specific expressions resulting in language acting as a confounding factor and thus leading to the observed in-group advantage.

More recently, Sauter, Eisner, Ekman and Scott (2010) examined the recognition of nonverbal emotion vocalizations such as screams and laughs produced by both Himba and British speakers. Recognition rates of Western participants were compared with recognition performance of the Himba who live in isolated villages in Namibia. They found evidence for recognition of these signals above chance in both cultural groups. The authors also report an in-group advantage with each group showing greater recognition accuracy for stimuli produced by members of their own cultural group. However, considering that stimulus duration has been argued to be an important factor when recognizing emotions cross-culturally (e.g. Beier & Zautra, 1972), the question remains whether this in-group advantage is replicable when using sentence materials

rather than short vocalizations as test stimuli.

Thus, evidence originating from studies with balanced designs concerning an in-group advantage in emotional tone of voice recognition is inconsistent. While some studies report an in-group advantage (Albas et al, 1976; Sauter et al., 2010), others fail to do so (McCluskey & Albas, 1981; McCluskey et al., 1975). In the current work, we aim to further test the in-group advantage hypothesis using a quasi-balanced design by comparing the ability of Chinese and British individuals to recognize emotional prosodic displays used by Chinese (Chinese speaking) and British (English speaking) speakers while addressing several methodological shortcomings of previous work in the following ways.

The Present Study

First, we aimed to control for 'stimulus equivalence' by ensuring that stimuli in both languages were highly recognizable. While some researchers (e.g., Matsumoto, 2002) have maintained that stimuli need not only be highly recognizable but actually *morphologically* identical, we argue that this approach is not necessarily ecologically valid for linguistic materials. If different language groups communicate emotional intention through different modulations of the voice, this feature of the language would be an important factor to also be considered in research context. Hence, these acoustic characteristics should not be influenced by forcing speakers to follow pre-determined display rules. In fact, Marsh, Elfenbein and Ambady (2003) postulate that it is almost impossible to eliminate cultural differences in emotion portrayals as they are hard to overcome. Moreover, as argued by others previously, researchers need to ensure that materials sound natural and not exaggerated. Thus, in the present study we controlled for "stimulus equivalence" by ensuring that stimuli are recognizable as determined by

native listeners and statistical models (discriminant analysis).

Second, it has been criticized (e.g. Elfenbein & Ambady, 2002) that few studies report error matrices which allow exploring whether emotional prosody recognition relies on similar mechanisms across culture and language groups. Thus, in the present study, we explored error patterns in detail. Moreover, we also report results from acoustical analyses and how acoustics can predict confusion patterns to assess whether misclassifications are based on (misleading) primary acoustic cues.

Finally, to control for the confounding factor of segmental information while testing recognition of supra-segmental information, we adopted a procedure employed in several comparative projects designed to examine vocal expressions in different languages (e.g., Castro & Lima, 2010; Pell, Paulmann, et al. 2009) and constructed sentences that were stripped off their semantic content (pseudo-sentences). These sentences were first entered into a perceptual rating study before being used in the main study designed to test the in-group advantage hypothesis.

Pilot Study

Participants

Participants were 31 native speakers of British English (21 female, M_{age} = 24.09, *SD* = 9.69) and 42 native speakers of Chinese (21 Female, M_{age} = 21.34, *SD* = 2.00), recruited on voluntary basis.

Materials and Procedure

Two native speakers of British English and two native speakers of Chinese created forty pseudo sentences in each language, which constituted the stimulus material.[1] All sentences retained natural phonological and morpho-syntactic properties

of the target language where meaningful content words were replaced with plausible pseudo words (e.g., English: Flotch deraded the downdary snat, Chinese: 寧溫雞吐不不非糖哈). Word and syllable length were controlled in both sets of sentences. A female native speaker of British English produced English sentences and a female native speaker of Chinese produced Chinese sentences to convey angry, happy, disgust, sad, pleasant surprise, fear and neutral affect. Each speaker was recruited from an acting school and had considerable acting experience. The 280 stimuli (40 sentences X 7 emotional displays) in each language were recorded in a soundproof booth using a high-quality fixed microphone while monitoring sound intensity to avoid clipping of files.

