Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal emotional stimuli

Pan Liu · Marc D. Pell

Abstract To establish a valid database of vocal emotional stimuli in Mandarin Chinese, a set of Chinese pseudosentences (i.e., semantically meaningless sentences that resembled real Chinese) were produced by four native Mandarin speakers to express seven emotional meanings: anger, disgust, fear, sadness, happiness, pleasant surprise, and neutrality. These expressions were identified by a group of native Mandarin listeners in a seven-alternative forced choice task, and items reaching a recognition rate of at least three times chance performance in the seven-choice task were selected as a valid database and then subjected to acoustic analysis. The results demonstrated expected variations in both perceptual and acoustic patterns of the seven vocal emotions in Mandarin. For instance, fear, anger, sadness, and neutrality were associated with relatively high recognition, whereas happiness, disgust, and pleasant surprise were recognized less accurately. Acoustically, anger and pleasant surprise exhibited relatively high mean $f0$ values and large variation in $f0$ and amplitude; in contrast, sadness, disgust, fear, and neutrality exhibited relatively low mean $f0$ values and small amplitude variations, and happiness exhibited a moderate mean $f0$ value and $f0$ variation. Emotional expressions varied systematically in speech rate and harmonics-to-noise ratio values as well. This validated database is available to the research community and will contribute to future studies of emotional prosody for a number of purposes. To access the database, please contact pan.liu@mail.mcgill.ca.

Keywords Speech production · Emotion recognition · Prosody · Acoustic features · Cross-cultural studies

P. Liu (✉) · M. D. Pell
School of Communication Sciences and Disorders,
McGill University,
1266 avenue des Pins Ouest,
Montréal H3G 1A8, Canada
e-mail: pliu261@gmail.com

Introduction

Humans can usually understand each other's emotions and intentions in a reliable manner from isolated speech cues, in the absence of visual information, as exemplified daily through routine telephone conversations (Paulmann & Pell, 2011). In such cases, the vocal parameters of speech, including fluctuations in pitch, loudness, speech rate, and so forth, which are referred to as *emotional prosody*, play a crucial role in conveying the speakers' affective disposition and emotional state to listeners (Banse & Scherer, 1996).

In the growing literature on emotional prosody, many researchers are conducting neuropsychological (e.g., Adolphs, Damasio, & Tranel, 2002; Adolphs, Tranel, & Damasio, 2001; Pell & Leonard, 2003; Ross & Monnot, 2008) and neuroimaging (e.g., Grandjean et al., 2005; Paulmann & Kotz, 2008; Paulmann, Pell, & Kotz, 2008) experiments to investigate the neuro-cognitive mechanisms underlying emotional prosody. Other studies are exploring how speakers use prosody to *encode* discrete emotions (e.g., happiness, anger, disgust, etc.; see Ekman, 1992a, b) and how listeners effectively *decode* these cues in speech, to advance knowledge of the particular perceptual-acoustic characteristics of vocal emotion expressions (e.g., Castro & Lima, 2010; Pell, Monetta, Paulmann, & Kotz, 2009; Pell, Paulmann, Dara, Alasseri, & Kotz, 2009; Scherer, Banse, & Wallbott, 2001; Thompson & Balkwill, 2006; for a review, see Juslin & Laukka, 2003). Given that emotional prosody is fully embedded in language and that it can be influenced by the linguistic properties of a specific language (Pell, 2001), it has been necessary for many researchers to construct valid emotional stimuli appropriate for the linguistic background of the participants under study. This time-consuming but essential step is necessary if emotional prosody research is to be conducted in different language contexts, and validated recordings are necessary for

researchers to successfully control the linguistic content of their stimuli to "isolate" the effects of prosody in relation to concurrent language features (e.g., lexico-semantic cues pertaining to emotion) in each language. While valid vocal emotional stimuli have been mostly established in Indo-European languages in the literature (e.g., Banse & Scherer, 1996; Baum & Nowicki, 1998; Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005; Castro & Lima, 2010; Juslin & Laukka, 2001; Nowicki & Duke, 1994; Pell, Monetta, et al., 2009; Pell, Paulmann, et al., 2009; Scherer et al., 2001; for a review, see Juslin & Laukka, 2003), this study will establish a validated database of vocal emotional stimuli in Mandarin Chinese, a major Sino-Tibetan language spoken by more than a billion people around the world.

Emotional prosody in Mandarin has been studied in various ways by researchers, although there have been few attempts to develop a well-controlled database of vocal emotional stimuli for future work in Mandarin Chinese (cf. You, Chen, & Bu, 2005). Many previous studies used emotional utterances taken from Chinese movies or TV shows/broadcasts as stimuli (Tao, Kang, & Li, 2006; You et al., 2005; Yu, Chang, Xu, & Shum, 2001; Zhang, 2008), which limits control of the linguistic content of speech for studying vocal emotions. Other researchers recruited speakers to produce emotional utterances that had a concurrent emotional semantic context (Anolli, Wang, Mantovani, & De Toni, 2008; You et al., 2005) or to produce "semantically neutral" but emotionally inflected sentences as vocal stimuli (e.g., Li, Shao, & Dang, 2009; Pao, Yeh, & Chen, 2008; You et al., 2005; Zhang, Ching, & Kong, 2006). A potential problem of the latter approaches is that prosody cannot be studied independently of corresponding semantic cues and, also, that "semantically neutral" sentences can sometimes promote unexpected interpretations by listeners when combined with different emotional prosodic meanings (see Pell, 2006, for a discussion). For these reasons, many researchers have constructed language-like *pseudoutterances* that can be produced by speakers to vocally encode emotions in a relatively naturalistic manner (e.g., Castro & Lima, 2010; Pell, Paulmann, et al., 2009; Scherer, Banse, Wallbott, & Goldbeck, 1991), an approach that was also adopted here.

