

Decoding speech prosody in five languages

WILLIAM FORDE THOMPSON and L-L. BALKWILL

Abstract

Twenty English-speaking listeners judged the emotive intent of utterances spoken by male and female speakers of English, German, Chinese, Japanese, and Tagalog. The verbal content of utterances was neutral but prosodic elements conveyed each of four emotions: joy, anger, sadness, and fear. Identification accuracy was above chance performance levels for all emotions in all languages. Across languages, sadness and anger were more accurately recognized than joy and fear. Listeners showed an in-group advantage for decoding emotional prosody, with highest recognition rates for English utterances and lowest recognition rates for Japanese and Chinese utterances. Acoustic properties of stimuli were correlated with the intended emotion expressed. Our results support the view that emotional prosody is decoded by a combination of universal and culture-specific cues.

1. Decoding speech prosody in five languages

Prosody refers to vocal qualities of speech, as distinct from verbal content, and includes intonation, stress, and timing. It signals points of emphasis, indicates a statement or question, and conveys emotional connotations (Darwin 1872; Frick 1985; Juslin and Laukka 2003). Prosodic communication of emotion can occur independently of verbal comprehension (Kitayama and Ishii 2002) and similar prosodic cues are used across cultures to convey emotions (Bolinger 1978). Evolutionary theorists have proposed that prelinguistic humans relied on prosody to communicate their intentions (e.g., mating, defense, aggression), and that this feature of speech has survived because of its evolutionary adaptive properties (Brown 2000).

Speakers communicate emotions through a combination of vocal qualities. Joy is expressed with a comparatively rapid speaking rate, high

average pitch, large pitch range, and bright timbre; sadness is conveyed with a slow speaking rate, low average pitch, narrow pitch range, and low intensity; anger is expressed with a fast speaking rate, high average pitch, wide pitch range, high intensity, and rising pitch contours; and fear is conveyed with a fast speaking rate, high average pitch, large pitch variability, and varied loudness (Scherer 1986). These emotional cues are well documented for English speech, but research is needed to evaluate whether similar vocal qualities are associated with emotion in other languages.

Most studies of emotional prosody have involved asking Western listeners to identify the intended emotion in samples of a Western language (e.g., English, Dutch, etc.). Results indicate that prosodic cues alone allow listeners to identify the emotion being conveyed, with identification rates of roughly four to five times that expected by guessing (Banse and Scherer 1996; Frick 1985; Scherer 1979, 1986; Standke 1992; van Bezooijen et al. 1983). Not all emotions are decoded equally well from speech prosody, however. Anger and sadness are typically decoded more reliably than joy and fear (Banse and Scherer 1996; Johnstone and Scherer 2000; Pittam and Scherer 1993).

Many theories of emotive speech assume that there are both universal and culture-specific prosodic cues to emotion (e.g. Matsumoto 1989; Mesquita and Fridja 1992; Scherer 1997), but refinement of such theories requires extensive cross-cultural data. Cross-cultural studies of emotion in speech are rare relative to the number of cross-cultural studies of emotion in facial expression (e.g. Biehl et al. 1997; Ekman 1972; Ekman et al. 1987; Izard 1971, 1994). In the case of facial expression, Russell (1994) argued that whereas the communication of emotional nuances depends on cultural display rules, broad characteristics of emotions (i.e., valence and arousal) are communicated according to universal principles.

Elfenbein and Ambady (2002) conducted a meta-analysis of 97 experiments exploring cross-cultural recognition of emotion in visual and auditory modes. Emotional stimuli (speech, facial expression) from one culture were presented to members and non-members of that culture. Cross-cultural emotion recognition accuracy was lower in studies of prosody than in studies of facial expression and body language, although most studies of prosody reported better-than-chance rates of recognition.

Few cross-cultural studies of emotional prosody have involved non-Western languages. Examination of non-Western languages is important because there is a long history of close contact between speakers of Western languages that might explain similar uses of prosody. Similar uses of prosody in Swedish and Norwegian, for example, can be explained by the history of interaction between speakers of these languages and would not

implicate universal principles of emotional prosody. Similar uses of prosody in Western and non-Western languages, on the other hand, would provide compelling evidence for universal principles of emotional prosody.