Stimuli

Digital recordings of all sentences were transferred to a computer, edited to isolate the onset and offset of each sentence, and then entered into a perceptual validation study to examine the extent to which the intended emotion of the speaker was successfully identified by individuals from the same linguistic background. In individual cubicles, participants listened to 40 pseudo-sentences expressed in 6 different emotions (anger, disgust, fear, happiness, surprise, and sadness) and in a neutral tone of voice, using Sennheiser headphones. They were asked to identify the emotion conveyed in each sentence as quickly and accurately as possible by clicking on one of seven response options that represented the types of emotional displays. The study instructions and response options were presented to both groups in English, but Chinese translations of the response options (i.e., emotion labels) were also provided to Chinese participants in case they needed to refer to those for clarification. The task started with 5 practice sentences followed by 280 sentences presented randomly in seven blocks; each block was followed by a mini-break.

We then computed the percentage of native listeners who accurately identified the target emotion communicated in each of the forty pseudo sentences produced in their own language. For both English and Chinese sentences, we selected sentences that were recognized at least three times above chance level (42.6% with a chance accuracy level of 14.2%)[2]. This selection procedure resulted in 28 sentences in each language category to be used in the main study. The average accuracy rates for selected English pseudo-sentences were: anger = 91%, disgust = 83%, fear = 64%, happiness = 55%, neutral=91%, sadness = 81%, pleasant surprise = 69%. The average accuracy rates for selected Chinese pseudo-sentences were: anger = 87%, disgust = 33%, fear = 55%, happiness = 56%, neutral = 98%, sadness = 80%, pleasant surprise = 80%.

Acoustics. We first acoustically analyzed the selected stimuli using *Praat* (Boersma & Weenink, 2009; see Table 1). To infer whether primary acoustic features (mean pitch, mean loudness, and mean duration) of stimuli could predict intended emotional categories, a discriminant analysis[3] was performed. Acoustic characteristics served as independent variables and intended emotional category served as the dependent variable. Results for Chinese materials revealed that for 69.4% of the sentences emotional category membership was predicted correctly: anger, 71.4%; disgust 66.7%; fear, 47.4%; happiness, 50.0%; neutral, 100%; sadness, 66.7%; pleasant surprise, 80%. Results for English sentences were similar as 59.2% of sentences were classified correctly: anger, 28.6%; disgust 60.7%; fear, 53.6%; happiness, 46.4%; neutral, 82.1%; sadness 71.4%; pleasant surprise, 71.4%.[4] Taken together, results from the discriminant analysis confirmed that stimuli contained detectable primary acoustic features that could be used by listeners to correctly differentiate between intended emotional categories.

Main Study

Method

In the following, we describe our original sample size, all data exclusions and manipulations, as well as all measures that we collected.

Participants and design

Hundred-and-ten students of East Asian background (60 women, M_{age} = 21.49, *SD* = 2.72) and 106 students of White British background (51 women, M_{age} = 21.44, *SD* = 5.61) participated in a study on emotion in voice recognition on voluntary basis. All White British participants were native speakers of English and did not speak Chinese and all East Asian students were native speakers of Chinese (mostly Cantonese) who spoke English (University level). Participants of East Asian background reported living in the U.K. for slightly less than 1.5 years on average (*M* = 17.48 months, *SD* = 20.04). No data was omitted from the analysis.

The study employed a 2 (cultural group: British vs. Chinese) by 2 (language: English vs. Chinese) by 7 (emotional display: anger, disgust, fear, happiness, neutral, sadness, pleasant surprise) mixed design. The main dependent variable was recognition accuracy.

Procedure

The procedure of the main study was identical to that of the pilot study, with the exception that this time the stimuli consisted of 28 preselected sentences expressed in one of seven emotions, resulting in 196 sentences. Participants listened to each sentence and identified the emotional display conveyed in each sentence by clicking on one of seven response options that represented the emotion types. Participants were not informed about the origin of the speakers whose voice they were about to hear in the task.

Results

The main recognition accuracy of six emotional expressions and neutral tone of voice is presented in Table 2 for each cultural group and language condition, along with associated error patterns for each emotional expression. For the British group, the overall emotional display recognition rates ranged between 48% and 91% in the English condition and between 25% and 86% in the Chinese condition. For the Chinese group, the overall emotional display recognition rates ranged between 28% and 87% in the English condition and between 39% and 98% in the Chinese condition.