Another key methodological consideration is the process for perceptually validating vocal emotional stimuli for future use. Many current studies of Mandarin provide only sparse details about this process (You et al., 2005; Zhang, 2008; Zhang et al., 2006) or refer to subjective assessments of their emotional speech corpus (Pao et al., 2008), often involving a very small group of listeners (e.g., two or four; Yu et al., 2001; Zhang, 2008). In some cases, valid emotional stimuli were selected on the basis of whether they sounded "typical" or "effective" to listeners (Thompson & Balkwill, 2006; Zhang, 2008). To establish a database that is suitable for different investigative purposes, it will be important to avoid subjective

assessments and to mitigate the influence of individual biases in emotion recognition by collecting a more robust set of perceptual data that define vocal emotional stimuli in Mandarin. As well, since most previous studies investigated a limited number of discrete emotions (e.g., three or four; Pao et al., 2008; Tao et al., 2006; Yu et al., ; Zhang, 2008; Zhang et al., 2006), a new database should seek to provide validated stimuli representing a broader set of emotional meanings currently being studied in the literature.

The primary aim of this study was to establish a validated database of vocal emotional stimuli in Mandarin Chinese using a well-controlled validation procedure, involving seven emotion categories: anger, happiness, sadness, fear, disgust, pleasant surprise, and a nonemotional category (neutrality). Within a discrete emotions framework, these six emotions (excluding neutrality) are typically considered *basic* human emotions, each with a distinct biological basis and expressive qualities that are universally shared across cultures and languages (Ekman, 1992a, b; Ekman, Sorenson, & Friesen, 1969).[1] While an inventory of vocal expressions of the basic emotions has been established in several languages (e.g., Castro & Lima, 2010; Pell, Paulmann, et al., 2009), this has not been fully achieved in Mandarin. Our inventory should therefore be a useful tool for conducting basic research on a range of vocal emotions in Mandarin and for conducting cross-cultural/cross-linguistic studies of vocal emotion communication, and these stimuli could be incorporated into assessments of emotional and social functions involving the Mandarin-speaking population.

To facilitate the comparability of our data with those in the published literature (e.g., Castro & Lima, 2010; Pell, Paulmann, et al., 2009), the present study adopted the procedures of Pell, Paulmann, et al. (2009) in their comparative study of vocal emotion expressions in English, Arabic, German, and Hindi. As was noted earlier, an important methodological issue in the literature concerns how to control the linguistic/semantic content of speech stimuli that carry vocal cues about emotion; following Pell, Paulmann, et al. (2009), we required speakers to produce emotionally inflected *pseudosentences* (e.g., in English: *The fector jabbored the tozz*) specially constructed for Mandarin Chinese. These stimuli were composed of pseudo content words conjoined by (real) function words, rendering them semantically meaningless but ensuring that the phonetic/segmental and suprasegmental

---

[1] The status of surprise as a basic emotion remains contentious, and this emotion can also take on different valences (i.e., negative or positive) that can influence perceptual-acoustic features of surprise expressions in the vocal channel. To control for this potential variability, we focused on positively valenced surprise (i.e., *pleasant surprise*) in Mandarin to facilitate comparisons with Pell, Paulmann, et al., (2009) data on English, German, Hindi, and Arabic. Eliciting *pleasant surprise* also served to increase the ratio of positive to negative emotions in our stimulus inventory, which could be useful to future researchers interested in the valence dimension.

properties were appropriate to native Mandarin speakers/ listeners. Similar pseudoutterances have been constructed for a range of languages (e.g., English, German, Hindi, Arabic, and Portuguese; see Castro & Lima, 2010; Pell, Paulmann, et al., 2009; Scherer et al., 1991). Here, Chinese pseudosentences were elicited from four native Mandarin speakers to convey the seven target emotional meanings (anger, happiness, sadness, fear, disgust, pleasant surprise, and neutrality). The recordings were then entered into a perceptual rating study where a group of 24 native Mandarin listeners judged the emotion being expressed by each item in a seven-alternative forced choice task. On the basis of the results of the perceptual study, items that reached a critical recognition consensus rate (i.e., three times chance performance, or 42.86 %) about the emotional meaning of the stimulus were included in the validated database, and acoustic analyses were conducted on these valid items to specify the acoustic characteristics of vocal emotions in Mandarin Chinese.

Method

There were four parts to the study. First, a set of pseudosentences in Mandarin Chinese were constructed and validated to ensure that they were perceived as "language-like" to native Mandarin-speaking listeners (*sentence construction*). Second, native Mandarin speakers were recruited to produce the pseudoutterances to express seven emotional meanings (*stimulus recording*). Third, these recordings were validated by a second group of Mandarin-speaking listeners who perceptually identified the emotion expressed by each item (*perceptual validation and selection*). Finally, the perceptually validated subset of stimuli was subjected to acoustic analysis to better understand the link between perceptual and acoustic features of the stimuli (*acoustic study*). As described below, all testing took place in Montréal, Canada but was conducted entirely in Mandarin Chinese at each stage of the investigation.

Sentence construction

Forty-five pseudosentences were created by the author, P.L., who is a native Mandarin speaker from North China. They were constructed by replacing content words with random Chinese characters that were semantically meaningless within the sentence context, while maintaining function words to convey grammatical information (see the Appendix for examples). To ensure that pseudosentences were both semantically meaningless and relatively plausible as Chinese sentences, a pilot study was conducted.