Elfenbien and Ambady (2003) proposed a cultural proximity hypothesis to predict how well people of different cultures recognize emotional expression. According to their hypothesis, members of cultures who share cultural elements such as degree of *individualism* or *collectivism*, power structure, and gender roles, should be more successful at decoding each other's emotional expressions than members of cultures that are less similar. The cultural proximity hypothesis predicts, for example, that Japanese people should be better at recognizing the emotional expressions of Chinese people than the emotional expressions of Americans, because Japanese and Chinese cultures are more similar to each other on relevant dimensions than Japanese and American cultures. Conversely, it predicts that English-speaking listeners should find it relatively difficult to decode emotional prosody in Japanese and Chinese speech.

The aim of this investigation was to compare the ability of Western listeners to decode speech prosody in a range of Western and non-Western languages. We investigated the ability of native speakers of English to identify the intended emotion in speech samples provided by native speakers of English, German, Chinese, Tagalog and Japanese. Speakers of each language uttered semantically neutral sentences in a way that communicated through prosodic cues each of four intended emotional expressions. Through pilot work we determined that the use of prosody was not exaggerated or dramatic but, rather, was typical for speakers of each language. The five languages represent two members of the Germanic branch of the Indo-European language family (English and German), one language from the Sino-Tibetan family (Mandarin-Chinese), one language from the Altaic family (Japanese), and one from the Austronesian family (Tagalog) (Katsiavriades and Qureshi 2003). German and English are Western languages spoken in individualistic societies, whereas Japanese and Chinese are Asian languages spoken in collectivist societies. Tagalog has been influenced by both the colonizing languages of Spanish and English as well as by the languages of its Asian neighbors, Japan, Korea and China. After so many years of diverse influences the culture of Tagalog speakers has been described as '... a blend of Asian, Islamic, and Amer-European cultures.' (Sundita 2003)

Our review of the literature motivated three predictions. First, assuming that sensitivity to emotional prosody is partially dependent on universal principles, we predicted that listeners would be able to decode emotional prosody in any language at rates higher than that predicted by

chance. Second, based on the idea that certain prosodic cues of emotion are culture specific, we predicted that listeners would have higher rates of recognition for emotional prosody in their own language (Albas et al. 1976; Kitayama and Ishii 2002; Scherer et al. 2001). Third, we predicted that listeners would be able to decode anger and sadness more reliably than joy and fear in all five languages (Banse and Scherer 1996; Johnstone and Scherer 2000; Pittam and Scherer 1993).

2. Method

Judges. 20 English-speaking judges (8 men, 12 women, mean age = 21.95) were recruited from the staff and students of York University (Toronto, Canada). Participants had minimal or no fluency with Chinese, German, Japanese, or Tagalog.

Materials. Laura-Lee Balkwill recorded all stimuli while staying in Sapporo, Japan as a visiting researcher at Hokkaido University from April 2000 to April 2001. Volunteers from Japan, China, the Philippines, Germany, and Canada (all of whom were living or studying in Sapporo, Japan) were recorded while speaking two sentences in four modes of expression: joyful, sad, angry, and fearful. Table 1 provides the closest word equivalent in each language for each of the target emotions. The sentences were two neutral statements (e.g., the bottle is on the table; the leaves are changing color). Speakers were given time prior to the recording session to translate the sentences into their first language and to rehearse each mode of expression.

Speakers were seated in a sound-attenuated booth in front of a Sony DCR-TRV900 digital video camera with an externally mounted Sony Electret Condenser Microphone (ECM-S959C). Each speaker was recorded separately in sessions lasting approximately 20 minutes. Each recording was transferred to a Macintosh PowerBook computer via the

Table 1. *Translation of emotion words*

Language	Joyful	Sad	Angry	Fearful
Chinese Male	xing fu	bei shang	fen nu	kong ju
Chinese Female	kai xin	shang xin	sheng qi	hai pa
German	fröhlich	traurig	ärgerlich	ängstlich
Japanese	shiawase	kanashii	ikari	kyoofu
Tagalog	masaya	malunglot	galit	takot

Adobe Premier software program. Audio files of each sentence in each expressive mode were extracted and then edited in the software program Sound Edit for unnecessary space before and after sentences.