Error patterns

As can be seen from Table 2, error patterns for the two groups were remarkably similar. When listening to Chinese sentences, both White British and Chinese listeners confused fear and sadness. Also, both groups most frequently misclassified surprise sentences as happy sentences and when listening to happy sentences, both groups most often misclassified these as neutral or pleasant surprise displays. Moreover, although generally well recognized, angry sentences were mistaken for disgust sentences when the wrong response alternative was chosen. Finally, for disgust sentences, no clear pattern emerged for White British listeners (all response alternatives received over 5% of responses), while Chinese listeners seemed to have recognized the valence (negative) of the stimuli but chose all negative response options with a roughly equal frequency.

As for English materials, fear and sadness were again confused though Chinese listeners also frequently chose neutral for both categories. Similar to the error pattern for Chinese sentences, happiness was most often misclassified as neutral or pleasant surprise by both groups. Again, pleasant surprise was mistaken for happiness most often.

Only responses for disgust do not show the same error pattern: For White British listeners, these sentences were relatively easy to recognize and if misclassified, pleasant surprise was the category chosen most often. Chinese listeners showed no clear pattern as neutral, surprise, angry and fear were chosen frequently. Taken together, these results suggest that both groups followed similar mechanisms when recognizing emotions from pseudo-sentences.

To infer whether primary acoustic features might have been used differently by the two groups when recognizing emotions, we entered errors made by participant into an additional discriminant analysis using mean pitch, intensity and duration as predictor variables. Sentences were grouped according to their most frequent misclassification; sentences that had an equal number of misclassifications were excluded from this analysis (29 sentences for Chinese materials and 28 for English materials). Results for Chinese sentences revealed similar patterns for Chinese and English participants in that 47.1% of mistakes could be predicted by acoustics for Chinese participants and 52.4% of misclassified sentences were identified correctly by the model for English participants. Results for English sentences were also similar although here English participants' misclassifications were predicted slightly more accurately by the model (50.8% for English participants vs. 42.2% correct for Chinese participants).[5] Overall, these findings suggest that misclassifications made by participants could be somewhat predicted by three of the main acoustic cues used to intone the sentences (i.e. pitch, intensity, and duration). Importantly, success rates of the model did not vary as a function of cultural group and/or materials tested.

Emotional prosody recognition accuracy

To account for possible stimulus or response biases, we first converted emotional prosodic display recognition accuracy rates to unbiased hit rates (H_U scores) following Wagner (1993). As recommended for proportional data, H_U scores were then arcsine-transformed before further analysis (Wagner, 1993). To examine whether individuals are more accurate recognizing emotional displays expressed by members of their own cultural/language[6] group than those expressed by members of a different cultural group, we conducted a mixed Analysis of Variance with cultural group (Chinese vs. White British) and language of materials (Chinese vs. English) as between-subjects factors and recognition accuracy (arcsine transformed H_U scores) for each emotional display (angry, disgust, fear, happiness, neutral, pleasant surprise, and sadness) as within-subjects factor. This analysis revealed a significant main effect of emotional display, F (6, 1266) = 339.03, $p < .001$, indicating that some displays were better recognized than others. Specifically, anger displays were recognized best (1.07), followed by neutral tone of voice (.91), and displays of sadness (.84), pleasant surprise (.77), fear (.65), disgust (.65), and happiness (.55; see Table 2 for unbiased hit rates). All pairwise comparisons were significant at $p < .001$, except for the difference between disgust and fear ($p = .89$). A second significant main effect emerged for language of materials, F (1, 211) = 7.05, $p < .01$, revealing that English sentences ($M = .80$, $SD = .16$) were more accurately recognized compared to Chinese sentences ($M = .75$, $SD = .16$). The main effect of cultural group was not significant, $F < 1$.

The significant two-way interaction between cultural group and language, F (1, 212) = 89.04, $p < .001$, confirmed that participants were more accurate recognizing emotional displays communicated in their native language as opposed to non-native language (see Figure 1). A simple effect analysis conducted to unfold this interaction effect revealed that British participants recognized emotional displays significantly more

accurately in English (M = .89, SD = .13) than in Chinese (M = .67, SD = .12), $p < .001$. The same pattern emerged for Chinese participants who performed significantly better when judging emotional displays in their native language (M = .84, SD = .15) than in English (M = .71, SD = .14). The difference in recognition accuracy between the two sets of materials was greater for British participants (Cohen's d = 1.60) compared to Chinese participants (Cohen's d = .90). A different reading of the data showed that Chinese participants (M = .71, SD = .14) performed slightly better than did British participants (M = .67, SD = .12) in recognizing emotional displays communicated in their non-native language (p = .07).