Participants Ten native Mandarin speakers (5 female/5 male; mean age = 25.2 ± 2.6 years) were recruited in the pilot study.

These participants were university students from China who had learned Mandarin from birth, had lived in China until at least 18 years of age, had been away from China for less than 2 years, and spoke English as a second language. Each participant was compensated $10 CAD per hour for their participation.

Procedure The 45 pseudosentences were presented to each participant in random order on a computer screen. Participants were asked to rate the degree of "language-likeness" (i.e., the extent to which the pseudosentence resembled a real Chinese sentence) on a 5-point scale from −2 to 2, in which −2 refers to *very unlike* while 2 refers to *very like*. All the instructions were given in Mandarin.

Results A rating score was calculated for each pseudosentence by averaging the ratings across the 10 participants. A subset of 35 pseudosentences with rating scores above 0 (mean rating, .53; standard deviation [*SD*], .43) were selected as the most language-like items of the original 45 items constructed. The selected pseudosentences had a mean length of 9.09 characters/syllables (range: 7–12 characters/syllables).

Stimulus recording

Participants Four native Mandarin speakers (2 female, 2 male) with a mean age of 24.3 (±4.6) years were recruited as *encoders* to produce vocal emotion expressions in Mandarin. Participants responded to an advertisement posted at McGill University, Montréal (Canada) and were selected for having lay experiences in broadcasting or public speaking in Mandarin Chinese when they were in China (e.g., member of the campus radio station in a Chinese university). All encoders were university students from China, had learned Mandarin from birth, and had lived in China until at least 18 years of age; none had been away from China for more than 2 years. They all spoke standard Mandarin without any accent and spoke English as a second language. Participants were compensated $10 CAD per hour for their participation.

Materials The 35 pseudosentences, selected in the pilot study, were used as materials to elicit emotional expressions from the 4 encoders in seven emotion categories: anger, disgust, fear, sadness, happiness, pleasant surprise, and neutrality. In addition to pseudosentences, a separate list of 45 real semantic Chinese sentences was constructed for each emotion category and was produced by the encoders in each target emotion as well. Following Pell, Paulmann, et al. (2009), the real semantic sentences were employed as practice to help encoders to produce pseudoutterances more effectively and naturally. Data pertaining to the semantic utterances were not the object of this study, which focused

on how emotional prosody operates independently of semantic context.

Elicitation and recording procedure Each encoder was recorded separately in a sound-attenuated recording booth. Utterances conveying the seven emotional meanings were recorded in separate blocks, which varied in order across encoders. For each category, the encoder produced the semantic utterances first to express the target emotion as practice, followed by the pseudoutterances. Each of the 35 pseudoutterances was presented 1 at a time on a computer screen. Encoders were instructed to produce the pseudoutterance in the target emotion as if talking to the experimenter, in a way that was as natural as possible. During recording, the main author (P.L.) and a research assistant, who are both native Mandarin speakers, monitored the recording process and provided clues (e.g., verbal scenarios) to help the encoder produce the target emotion effectively; however, they never provided vocal examples of the target emotion to the encoders. All instructions and communications during testing were in Mandarin. Breaks were inserted in between blocks to ensure the transition between different emotions. All pseudoutterances were recorded using a Tascam digital recorder and a high-quality head-mounted microphone; the digital recordings were then transferred and saved onto a computer and were edited into individual .wav sound files for each utterance, using Praat speech analysis software (Boersma & Weenink, 2001). The average duration of the edited pseudoutterances was 1.65 ± 0.41 s, although this varied considerably by emotion type, as was expected (happiness, 1.62 s; anger, 1.40 s; sadness, 1.86 s; disgust, 2.14 s; fear, 1.47 s; surprise, 1.58 s; neutrality, 1.48 s).

Perceptual validation and selection

The edited pseudoutterances were entered into a perceptual validation study to evaluate how they were perceived by a group of native listeners. On the basis of the perceptual data, a valid subset of the utterances that reliably conveyed each target emotion could be identified.

Participants Twenty-four native listeners, or *decoders* (12 female, 12 male), with a mean age of 25.5 (±3.3) years were recruited for the perception study. Again, they were students from China who had learned Mandarin from birth, had lived in China until at least 18 years of age, had been away from China for less than 2 years, and spoke English as a second language. Each participant was compensated $10 CAD per hour for their participation.

Materials and procedure The total number of pseudoutterances produced by the speakers was 980 (35 pseudosentences × 7 emotions × 4 speakers). However, 1 item from the sadness

category of a male speaker was removed due to a recording artifact, leaving 979 items to be entered into the validation study. Following Pell, Paulmann, et al., 2009, the peak amplitude of all utterances was normalized to 75 dB to mitigate gross differences in perceived loudness for utterances recorded during different testing sessions. Using Superlab presentation software (Cedrus, U.S.), the 979 utterances were randomly combined and divided into four blocks, which were presented in two testing sessions, two blocks during each session (sessions were usually separated by a day). During the testing, each utterance was played once over headphones controlled by the Superlab program, which recorded mouse click responses. Decoders rendered two judgments following each item: First, they identified which emotion was being expressed by the speaker from a list of the seven categories presented on the computer screen; then, with the exception of items identified as "neutral," participants were immediately presented a 5-point rating scale on the screen to rate the intensity of the emotion that had been recognized (where 1 referred to *very weak* and 5 referred to *very intense*). All participants received practice trials prior to the first block during each testing session and frequent breaks during each session. All instructions and communication during the perceptual testing were conducted entirely in Mandarin.