The recording sessions resulted in up to six recordings of each sentence by speakers of each language. Pilot work was conducted in order to select the speech samples to be used in the experiment. For each language, two native speakers judged the speech samples recorded in terms of its suitability for use in the experiment. The criterion for selecting speech samples was that they should be exemplary of how prosody is typically used in each language. That is, we were not interested in exaggerated or dramatic uses of prosody, but in *typical* uses of prosody. This procedure resulted in a final stimulus set that consisted of 2 sentences \times 4 emotions \times 2 speakers \times 5 languages for a total of 80 sentences. The sentences were imported into an automated forced-choice selection experiment that was created in Hypercard for the purpose of the investigation.

Procedure. Each judge was tested separately, beginning with a short demographic questionnaire (age, gender, and education). They were seated in front of a Macintosh G3 Powerbook computer inside a sound-attenuated booth. The following instructions were presented on the computer monitor in their native language:

In this experiment you will hear various short phrases, in different languages. Your task is to label the emotionality of the phrase, between the four alternatives of joyful, sad, angry, or fearful. You must choose one of these alternatives only.

Judges were given an opportunity to practice the forced-choice task and to seek clarification. Practice trials were selected randomly from the set of experimental trials. Judges typically made 3–4 practice judgments and received no feedback. During the practice and testing sessions they wore a set of Sennheisser HD-480 headphones, adjusted to a comfortable loudness level (approximately 60–70 dB). Presentations were triggered and responses made using the computer mouse. The 80 speech samples were presented in random order. Each testing session took approximately 15 minutes.

3. Results

Our first prediction was that listeners would be able to recognize emotion at rates above chance (25%) in each of the five languages. Table 2 shows the mean recognition accuracy rates collapsed across intended emotions. Overall recognition accuracy was above chance performance levels for utterances in all five languages.

Table 2. *Mean recognition accuracy of overall emotion*

Language set (n's = 16)	Mean accuracy	Standard error
English	.944	.020
German	.675	.026
Chinese	.594	.027
Tagalog	.722	.020
Japanese	.541	.018

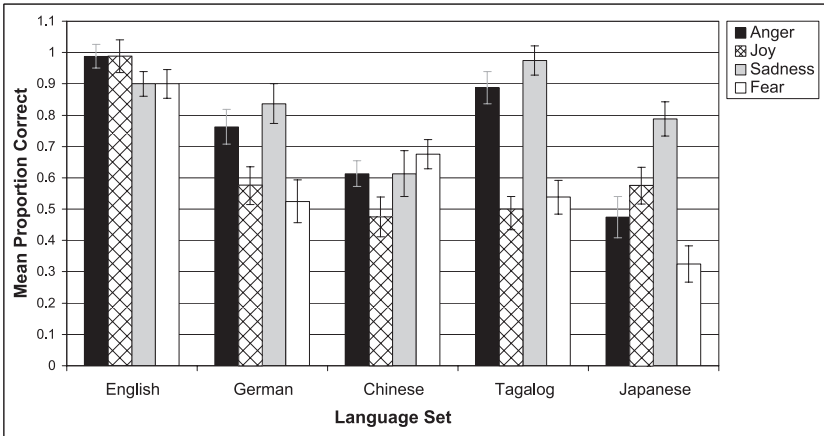


Figure 1. *The bars represent the mean percent correct identification of each intended emotion for each of the five languages. The rate of guessing correctly by chance was 25 percent. Listeners recognized all four emotions at rates that were significantly higher than chance performance in all five languages.*

Figure 1 shows mean recognition rates for each emotion in each language. Listeners recognized all four emotions at rates that were significantly higher than chance performance in all five languages. Responses were entered into a two-way ANOVA with language (English, German, Chinese, Tagalog, Japanese) and intended emotion (anger, joy, fear, and sadness) as within subjects variables. The dependent variable was the percent correct recognition of intended emotion in each language.

Our second prediction was that listeners would recognize emotions best in their own language. An in-group advantage was confirmed by a main effect for language, $F(4, 395) = 22.22$, $p < .01$. Sheffé tests confirmed that accuracy was higher for English speech (mean = .94) than for German, Chinese, Japanese, or Tagalog speech (mean = .60) (p 's < .001).