The analysis also revealed significant two-way interactions between emotional display and cultural group, F (6, 1266) = 5.98, $p < .001$, and emotional display and language, F (6, 1266) = 156.61, $p < .001$. These two-way interactions effects were qualified by a significant three-way interaction between emotional display, cultural group, and language, F (6, 1266) = 13.25, $p < .001$. A simple effects analysis by cultural group revealed that White British participants showed significantly higher recognition accuracy rates for English as opposed to Chinese sentences, ps < .01, except when sentences were intoned in a neutral, p = .94, and happy, p = .16, tone of voice. Similarly, Chinese participants were better at recognizing emotions from Chinese sentences as opposed to English sentences, ps < .01, except for stimuli intoned in fear, p = .44, and sad, p = .59, for which no in-group advantage was observed. A second simple effects analysis by language revealed that English sentences, regardless of the type of emotional display, were better recognized by White British participants than by Chinese participants, ps <.01. Comparably, all emotional displays were better recognized from Chinese materials by Chinese participants than by White British participants, ps < .01, except for angry tone of voice, p = .56. Taken together, these results confirm an in-group

advantage for most of the emotional categories tested in the current study.

A separate ANOVA conducted to examine sex differences with the average recognition accuracy as the dependent variable and participants' sex, cultural group, and language of materials as between-subjects variables revealed a significant main effect of sex, F (1, 207) = 4.67, $p < .05$, with women (M = .79, SD = .16) showing slightly higher recognition scores than men (M = .76, SD = .15). Sex did not interact with any of the other variables in the model, Fs < 1.

Discussion

Emotional prosody recognition: Exploring the In-group advantage

The goal of the present study was to investigate cross-cultural emotional prosody recognition in a quasi-balanced design. Overall, the findings provide support for the 'dialect theory' (Elfenbein & Ambady, 2003), which postulates that both universal and culture-specific affect programs modulate emotion recognition.

Support for the notion of universal affect programs comes from two sets of findings. First, recognition of emotional displays were generally high: participants were able to successfully infer emotional displays from pseudo-sentences communicated in their native language and a foreign language, with recognition accuracy rates three to six times higher than predicted by chance. These findings add to previous evidence demonstrating similarly high accuracy rates in tasks requiring recognition of emotional displays in languages other than one's own (e.g. Albas et al., 1976; Pell et al., 2009; Scherer et al. 2001; Thompson & Balkwill, 2006; Van Bezooijen et al. 1983).

Second, Chinese and White British listeners' response patterns show considerable similarity; we found no clear evidence that acoustic cues signaling distinct emotion displays were used differently by the two groups (c.f. Table 2). In particular,

error patterns show that fear and sadness were confused frequently. One potential reason for this confusion may be due to these expressions sharing specific primary acoustic characteristics (e.g. similar fundamental frequency and intensity; see Table 1) that are easily mistaken for one another when other cues such as semantics or facial expressions are missing. Similarly, surprise was often mislabeled as happiness, suggesting that both listener groups recognized valence (positive) similarly and acoustics were misinterpreted (e.g. high pitch). The results also show that for both groups, disgust misclassifications were most variable. Disgust was also the second-most difficult emotional display to recognize, presumably because this emotion is often expressed in short interjections (e.g. yuck) rather than in sentential context (Banse & Scherer, 1996). Alternatively, low recognition rates for disgust stimuli can be attributed to difficulties by decoders as the Chinese speaker was not well recognized when expressing disgust in the pilot study (but also see discussion below).

In line with the assumption that universal affect programs play a role during emotional prosody recognition, results also fail to provide clear evidence for the suggestions that Asians recognize angry (facial) expressions less accurately than Caucasians (Matsumoto, 1992). The only indication that collectivistic cultures (e.g. Asian) show minor culture-specific influences during angry prosody recognition comes from the fact that Asian participants recognized angry English sentences *better* than angry Chinese sentences (87% vs. 82% [1.09 vs. .98 arcsine transformed H_U scores]). This may be due to the distinctiveness of angry expressions at the acoustic level (e.g. loud tone of voice), which is recognized cross-culturally but perhaps slightly less acceptable in Asian cultures. However, given that recognition was generally high for angry materials, latter finding requires replication. We believe that the current data from vocal materials provide little support for the claim that Asian recognize anger less

successfully than Caucasians.