Selection The accuracy of each decoder in identifying utterances conveying each emotion was first calculated; data for 1 male decoder was subsequently removed from the following analyses, since he selected "neutral" for the vast majority of items. On the basis of the accuracy data, a subset of perceptually robust or valid items was selected from the original 979 recordings (Pell, Paulmann, et al., 2009). To eliminate items that were poorly encoded in the recording sessions, two criteria were adopted to select valid items: (1) a recognition rate of at least 42.86 % (i.e., three times chance performance in the seven-choice emotion recognition task) for the target emotion (Castro & Lima, 2010; Pell, Paulmann, et al., 2009) and (2) recognition rates of less than 42.86 % in any other emotion categories for that item (Castro & Lima, 2010). For each emotion category, items that failed to reach these two criteria were removed from the set of "valid" utterances. Among the removed items, those with a recognition rate lower than 42.86 % on the target emotion, but with a recognition rate of at least 42.86 % on another nontarget emotion, were *regrouped* into that nontarget emotion category.

The frequency and proportion of valid items for each emotion and speaker are summarized in Table 1. The selection and regrouping of items led to the inclusion of 89 % (874/979) of the original items in the seven emotion categories (speaker C.C., 83 %; G.R.W., 90 %; N.Z., 90 %; T.F.G., 94 %). As is shown in Table 1, pleasant surprise was associated with the lowest proportion of valid items (34 %) among the seven categories and was frequently identified as sounding happy for all four speakers. As is also evident in

Table 1 Frequency of perceptually valid items observed for each emotion category for each of the four speakers

Speaker (sex)	No. of items	Emotion								
		Anger	Happiness	Sadness	Fear	Disgust	Surprise	Neutrality	All Emotions	% Valid
C.C. (female)	original	35	35	35	35	35	35	35	245	
	removed	0	5	0	0	14	25	30		
	regrouped	5(dis)	8(sur)	20(neu)	0	0	0	0		
	valid	40	38	55	35	21	10	5	204	83 %
G.R.W. (male)	original	35	35	35	35	35	35	35	245	
	removed	3	3	12	0	0	23	0		
	regrouped	0	11(sur)	0	4(sad)	0	0	1(hap)		
	valid	32	43	23	39	35	12	36	220	90 %
N.Z. (female)	original	35	35	35	35	35	35	35	245	
	removed	7	4	0	1	5	22	0		
	regrouped	0	4(ang) 11(sur)	0	0	0	0	0		
	valid	28	46	35	34	30	13	35	221	90 %
T.F.G. (male)	original	35	35	34	35	35	35	35	244	
	removed	0	3	0	0	0	22	0		
	regrouped	5(sur)	3(sur)	2(sur)	0	0	0	0		
	valid	40	35	36	35	35	13	35	229	94 %
All speakers	original	140	140	139	140	140	140	140	979	
	valid	140	162	149	143	121	48	111	874	89 %

Note. The emotion terms in parentheses refer to the categories from which the regrouped items were adopted: ang, anger; dis, disgust; fea, fear; hap, happiness; neu, neutrality; sad, sadness; sur, pleasant surprise.

Table 1, individual differences in the ability to express particular emotions were also observed among the 4 encoders.

Acoustic study

The selected subset of 874 perceptually validated items was subjected to acoustic analyses to evaluate basic acoustic patterns that differentiate vocally expressed emotions in Mandarin. On the basis of the previous literature, these analyses focused on six major acoustic parameters of vocal emotion expressions that are widely studied: mean fundamental frequency (mean $f0$, in Hertz), fundamental frequency variation ($f0$ range, in Hertz), amplitude variation (amplitude range, in decibels), mean harmonics-to-noise ratio (mean HNR, in decibels), HNR variation (SD of HNR, in decibels), and speech rate (in syllables per second). The observed values of mean $f0$, maximum $f0$, minimum $f0$, maximum amplitude, minimum amplitude, mean HNR, SD of HNR, and utterance duration for each item were obtained in Praat, allowing the six parameters of interest to be calculated prior to statistical analyses. Following Pell, Paulmann, et al. (2009), in order to correct for individual differences in a speaker's mean voice pitch, all $f0$ measures (mean, maximum, and minimum $f0$) were normalized in relation to the individual "resting frequency" of each speaker (i.e., the average minimum $f0$ value of all neutral

utterances produced by that speaker; see Pell, Paulmann, et al., 2009, for details). Measures of $f0$ range were then calculated by subtracting the normalized minimum $f0$ values from the normalized maximum $f0$ values. For both normalized values of mean $f0$ and $f0$ range, a value of 1 for an utterance represents a 100 % increase in the speaker's resting frequency, which, as a proportional value, could be compared across speakers. Similarly, in order to correct for individual differences in a speaker's intensity, the amplitude values (maximum amplitude, minimum amplitude) of each speaker were normalized in relation to the average minimum amplitude value of all neutral utterances produced by that speaker. Then the measures of amplitude range were calculated by subtracting the normalized minimum amplitude values from the normalized maximum amplitude values. Measures of speech rate were calculated by dividing the number of syllables of each utterance by the corresponding utterance duration, in syllables per second.[2]

---

[2] Mean amplitude was not included as an acoustic parameter, because these measurements would not necessarily reflect the natural mean values of the speakers' intensity, due to variations in recording volume across multiple recording sessions and due to further alterations in peak volume necessary for the perceptual study. Note that measures of speech rate, mean HNR, and SD of HNR were not normalized, because these ratios could already be compared directly across speakers and recording sessions.