Our third prediction was that recognition of vocal emotion would be better for sadness and anger than for joy and fear. Scheffé tests revealed that recognition accuracy was better across languages for anger than for joy ($p < .05$) or fear ($p < .01$). Recognition accuracy was also significantly better for sadness than for joy ($p < .01$) or fear ($p < .01$). There were no significant differences in recognition rates for anger and sadness, and no significant differences in recognition rates for joy and fear.

Recognition accuracy did not differ reliably as a function of intended emotion for all five languages, however. There was a main effect of intended emotion only for German, Tagalog, and Japanese stimuli.

To summarize, our first prediction, that listeners would be able to decode emotional prosody in all five languages, was supported. Recognition rates for all four emotions were above chance in all languages. Our second prediction, that listeners would recognize emotional prosody more easily in their own language than in another language, was also supported. Our third prediction, that sadness and anger should be better recognized than joy and fear, was also supported. Recognition accuracy across languages was significantly higher for anger and sadness than for joy and fear.

3.1. *Error analysis*

We also identified common confusions in emotional decoding. Were errors evenly distributed between the three incorrect choices or were some emotions more likely to be mistaken for each other? To examine this issue we created error matrices for the five languages (see Table 3). A Chi-Square analysis for each matrix indicated that the errors were not equally distributed in any case (p 's $> .05$). As shown in Table 3, listeners frequently confused sadness and fear with one another in all five languages. Listeners also frequently confused joy with anger in the English stimuli, and anger with joy in the Japanese and Tagalog stimuli.

3.2. *Acoustic analysis*

We next explored how much of the variance in recognition rates could be explained by similarities or differences in the acoustic attributes of emotive speech. Acoustic analyses of speech samples were performed to isolate vocal qualities that may have been used by listeners to decode emotional meaning. It is known that vocal qualities of speech are dependent on emotional intent (for reviews see Frick 1985; Murray and Arnott 1993; Scherer 1986). For example, analyses of sad speech tend to reveal

Table 3. *Error matrices: Percentage of incorrect judgments of anger, joy, sadness, and fear in five languages*

Language	Intended emotion	Response			
		Anger	Joy	Fear	Sadness
English	Anger	—	0.00	0.00	1.25
	Joy	1.25	—	0.00	0.00
	Fear	1.25	0.00	—	8.75
	Sadness	0.00	0.00	10.00	—
German	Anger	—	5.00	7.50	11.25
	Joy	3.75	—	17.50	21.25
	Fear	1.25	13.75	—	32.50
	Sadness	5.00	0.00	11.25	—
Chinese	Anger	—	16.25	11.25	11.25
	Joy	12.50	—	26.25	13.75
	Fear	8.75	2.50	—	21.25
	Sadness	1.25	1.25	36.25	—
Tagalog	Anger	—	8.75	1.25	1.25
	Joy	22.50	—	22.50	6.25
	Fear	10.00	17.50	—	18.75
	Sadness	0.00	0.00	2.50	—
Japanese	Anger	—	35.00	11.25	6.25
	Joy	8.75	—	30.00	3.75
	Fear	3.75	12.50	—	51.25
	Sadness	2.50	3.75	15.00	—

relatively low intensity, whereas anger is characterized by relatively high average intensity. We used Praat software (Boersma and Weenink 2001) to obtain measures of duration (in seconds), the mean and range of fundamental frequency (F_0), and the mean and range of intensity (dB). A measure of event density was also obtained by having two independent raters count the events (peaks in the waveform) of each utterance visually and auditorally. The number of events in each utterance was divided by the duration of the utterance to arrive at a measure of events per second, or event density.

The means of these measures for each emotion in each language set are shown in Table 4. We first assessed whether each acoustic parameter varied significantly depending on the intended emotion. If so, then listeners may have used these acoustic parameters as a perceptual cue for identifying the intended emotion. We also determined if the patterns of association between these measures and the four emotions differed as a function of language.