Similar to previous studies, one of the caveats of the current study is that it does not make it easily possible to distinguish between 'strong' and 'weak' versions of universality. For instance, Russel (1995) proposed the principle of *minimal universality* for facial expressions of emotions. Minimal universality assumes that specific patterns of muscle movements can be used by both in- and out-group members to infer something about the state (mental, physical, cognitive) of the poser. However, the inferences made by in- and out-group members do not necessarily need to coincide. For vocal emotional displays, a similar assumption can be made, i.e. inferences about someone's state can be made based on variations of speech sounds resulting from vocal-production-related physiology. Still, in- and out-group members might differ in the assumptions they make about the mental state of someone depending on cultural display rules.

Support for culture-specific affect programs comes from the findings that demonstrate a clear overall in-group advantage (Elfenbein & Ambady, 2002), such that members of each cultural group were more accurate in recognizing the emotions displayed by a member of their own cultural group than by a member of the other cultural group, contributing to previous evidence demonstrating a similar advantage when recognizing emotions from the voice (Albas et al. 1976; Pell et al., 2009; Scherer et al. 2001; Thompson & Balkwill, 2006; Van Bezooijen et al. 1983). Although Beier and Zautra (1972) hypothesized that the in-group advantage should be particularly pronounced when stimulus duration is short (such as in Sauter et al.'s (2010) study), the current results support an in-group advantage with *sentence-long* stimuli.

The current data also provide some evidence that the in-group advantage is more pronounced for some emotional displays than others. Specifically, British participants

showed an in-group advantage for all displays except for those expressed in a neutral or happy tone of voice and Chinese participants showed an in-group advantage for all displays except for those expressed in a fearful or sad tone of voice. Previous evidence has suggested that emotions are displayed through specific use of acoustic cues, i.e. emotional displays have distinct acoustic profiles. Thus, it can be hypothesized that when recognition accuracy is equally good for native and non-native speakers (e.g. sadness expressed in English), cues used by the speakers were universally decoded by listeners based on the cues' distinctiveness at the acoustic level (e.g. low tone of voice is universally used for sad displays, or wide pitch range as is common across languages for happy displays). In turn, these commonalities render it relatively easy to infer these categories (c.f. Pell, Paulmann et al., 2009) in a universal manner.

Finally, the current study provides tentative evidence for the claim that the in-group advantage is subject to learning (Elfenbein, Beaupré, Lévesque, & Hess, 2007). Specifically, we find that, compared to British participants, Chinese participants were slightly more accurate recognizing emotions displayed by a member of another cultural group. Moreover, the difference between the British participants' recognition rates of English and Chinese emotional displays was greater compared to that of Chinese participants. Together these findings suggest that Chinese participants showed a lower in-group advantage compared to British participants. This finding is consistent with the suggestion that individuals who live in a foreign country (e.g. to attend university) should show a lower in-group advantage than those who do not (e.g. Elfenbein et al., 2007). The current study does not allow to comment on what kind of learning leads to higher recognition rates. It has previously been suggested that *amount of exposure* to the out-group correlates with recognition rates. In other words, the more exposure an individual has to members of the other culture, the better their recognition of the out-

group emotional displays (e.g. Elfenbein & Ambady, 2003b). Thus, improved cross-cultural emotion recognition could result from familiarity with the out-group display rules. Our sample of Chinese participants consisted of individuals attending a British university and some of them came from locations where English is widely used (e.g., Hong Kong). A future study that includes Chinese participants with no prior exposure to English would be expected to yield a similar in-group advantage as that shown by the British participants in the current study. An ideal design of a future study fully exploring the relationship between language learning and emotional prosody recognition should include both native English and Chinese speakers with no/some familiarity of Chinese and English respectively.

In short, the present data provide evidence for the claim that cross-cultural emotional prosodic display recognition is influenced by both universal and culture-specific characteristics. While the exact nature of these culture-specific influences needs to be confirmed in future studies, the present data allude to the possibility that they stem from at least two sources. First, culture-specific emotionally relevant rules (e.g. to hide the feeling of disgust) from the native culture are transferred and (mis)applied when recognizing emotional displays from non-native stimuli. In addition, emotion irrelevant factors such as linguistic differences (e.g. unit size of tones, attention naturally paid to pitch height vs. direction) between the native and non-native language can impact cross-cultural emotion recognition (e.g. Scherer et al., 2001). Latter point would imply that familiarity with the language (and its linguistic properties) leads to increased emotional prosody recognition accuracy (or at least that those who know the language are less susceptible to interference from linguistic features during emotion recognition) as was observed in the current study.