Results and discussion

Perceptual results

Recognition rates On the basis of the selected subset of 874 items, emotion recognition rates were calculated for each emotion category across speakers. The *hit rate* was calculated as an uncorrected measure of target category recognition (percent correct), according to which neutrality (86 %) was recognized best, followed by anger (82 %), sadness (81 %), fear (80 %), happiness (70 %), disgust (67 %), and finally, pleasant surprise (56 %). These data were then converted to *Hu* scores (Wagner, 1993) to correct for differences in item frequency among categories and individual-participant response biases (i.e., relative use of specific response alternatives). *Hu* scores were calculated for each participant as the joint probability that a presented item was correctly recognized and that the corresponding target category was correctly used.

To evaluate whether the seven emotion categories could be differentiated perceptually, a one-way ANOVA performed on the *Hu* scores as a function of emotion category showed a significant effect of emotion category, $F(6, 154) = 22.95$, $p < .0001$. Post hoc (Tukey's) comparisons revealed that fear (.673) had the highest recognition among the seven categories, followed by neutrality (.667), sadness (.643), anger (.643), and disgust (.573); this was followed by happiness (.541), which was significantly lower than fear (.673; $p < .05$). Pleasant surprise (.253) was recognized significantly less accurately than the other six categories (ps $< .01$; see Fig. 1 for an illustration).

On the basis of the analysis of unbiased accuracy rates (i.e., *Hu* scores; Wagner, 1993), there was expected variation in the recognition of discrete emotions from vocal cues in Mandarin. Specifically, recognition was most accurate for fear, neutrality, sadness, and anger, compatible with previous data in other languages (Banse & Scherer, 1996; Castro & Lima, 2010; Pell, Paulmann, et al., 2009; Thompson & Balkwill, 2006). A general advantage for recognizing *negative* emotions
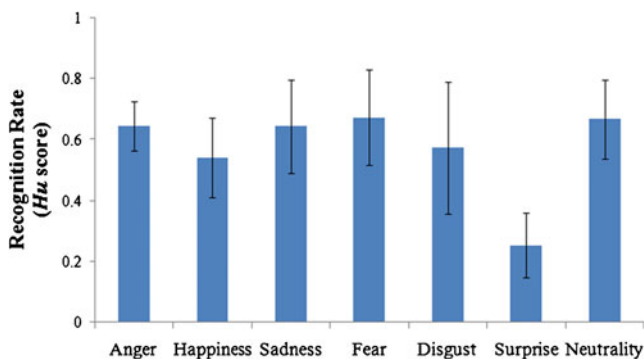
from vocal speech cues, independently of language, is compatible with evolutionary views that vocal signals associated with threat must be highly salient to ensure human survival (Ohman, Flykt, & Esteves, 2001; Tooby & Cosmides, 1990). Disgust and happiness were recognized relatively poorly in Mandarin, which is also compatible with previous findings (e.g., Banse & Scherer, 1996; Castro & Lima, 2010; Pell, 2002; Pell, Paulmann, et al., 2009; Scherer et al., 1991). It seems likely that happiness and disgust are expressed more saliently in other communication channels, such as through facial cues (Ekman, 1994; Elfenbein & Ambady, 2002; Russell, 1994; Wallbott, 1988), yielding lower accuracy rates when these two emotions must be recognized from isolated vocal cues (Elfenbein & Ambady, 2002; Pell & Kotz, 2011; Pell, Paulmann, et al., 2009). In the case of disgust, this emotion may be conveyed predominantly by nonverbal vocalizations, rather than by vocal inflections of the whole utterance (e.g., Castro & Lima, 2010; Paulmann & Pell, 2011; Scherer et al., 1991). As was expected, pleasant surprise turned out to be the most difficult emotion to identify in Mandarin, with the lowest recognition rate (see also Pell, Paulmann, et al., 2009). It is widely held that surprise is especially difficult to produce and to recognize naturally in a simulated experimental context such as the one employed here; moreover, low accuracy rates for pleasant surprise observed here and by Pell, Paulmann, et al. (2009) may be due to the positive valence of these expressions, which were frequently identified in the perceptual study as intense forms of "happiness," rather than surprise (see Abelin & Allwood, 2000; Castro & Lima, 2010; Montero et al., 1999; Navas, Hernáez, Castelruiz, & Luengo, 2004).

Intensity ratings The intensity ratings assigned to the 763 utterances recognized as conveying an emotion (rather than neutrality) were calculated for each emotion category across speakers and were submitted to a one-way ANOVA with six levels of emotion (anger, disgust, fear, sadness, happiness, pleasant surprise). The ANOVA revealed a significant effect of emotion category on intensity ratings, $F(5, 132) = 3.72$, $p < .005$. Post hoc (Tukey's) elaboration of the emotion effect indicated that pleasant surprise (3.57 ± 0.61) and anger (3.41 ± 0.52) were rated relatively high on this scale, followed by fear (3.28 ± 0.53), disgust (3.26 ± 0.67), and sadness (3.20 ± 0.51); however, none of these differences reached statistical significance. Happiness (2.86 ± 0.66) was rated as the least intense emotion, differing significantly from both surprise (3.57 ± 0.61) and anger $(3.41 \pm 0.52; p$s $< .05)$.