For each measure, we conducted a two-way ANOVA with language and intended emotion as independent variables. Mean fundamental

Table 4. Means of prosodic features by emotion and language set

Language	Intended emotion	Mean F ₀ (Hz)	Range F ₀ (Hz)	Mean (dB)	Range (dB)	Event density
English	Joy	256.52	272.97	74.73	28.09	4.93/s
	Sadness	156.05	140.41	61.31	24.07	4.19/s
	Anger	178.42	138.64	71.46	35.33	4.99/s
	Fear	179.60	127.36	64.48	26.20	4.74/s
German	Joy	203.80	248.97	63.91	28.23	4.72/s
	Sadness	142.28	175.61	55.30	18.24	4.81/s
	Anger	158.66	175.61	68.30	34.71	5.21/s
	Fear	176.19	176.87	61.35	23.85	4.90/s
Chinese	Joy	252.09	205.15	71.86	26.67	4.67/s
	Sadness	159.96	124.87	64.10	22.62	4.91/s
	Anger	259.91	225.57	75.29	25.19	5.62/s
	Fear	197.52	198.71	67.19	29.54	4.74/s
Tagalog	Joy	303.63	272.59	73.74	35.44	4.71/s
	Sadness	174.98	106.66	65.14	27.22	3.63/s
	Anger	265.90	210.14	76.11	34.38	5.16/s
	Fear	217.81	177.92	65.99	34.58	4.87/s
Japanese	Joy	216.83	193.38	68.69	34.69	4.48/s
	Sadness	154.08	99.48	63.02	29.63	4.14/s
	Anger	179.32	134.23	70.73	34.57	5.26/s
	Fear	169.17	197.77	64.54	33.70	4.15/s

Notes: Mean and Range F₀ = mean and range of fundamental frequency measured in Hertz (Hz). Mean and Range dB = mean and range of intensity measured in decibels (dB). Event density measured in events per second (s).

frequency (F₀) varied significantly as a function of intended emotion, $F(3, 60) = 14.82$, $p < .001$, with higher means associated with the emotion of joy ($M = 246.57$ Hz) and anger ($M = 208.44$ Hz) than with fear ($M = 188.06$ Hz) and sadness ($M = 157.47$ Hz). There was also a significant main effect of language, $F(4, 60) = 6.99$, $p < .001$ with the highest fundamental frequency associated with Tagalog speech (mean F₀ = 240.58 Hz) and the lowest fundamental frequency associated with German speech (mean F₀ = 170.23 Hz). We are reluctant to interpret this difference, however, because we only recorded two voices for each language. There was no significant interaction between intended emotion and language set for mean F₀.

For range F₀, there was again a significant main effect of intended emotion, $F(3, 60) = 5.020$, $p < .01$. Joy had the greatest F₀ range ($M = 238.62$ Hz) and sadness had the smallest F₀ range ($M = 129.41$ Hz). There was no significant main effect of language and no interaction between language and intended emotion.

The analysis of mean intensity also revealed a significant main effect of intended emotion, $F(3, 60) = 33.62, p < .001$. Anger and joy had the highest mean amplitudes (M 's = 72.43 dB and 70.59 dB, respectively) followed by fear ($M = 64.71$ dB) and then sadness ($M = 61.78$ dB). There was also a main effect of language, $F(4, 60) = 10.87, p < .001$, with lower intensity associated with German speech (mean = 62.28 dB) than with other languages (means ranged from 66.75 to 69.61 dB). There was no significant interaction between intended emotion and language.

For range dB, there was a significant main effect of intended emotion, $F(3, 60) = 17.20, p < .001$, with a lower range of amplitude for sad speech ($M = 24.36$ dB) than for the other three emotions (M 's = 29.58 to 32.84 dB). There was also a main effect of language, $F(4, 60) = 12.95, p < .001$, with greater amplitude ranges for Tagalog and Japanese speech (mean range = 32.91 dB and 33.15 dB respectively) than for the other three languages (all means < 28.50 dB). For this measure, there was a significant interaction between language and intended emotion, $F(12, 60) = 2.42, p = .02$, suggesting that the association between range dB and intended emotion was not consistent across languages.

For event density, there was a main effect of intended emotion, $F(3, 60) = 3.85, p < 0.02$, with higher event density for angry speech ($M = 5.23/s$) than for sad speech ($M = 4.34/s$). There was no main effect of language and no significant interaction between language and intended emotion.

To summarize, the analysis indicated that all acoustic measures — the mean and range of fundamental frequency, the mean and range of intensity, and event density — differed as a function of intended emotion. Only the association between range of intensity and emotion significantly varied as a function of language. Across languages, joyful and angry utterances were characterized as having a higher mean pitch and mean intensity than sad or fearful utterances, and angry speech had higher event density than sad speech. In short, our selected acoustic measures varied significantly as a function of intended emotion and could have been used by listeners as cues for determining the intended emotion. We therefore examined the extent to which each of these emotive elements could predict the judgments.