Emotional prosody recognition: Why some emotions are better recognized than others

Comparable to previous studies, the present results confirm that some emotion displays are recognized more accurately than others. For instance, irrespective of language, angry sentences were recognized much better than sentences conveying happiness (see Table 2). It has been previously argued that differences in recognizing individual emotions may be due to biological or evolutionary factors as it may be more advantageous to recognize potential danger (anger/fear) than a non-threatening situation (e.g. happy/pleasant surprise; e.g. Öhman & Mineka, 2001; Öhman, 2002). However, note that in the facial expression literature, anger *and* happiness are usually among the emotions best recognized (c.f. meta-analysis by Elfenbein & Ambady, 2002) which has led to the assumption that emotional facial expressions displaying approach and avoidance behavior are most likely to be recognized across cultures (e.g. Baron & Boudreau, 1987; McArthur & Baron, 1983; Elfenbein & Ambady, 2002). Interestingly, however, difficulties have been documented in recognizing happiness or joy in the vocal channel (e.g. Pell et al., 2009). Specifically, it has been argued that the "smile" associated with happy faces is a non-ambiguous cue when recognizing faces cross-culturally (Scherer et al., 2001), whereas vocally expressed joy or happiness seems to be more strongly modulated by language and culture differences (Juslin & Laukka, 2003) and might not contain discernible acoustic characteristics that help listeners easily categorize this emotion (recall that the discriminant analysis categorized happy sentences least accurately). A visual inspection of the present data confirms the suggestion that happiness may be more susceptible to culture differences: while Chinese listeners were relatively good at recognizing happiness from Chinese speech (68% [.83]) they experienced difficulty when recognizing happiness from English (28% [.32]).

Interestingly, British listeners performed similar in their native language (48% [.56]) and non-native language (42% [.50]). This would suggest that while Chinese listeners recognized signals provided by the Chinese speaker, they were not as successful in identifying the cues used by the British speaker. In contrast, English listeners were equally able to decipher cues provided by the Chinese speaker and British speaker. This might mean that Chinese participants needed a cue not used by the British speaker to detect happiness. This cue (which seems to be present for stimuli expressed by the Chinese speaker), however, doesn't hinder British listeners to detect happiness from the Chinese speaker. Alternatively, Chinese listeners fail to correctly detect/interpret a cue used by the British speaker. Future studies focusing on more elaborate acoustical analyses might pinpoint crucial acoustic/suprasegmental differences between the two languages which could explain why Chinese listeners have difficulties to accurately detect happiness from British stimuli. Finally, it becomes clear from Table 2 that happy sentences were often mistaken for neutral sentences. The difficulty in recognizing happiness could thus also be due to encoding 'style' of our speakers; without semantic content complementing the expression, speakers are perceived using a neutral (perhaps 'daily-life' friendly) tone of voice.

In addition, emotion display recognition rates revealed particularly poor recognition of disgust expressions. Presumably, this effect is driven by the poor recognition of disgust displays of Chinese materials. In a study by Beaupré and Hess (2005), Asian and Caucasian participants were asked to indicate how often they think a specific emotional category is displayed in daily-life. Asian participants considered disgust expressions less probable in daily-life than Caucasian participants. The difficulty in recognizing disgust displays from Chinese materials could thus be related to such a difference in experience. If less often exposed to specific emotional expressions, it may

be more difficult to recognize them (Beaupré & Hess, 2005), and it may also be more difficult to express the emotion in the first place (recall that Chinese disgust stimuli were not very well recognized in the pilot study). These points are strengthened by the observation that Chinese participants were worse at recognizing disgust from Chinese sentences (40% [.51]) than from English sentences (56% [.74]). In line with the assumption that prevalence in daily-life may influence recognition rates, neutral tone of voice was generally well recognized (in fact, British listeners recognized neutral tone of voice very successfully from both Chinese and English materials suggesting that a neutral tone of voice which is arguably prevalent in daily-life can easily be deciphered in a universal manner).