Fig. 1 Mean recognition rate (*Hu* score) for seven vocal expressions of emotion in Mandarin Chinese (across speakers; error bars indicate the standard deviations)

Acoustic results

The six acoustic measures of the 874 valid expressions are presented in Table 2 for each emotion category, averaged

Table 2 Acoustic measures (normalized) for each emotion category averaged across speakers

Acoustic measures	Emotion						
	Anger	Happiness	Sadness	Fear	Disgust	Surprise	Neutrality
$f0$ mean (Hz)	1.43	1.32	0.65	0.99	0.73	1.69	0.47
$f0$ variation (range in Hz)	1.86	1.85	1.03	1.22	1.72	2.06	1.09
Amplitude variation (range in dB)	1.08	1.04	0.99	0.99	1.07	1.14	0.94
HNR mean (dB)	9.62	11.5	13.5	11.18	9.75	11.52	10.6
HNR variation (SD in dB)	5.75	5.96	7.05	6.75	6.62	6.39	6.26
Speech rate (syllables/s)	6.67	5.65	4.95	6.25	4.48	5.9	6.16

across speakers. To explore how vocal expressions of the seven emotion categories differed along these dimensions in Mandarin, a one-way MANOVA was performed on the acoustic data as a function of emotion category, with the six acoustic parameters (normalized mean $f0$, normalized $f0$ range, normalized amplitude range, mean HNR, SD of HNR, and speech rate) serving as the dependent variables. The MANOVA indicated that the effect of emotion category on the six acoustic parameters was significant, Wilk's $\Lambda = 0.205$, $F(36, 3788) = 45.71$, $p < .0001$. Subsequent univariate analyses revealed that the effect of emotion category was significant for normalized mean $f0$, $F(6, 867) = 106.22$, $p < .0001$, normalized $f0$ range, $F(6, 867) = 62.78$, $p < .0001$, normalized amplitude range, $F(6, 867) = 8.79$, $p < .0001$, mean HNR, $F(6, 867) = 44.33$, $p < .0001$, SD of HNR, $F(6, 867) = 30.61$, $p < .0001$, and speech rate, $F(6, 867) = 108.27$, $p < .0001$.

Post hoc (Tukey's) comparisons were carried out on each acoustic parameter to examine the differences among emotion categories. The observed acoustic differences demonstrated a number of consistencies with those reported for other languages (e.g., Castro & Lima, 2010; Jaywant & Pell, 2012; Juslin & Laukka, 2001; Pell, Paulmann, et al., 2009; Scherer, London, & Wolf, 1973; Thompson & Balkwill, 2006). For instance, sadness was expressed in a low mean $f0$, a small $f0$ and amplitude variation, a high mean HNR, and a slow speech rate. Disgust exhibited a low mean $f0$, a low mean HNR, a large HNR variation, and the slowest speech rate. Pleasant surprise exhibited the highest mean $f0$, the largest $f0$ variation, and the largest amplitude variation. Neutrality was conveyed by a relatively low mean $f0$, a small $f0$ variation, the smallest amplitude variation, and a moderate speech rate. These cross-language consistencies are compatible with the idea that emotional communication is constrained to a large extent by biological factors and is shared across cultures (Scherer, 1986). On the other hand, greater cross-language variability was observed in the expression of fear, anger, and happiness, which might be due to methodological variability in this literature (e.g., the length of the emotional stimuli that were used, different subtypes of the target emotion that were elicited in different languages, etc.). As well, due to the small number of

speakers who produced the emotional stimuli, it is certain that individual variability in emotion expression is partly responsible for many of the observed acoustic differences between speakers and across languages.

To briefly evaluate how well the six acoustic parameters predicted the perceptual classification of items conveying each of the seven emotional meanings, a stepwise discriminant analysis was performed. The analysis revealed six significant canonical functions. Function 1, $F(36, 3788) = 45.71$, $p < .0001$, accounted for 47.3 % of the variance and correlated positively with mean $f0$ ($r = .81$), speech rate ($r = .59$), and $f0$ variation ($r = .52$). Function 2, $F(30, 3454) = 54.23$, $p < .0001$, explained 33.8 % of the remaining variance and correlated positively with speech rate ($r = .73$) and negatively with $f0$ variation ($r = -.47$). Function 3, $F(24, 3015) = 64.95$, $p < .0001$, accounted for 14.4 % of the remaining variance and correlated positively with mean HNR ($r = .82$) and mean $f0$ ($r = .50$). Finally, function 4, $F(18, 2447) = 80.59$, $p < .0001$, function 5, $F(12, 1732) = 99.18$, $p < .0001$, and function 6, $F(6, 867) = 108.27$, $p < .0001$, accounted for relatively lower percentages of the remaining variance (4.0 %, 0.4 %, and 0.1 %, respectively) and correlated positively with HNR variation ($r = .73$), amplitude variation ($r = .83$), and $f0$ variation ($r = .67$), respectively.

This model of six acoustic measures correctly predicted the classification of validated items into the seven emotion categories at an overall rate of 59.4 % (519/874).[3] These results underscore that the six selected acoustic measures are essential cues in communicating vocal emotions (Bachorowski & Owren, 1995; Castro & Lima, 2010; Mozziconacci, 2001; Pell, Paulmann, et al., 2009; Thompson & Balkwill, 2006; Williams & Stevens, 1972). However, there were major differences among the specific emotion categories in how well they

---

[3] To allow comparisons with Pell, Paulmann, et al. (2009), the discriminant analysis was rerun including only three acoustic parameters: $f0$ mean, $f0$ variation, and speech rate. The analysis yielded three significant canonical functions that correctly predicted the seven emotion categories in Mandarin at an overall rate of 49.9 %. These classification results are relatively similar to those reported by Pell, Paulmann, et al. for English (58 %), German (49 %), Hindi (56 %), and Arabic (53 %) when the same analysis was performed.

were classified on the basis of the acoustic data: disgust = 76 % (92/121), neutrality = 75.7 % (84/111), sadness = 65.1 % (97/149), pleasant surprise = 54.2 % (26/48), anger = 53.6 % (75/140), fear = 51 % (73/143), and happiness = 44.4 % (72/162). Furthermore, since approximately 40 % of the items could not be classified by the six acoustic measures, future work will need to include additional parameters to fully capture how listeners use acoustic cues to recognize emotions from speech prosody.[4]

Conclusion

The present study sought to establish a well-controlled, validated database of vocal emotional stimuli in Mandarin Chinese for use in future research. Four native Mandarin speakers produced Chinese pseudoutterances in seven emotional categories (anger, happiness, sadness, fear, disgust, pleasant surprise, and neutrality). Their expressions were validated by a new group of 24 native listeners in an emotion recognition task, on the basis of which a valid subset of items was selected and subjected to acoustic analysis. Expected variations were observed among the seven emotion categories in both perceptual and acoustic patterns.