3.3. *Multiple regression analysis*

Multiple regressions were conducted to assess the extent to which variation in the acoustic measures could predict the emotion judgments of listeners. The dependent variable was the percentage of joy, anger, sadness,

Table 5. Summary: Multiple regression analyses of prosodic features on emotion judgments across languages

Emotion category	Significant predictors	Adj. R ²
Anger	+Mean dB, +Range dB, +Edensity	.413
Joy	+Range F ₀	.201
Sadness	–Mean dB, –Range dB, –Edensity	.557
Fear	–Mean dB	.022 (ns)

Note: Mean and Range dB = mean and range of amplitude. Mean and Range F₀ = mean and range of fundamental frequency. Edensity = event density.

Table 6. Summary: Stepwise Multiple regression analyses of prosodic features on emotion judgments for the English language set

Emotion category	Significant predictors	Adj. R ²
Anger	+Range dB	.491
Joy	+Mean dB, +Range F ₀	.629
Sadness	–Mean dB	.363
Fear	None	n/a

Table 7. Summary: Stepwise Multiple regression analyses of prosodic features on emotion judgments for the German language set

Emotion category	Significant predictors	Adj. R ²
Anger	+Range dB	.488
Joy	None	n/a
Sadness	–Range dB	.674
Fear	None	n/a

and fear responses for each utterance, resulting in four regression analyses. The results are summarized in Table 5.

Greater average intensity, greater range of intensity and greater event density were all associated with greater percentage of anger responses by listeners. Greater range of fundamental frequency was associated with a greater percentage of joy responses. A lower mean and range of intensity and a fewer number of events/second were associated with increased percentage of sad responses. A lower mean intensity was associated with fear responses.

To determine if these patterns were consistent within our five language sets, we conducted separate multiple regressions for each language set. Each stepwise equation tested the predictive value of these cues for emotion category selection in each language. The results are summarized in Tables 6, 7, 8, 9, and 10.

Table 8. *Summary: Stepwise Multiple regression analyses of prosodic features on emotion judgments for the Chinese language set*

Emotion category	Significant predictors	Adj. R ²
Anger	+Mean dB	.521
Joy	None	n/a
Sadness	–Mean dB, –Range dB	.841
Fear	+Range dB, –Mean F ₀	.695

Note: Mean and Range dB = mean and range of amplitude. Mean and Range F₀ = mean and range of fundamental frequency. Edensity = event density.

Table 9. *Summary: Stepwise Multiple regression analyses of prosodic features on emotion judgments for the Tagalog language set*

Emotion category	Significant predictors	Adj. R ²
Anger	+Mean dB	.507
Joy	+Mean F ₀	.341
Sadness	–Mean F ₀	.534
Fear	None	n/a

Table 10. *Summary: Stepwise Multiple regression analyses of prosodic features on emotion judgments for the Japanese language set*

Emotion category	Significant predictors	Adj. R ²
Anger	+Edensity	.486
Joy	None	n/a
Sadness	–Mean dB	.408
Fear	None	n/a

For utterances spoken in English, anger responses were associated with a greater range of intensity. Joy responses were associated with greater range of pitch and greater mean intensity. Sadness responses were associated with lower mean intensity. Fear responses were not associated with any of the acoustic measures.

For utterances spoken in German, anger responses were associated with a greater range of intensity. None of the acoustic measures was predictive of joy responses. Similar to the results for the English speech set, sad responses were associated with lower mean intensity. None of the cues was associated with fear responses.

For utterances spoken in Chinese, anger responses were associated with a greater mean intensity. None of the acoustic measures was associated with joy responses. Sad responses were associated with lower mean inten-

sity and lower range of intensity. Fear responses were associated with greater range of intensity and lower pitch range.

For utterances spoken in Tagalog, anger responses were associated with greater mean intensity. Joy responses by both groups were associated with a higher mean pitch. Sadness responses were associated with lower mean pitch. None of the acoustic measures were predictive of fear responses.

In the Japanese set, greater event density was associated with anger responses. None of the acoustic measures was associated with joy responses. Sadness responses were associated with lower mean intensity. None of the acoustic measures was predictive of fear responses.