While our results confirm that some emotional displays can be recognized more easily than others, individual emotional prosodic display recognition rates should be interpreted with caution. Similar to previous studies, the present investigation presented materials spoken from one (female) speaker from each cultural group only. This raises the question whether results can generalize to male speakers, too. Some of our previous research (e.g. Paulmann, Pell, Kotz, 2008; Pell, Paulmann et al., 2009) in which we used male and female speakers from different age groups shows similar acoustic profiles of stimuli when compared to profiles of the stimuli used in the current study. Moreover, past studies using event-related brain potentials (ERPs) (e.g. Paulmann & Kotz, 2008; Paulmann, Schmidt, Pell, Kotz, 2009) fail to find meaningful differences in emotion decoding for different speakers, that is ERP results suggest that rapid emotional evaluation of stimuli takes place irrespective of speaker voice, age or style. These findings suggest that listeners can generally infer emotions from a wide variety of speakers irrespective of the expression ability of speakers (as needs to be done in real life where a high degree of proto-typicality can not readily be expected, either). In light

of this evidence, we suggest that the use of only one speaker voice in the current study is unlikely to affect the general interpretation of results. However, given that past research suggests that speakers' expression abilities might differ for *specific* emotions (e.g. one speaker is "better" at expressing anger than disgust whereas another speaker is particularly "good" at expressing disgust), we suggest that single emotion recognition rates should always be interpreted with caution. Nevertheless, to address this potential confound, future studies should include a wider range of speakers (both male and female from different age groups) to add to the generalizability of the current findings.

Concluding Remarks

To conclude, while the majority of previous studies revealed an in-group advantage in unbalanced designs, our study provides evidence for the in-group advantage employing a quasi-balanced design, where each cultural group in the study judged emotional displays expressed by a member of their own and another cultural group. In addition to using a quasi-balanced design, we also controlled for stimulus equivalence and employed materials stripped off from their semantic content to address some methodological limitations of the existing work. Our findings add to accumulating evidence that both universal and culture-specific emotion processes impact emotional prosody recognition (e.g. Elfenbein & Ambady, 2002) by demonstrating that, while emotion recognition rates for non-native materials were above-chance, listeners are significantly better at recognizing vocal emotions expressed by speakers of their own language than by speakers of an unknown or foreign language.

References

Albas, D., McCluskey, K., & Albas, C. (1976). Perception of the emotional content of speech: A comparison of two Canadian groups. *Journal of Cross-Cultural Psychology, 7,* 481–489.

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

Baron, R. M. & Boudreau, L. (1987). An ecological perspective on integrating personality and social psychology. *Journal of Personality and Social Psychology, 53,* 1222 – 1228.

Beaupré, M. G., & Hess, U. (2005). Cross cultural emotion recognition among Canadian ethnic groups. *Journal of Cross Cultural Psychology, 36,* 355-370.

Beier, E. G., & Zautra, A. J. (1972). Identification of vocal communication of emotions across cultures. *Journal of Consulting and Clinical Psychology, 39,* 166.

Boersma, P., & Weenink, D. (2009). Praat: doing phonetics by computer [Computer program]. Version 5.1.25, retrieved from http://www.praat.org/.

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

Ekman, P. (1972). Universals and cultural difference in facial expressions of emotion. In Cole, J. (Eds.), *Nebraska Symposium on Motivation*. 1971, (Vol. 19, pp. 207-283). Lincoln: University of Nebraska Press.

Ekman, P. (1973). *Darwin and facial expression: A century of research in review*. New York: Academic Press.

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

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128,* 203–235.

Elfenbein, H. A., & Ambady, N. (2003). Universals and cultural differences in recognizing emotions. *Current Directions in Psychological Science, 12,* 159-164.

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.

Elfenbein, H. A., Beaupré, M. G., Lévesque, M. & Hess, U. (2007). Toward a dialect theory: Cultural differences in the expression and recognition of posed facial expressions. *Emotion,* 7, 131-146.

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

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

Kramer, E. (1964). Elimination of verbal cues in judgments of emotion from voice. *Journal of Abnormal and Social Psychology, 68,* 390-396.

Matsumoto, D. (1992). American-Japanese cultural differences in the recognition of universal facial expressions. *Journal of Cross-Cultural Psychology, 23,* 72-84.

Marsh, A. A., Elfenbein, H. A., & Ambady, N. (2003). Nonverbal "accents": Cultural differences in facial expressions of emotion. *Psychological Science,* 14, 373-376.

Matsumoto, D. (2002). Methodological requirements to test a possible ingroup advantage in judging emotions across cultures: Comments on Elfenbein and Ambady and evidence. *Psychological Bulletin, 128,* 236–242.

Matsumoto, D. (2006). Culture and nonverbal behavior. In V. Manusov, & M. Patterson (Eds.), *Handbook of Nonverbal Communication*. Thousand Oaks, CA: Sage

Publication.