There are several methodological limitations of this study. The first problem is that only a small number of encoders (i.e., 4) were recruited to produce the emotional expressions; these participants were not professional actors but had lay experiences in broadcasting and/or public speaking. The fact that our encoders did not have professional training, as is true of most related studies, could have led to greater variability in the ability of individual speakers to produce vocal emotion expressions, influencing the perceptual and acoustic measures to a certain degree. Another potential shortcoming is that only 24 decoders were recruited for the perceptual validation task; while this sample size is typical of comparable studies in the emotion perception literature (e.g., Burkhardt et al.,

2005; Pell, Paulmann, et al., 2009; Thompson & Balkwill, 2006; for a review, see Juslin & Laukka, 2003) and likely robust, a larger sample of decoders could potentially improve the reliability of our data. Finally, it is clear that only a small number of acoustic measures were employed in the acoustic analyses presented and that many additional parameters would be needed to characterize all relevant features of emotional speech in Mandarin Chinese (or any other language). Future studies that include a larger set of acoustic measures will prove important to elaborate current evidence of the acoustic characteristics of Chinese emotional speech.

Despite these limitations, the selected vocal emotion expressions in this study were perceptually validated and exhibited systematic acoustic patterns that were similar to those found in other languages to a certain extent (e.g, Castro & Lima, 2010; Pell & Kotz, 2011; Pell, Paulmann, et al., 2009; Scherer et al., 2001). Therefore, this database, which contains 874 items conveying the seven emotional meanings, could be a valid and useful tool for future research and is currently available to the research community. It will contribute to future behavioral, neuropsychological, and neuroimaging studies on vocal emotions in Mandarin and will also contribute to future cross-cultural/cross-linguistic studies of vocal emotion communication that shed light on both the "universal" and unique features of this communication subsystem. In addition, our new stimuli could be incorporated into assessments of emotional functions involving the Mandarin-speaking population (e.g., to evaluate emotion communication functions in Mandarin-speaking brain-damaged patients). To access the database and the relevant information, please contact pan.liu@mail.mcgill.ca.

---

[4] To ensure that our criterion for including valid items in our database (i.e., minimum 42.86 % recognition) was not too liberal, a new subset of 724 items reaching a higher recognition rate of 60 % was selected and subjected to the same set of analyses. In general, similar results were found to those described in the text for the larger subset of 874 items. The new MANOVA demonstrated a significant effect of emotion category on the six acoustic measures (*p*s < .0001); for each measure, post hoc comparisons revealed a pattern of pairwise differences similar to that found in the subset of 874 items. Discriminant analysis indicated that the six acoustic parameters predicted the perceptual classification of the items at an overall rate of 60.8 % (440/724), which is almost identical to that observed for the original subset (59.4 %, 519/874). It can be concluded that our approach of describing the acoustic and perceptual qualities of our 874 "valid" items on the basis of a minimum recognition rate of 42.86 % (i.e., three times chance performance in the seven-alternative forced choice task) yields stable and replicable data on how vocal emotions are communicated in Mandarin Chinese.

Appendix

List of Chinese pseudosentences included in the database
1. 她在一个门文上走亮
2. 我们在两上投了一个绳
3. 他们拉摇了我的绳雨
4. 我马上就投量你
5. 他们抄组了一个明平的文柱
6. 我扭了一个非常圆良的春某
7. 我在腔红之前吐店了
8. 他把雪皮提在广田上
9. 我扶打过这个皮魂
10. 我将在六馆之后打关
11. 他在天某里揣写着屯八

12. 她在地泉里提了头山
13. 他把我的铺瓶缠过了
14. 他们让我拦了一个夫纸
15. 我被他们游讨了
16. 他们刨提了我的偶连
17. 他们在楼谷中投玩
18. 我从路瓶里搭乱了
19. 她堵捧了一个白丽的平本
20. 我们在签木中托亮了
21. 我的班桃被拢右了
22. 她昨天洗安了门顿
23. 他没有挑跑我们
24. 他在我们的窗水里抚摇了
25. 他在地车上拔冲
26. 我在果体里翻移了桌风
27. 他从场东里打了一个头云
28. 简单里有一个组点
29. 他在提哄着一个单雪
30. 我拐收了一个豆腿
31. 我在辉轩里拢饭了
32. 她在格边里赏了一个天阳
33. 我把这个空书摇了十个
34. 我在月点里攒了个头文
35. 他在阔线里踢了米

References

Abelin, Å., & Allwood, J. (2000). Cross linguistic interpretation of emotional prosody. In *Proceedings of the ISCA Workshop on Speech and Emotion* (pp. 110–113). Newcastle, Northern Ireland.

Adolphs, R., Damasio, H., & Tranel, D. (2002). Neural systems for recognition of emotional prosody: A 3-D lesion study. *Emotion, 2,* 23–51.

Adolphs, R., Tranel, D., & Damasio, H. (2001). Emotion recognition from faces and prosody following temporal lobectomy. *Neuropsychology, 15,* 396–404.

Anolli, L., Wang, L., Mantovani, F., & De Toni, A. (2008). The voice of emotion in Chinese and Italian young adults. *Journal of Cross-Cultural Psychology, 39,* 565–598.

Bachorowski, J. A., & Owren, M. J. (1995). Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological Science, 6,* 219–224.

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

Baum, K. M., & Nowicki, S., Jr. (1998). Perception of emotion: Measuring decoding accuracy of adult prosodic cues varying in intensity. *Journal of Nonverbal Behavior, 22,* 89–107.

Boersma, P., & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5,* 341–345.

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

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

Ekman, P. (1992a). Are there basic emotions? *Psychological Review, 99,* 550–553.

Ekman, P. (1992b). An argument for basic emotions. *Cognition and Emotion, 6,* 169–200.

Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin, 115,* 268–287.

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science, 164,* 86–88.

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

Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M. L., Scherer, K. R., & Vuilleumier, P. (2005). The voices of wrath: Brain responses to angry prosody in meaningless speech. *Nature Neuroscience, 8,* 145–146.

Jaywant, A., & Pell, M. D. (2012). Categorical processing of negative emotions from speech prosody. *Speech Communication, 54,* 1–10.

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.

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

Li, A., Shao, P., & Dang, J. (2009). Cross-cultural and multi-modal investigation of emotion expression. *Journal of Tsinghua University, 49*(Suppl. 1), 1393–1401.

Montero, J. M., Gutiérrez-Arriola, J. M., Colás, J., Guarasa, J. M., Enríquez, E., & Pardo, J. M. (1999). Development of an emotional speech synthesiser in Spanish. In *Proceedings of Eurospeech'99* (pp. 2099–2102). Budapest, Hungary.

Mozziconacci, S. J. L. (2001). Modeling emotion and attitude in speech by means of perceptually based parameter values. *User Modelling and User-Adapted Interaction, 11,* 297–326.

Navas, E., Hernáez, I., Castelruiz, A., & Luengo, I. (2004). Obtaining and evaluating an emotional database for prosody modelling in standard Basque. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science), 3206,* 393–400.

Nowicki, S., Jr., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The diagnostic analysis of nonverbal accuracy scale. *Journal of Nonverbal Behavior, 18,* 9–35.

Ohman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology. General, 130,* 466–478.

Pao, T., Yeh, J., & Chen, Y. (2008). Emotion recognition via continuous Mandarin speech. In S. Pinder (Ed.), *Advances in human computer interaction* (pp. 365–384). InTech.

Paulmann, S., & Kotz, S. A. (2008). An ERP investigation on the temporal dynamics of emotional prosody and emotional semantics in pseudo- and lexical-sentence context. *Brain and Language, 105,* 59–69.

Paulmann, S., & Pell, M. D. (2011). Is there an advantage for recognizing multi-modal emotional stimuli? *Motivation and Emotion, 35,* 192–201.

Paulmann, S., Pell, M. D., & Kotz, S. A. (2008). Functional contributions of the basal ganglia to emotional prosody: Evidence from ERPs. *Brain Research, 1217,* 171–178.

Pell, M. D. (2001). Influence of emotion and focus location on prosody in matched statements and questions. *Journal of the Acoustical Society of America, 109,* 1668–1680.

Pell, M. D. (2002). Evaluation of nonverbal emotion in face and voice: Some preliminary findings on a new battery of tests. *Brain and Cognition, 48,* 499–504.

Pell, M. D. (2006). Cerebral mechanisms for understanding emotional prosody in speech. *Brain and Language, 96,* 221–234.

Pell, M. D., & Kotz, S. A. (2011). On the time course of vocal emotion recognition. *PLoS One, 6,* e27256.

Pell, M. D., & Leonard, C. L. (2003). Processing emotional tone from speech in Parkinson's disease: A role for the basal ganglia. *Cognitive, Affective, & Behavioral Neuroscience, 3,* 275–288.

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

Ross, E. D., & Monnot, M. (2008). Neurology of affective prosody and its functional- anatomic organization in right hemisphere. *Brain and Language, 104,* 51–74.

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.

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

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

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

Scherer, K. R., London, H., & Wolf, J. J. (1973). The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality, 7,* 31–44.

Tao, J., Kang, Y., & Li, A. (2006). Prosody conversion from neutral speech to emotional speech. *IEEE Transactions on Audio, Speech and Language Processing, 14,* 1145–1153.

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

Tooby, J., & Cosmides, L. (1990). The past explains the present: Emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology, 11,* 375–424.

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

Wallbott, H. G. (1988). Big girls don't frown, big boys don't cry—Gender differences of professional actors in communicating emotion via facial expression. *Journal of Nonverbal Behavior, 12,* 98–106.

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

You, M., Chen, C., & Bu, J. (2005). CHAD: A Chinese affective database. In *Proceedings of 1st International Conference on Affective Computing and Intelligent Interaction* (pp. 542–549). Beijing, China.

Yu, F., Chang, E., Xu, Y., & Shum, H. (2001). Emotion detection from speech to enrich multimedia content. In *Proceedings of 2nd IEEE Pacific-Rim Conference on Multimedia* (pp. 550–557). Beijing, China.

Zhang, S. (2008). Emotion recognition in Chinese natural speech by combining prosody and voice quality features. *Lecture Notes in Computer Science, 5264,* 457–464.

Zhang, S., Ching, P. C., & Kong, F. (2006). Acoustic analysis of emotional speech in Mandarin Chinese. In *International Symposium on Chinese Spoken Language Processing* (pp. 57–66). Kent Ridge, Singapore.