McArthur, L. Z., & Baron, R. M. (1983). Toward an ecological theory of social perception. *Psychological Review, 90*, 215-238.

McCluskey, K. W., & Albas, D. C. (1981). Perception of the emotional content of speech by Canadian and Mexican children, adolescents, and adults. *International Journal of Psychology, 16*, 119-132.

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*, 15-21.

Mesquita, B., & Frijda, N. (1992). Cultural variations in emotions: A review. *Psychological Bulletin, 112*, 179 – 204.

Keppel, G. (1991*). Design and analysis: A researcher's handbook* (3rd edition). Englewood Cliffs NJ: Prentice-Hall. 672 p.

Öhman, A. & Mineka, S. (2001). Fear, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review, 108*, 483-522.

Öhman, A. (2002). Automaticity and the amygdala: Nonconscious responses to emotional faces. *Current Directions in Psychological Science, 11*, 62-66.

Paulmann, S., Pell, M.D., Kotz, S.A. (2008). How aging affects the recognition of emotional speech. *Brain and Language, 104, 262-269.*

Paulmann, S. & Kotz, S.A. (2008). Early emotional prosody perception based on different speaker voices. *Neuroreport*, 19, 209-213.

Paulmann, S., Schmidt, P., Pell, M.D., Kotz, S.A. (2008). Rapid processing of emotional and voice information as evidenced by ERPs. In Barbosa, P.A., Madureira, S., Reis, C. (Eds.) *Proceedings of the Conference on Speech Prosody 2008* (pp. 205-209), Campinas, Brazil.

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, A. (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behaviour, 33*, 107-120.

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.

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

Sauter, D., 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, 107*, 2408-2412.

Sauter, D. A., & Scott, S. K. (2007). More than one kind of happiness: Can we recognize vocal expressions of different positive states? *Motivation and Emotion, 31*, 192–199.

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

Scherer, K. R. (2013). Vocal markers of emotion: Comparing induction and acting elicitation. *Computer Speech & Language, 27,* 40-58.

Thompson, W., & Balkwill, L.-L. (2006). Decoding speech prosody in five languages. *Semiotica, 158,* 407–424.

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.

Wagner, H. L. (1993). On measuring performance in category judgement studies of nonverbal behavior. *Journal of Nonverbal Behavior*, *17*, 3–28.

Footnotes

[1] The use of acted speech samples has previously been challenged. However, there is clear empirical evidence that speech samples obtained from well trained actors contain very comparable acoustic features to samples obtained from non-posed situations (see e.g. Scherer, 2013 for a recent comparison of mood induced and acted materials). Given that the current study set out to compare vocal emotion recognition across languages, it was crucial to use controlled, good quality recordings which could not have been achieved if spontaneous speech had been used (also see e.g. Banse & Scherer (1996) for a discussion on the advantages/disadvantages to use acted speech samples).

[2] It was not possible to meet this selection criterion for Chinese sentences intoned in a tone conveying disgust. Thus, for this condition the best 28 sentences were included while ignoring the selection criterion.

[3] All reported discriminant analyses were successfully cross-validated in SPSS with randomly selected subsamples.

[4] When analyses are repeated using the acoustic parameter range dB, results showed a similar pattern: Results for Chinese materials revealed that for 61.2% of the sentences emotional category membership was predicted correctly: anger, 46.4%; disgust 52.4%; fear, 42.1%; happiness, 44.4%; neutral, 100%; sadness, 66.7%; pleasant surprise, 80%. Results for English sentences were similar as 70.4% of sentences were classified correctly: anger, 96.4%; disgust 64.3%; fear, 25.0%; happiness, 57.1%; neutral, 78.6%; sadness 96.4%; pleasant surprise, 75.0%.

[5] Again, similar results were found when range dB is used as predictor (instead of mean dB): For Chinese sentences, the model could predict 37.4% of mistakes made by Chinese participants while 41.3% of misclassified sentences were identified correctly by the model for English participants. Results for English sentences revealed that 43.2% of

mistakes were predicted correctly for English participants vs. 46.9% for Chinese participants.

[6] For the ease of reading, we will from now on use the term "cultural group".

[7] We present percentage accuracy in this table (rather than H_U scores) to allow for comparisons with previously reported findings using percentage accuracy. Unbiased hit rates (H_U scores) are available from the authors.