4. Discussion

The results of this investigation indicate that English-speaking listeners are capable of decoding emotional prosody not only in their own language but also in a range of unfamiliar Western and non-Western languages. Other researchers have observed that listeners can decode emotional prosody in unfamiliar languages at rates better than expected by chance (Banse and Scherer 1996; Johnstone and Scherer 2000), but these studies have primarily focused on Western languages. Our results complement previous investigations by revealing that listeners can decode emotional prosody in both Western and non-Western languages. For the five languages tested, listeners recognized all four emotions at rates that are well above that expected by chance (mean proportion correct = .70).

Listeners recognized emotions most successfully in their own language, supporting predictions of an in-group advantage. They were most accurate at identifying emotions expressed in English (mean proportion correct = .94), and least accurate at identifying emotions expressed in Japanese and Chinese (mean proportion correct = .54 and .59, respectively). This specific difficulty in decoding Chinese and Japanese prosody supports Elfenbein and Ambady's (2003) cultural proximity hypothesis, in that Chinese and Japanese represent very different cultures than that of our English-speaking judges. Although the latter implications are intriguing, it would be premature to draw strong conclusions about differences in recognition accuracy between languages considering the limited number of speech samples used in this study.

Recognition of sadness and anger in speech was better overall than recognition of fear and joy, as reported in previous research (e.g. Juslin and Laukka 2003; Banse and Scherer 1996; Johnstone and Scherer 2000; Pitam and Scherer 1993). Evolutionary factors have often been proposed to

account for enhanced sensitivity to some emotions over others. For example, it is highly adaptive to recognize and react to anger as a potential threat to one's safety (Tooby and Cosmides 1990; Ohman et al. 2001). Recognizing a pleasant sound may not bear on one's survival, but recognizing and locating a threatening sound may mean the difference between life and death. Supporting this view, Ohman et al. (2001) reported faster detection of a threatening angry face than a non-threatening happy face. Sensitivity to sadness, on the other hand, may be adaptive for group cohesion, because displays of sadness signal to group members the need for help, support and protection (Barbee et al. 1998; Miceli and Castelfranchi 2003; Sadoff 1996; Sarbin 1989). The overt expression of fear may have the adaptive purpose of alerting conspecifics to the existence of a threat and enlisting their aid (Parr 2003). However, individuals may also wish to disguise their fear reactions to avoid being perceived as a low-dominance member of the social group (Montepare and Dobish 2003).

Our analysis of speech stimuli revealed several acoustic cues associated with specific emotions, providing potential cues to decode emotional meaning. Regression analyses confirmed the predictive power of these cues for emotive judgments. Across languages, the ability to decode emotions was associated with acoustic qualities such as mean frequency, intensity, and event density. The current data complement other acoustic analyses of emotional speech (Banse and Scherer 1996; van Bezooijen et al. 1983), and extend those analyses to both Western and non-Western languages.

How might one account for relationships between acoustic cues and specific emotions? Scherer's *component process model* describes physiological changes that occur in the vocal apparatus during the expression of emotions. For example, speaking of something that is deeply unpleasant often manifests as faucal and pharyngeal constriction and a tensing of the vocal tract. The acoustic outcome is higher frequency energy. Based on this model, Banse and Scherer (1996) predicted several associations between the acoustic attributes of speech and emotional intent that are similar to those observed here. Scherer (1989) also distinguished between the 'push effects' of the acoustic features of communication and the 'pull effects' of the social norms constraining both the expression and recognition of specific emotions, advocating further exploration of both types of factors.

The Brunswikian lens model (Brunswik 1956) also provides a framework for understanding the ability to decode emotional meaning from speech (Scherer 1982; Juslin and Laukka 2003). Originally designed to describe visual perception, the model has been adapted to many types of human judgments. In a Brunswikian framework the intent of the encoder to

express an emotion is facilitated by the use of a large set of cues that are probabilistic and partially redundant. Each cue by itself is not a reliable indicator of the expressed emotion, but combined with other cues, each one contributes in a cumulative fashion to the communication and recognition of emotion. The model incorporates the flexibility of the decoder to shift expectations from unavailable to available cues (Juslin 2000). The results of our investigation are consistent with a Brunswikian model in that multiple and overlapping cues were related to judgments of the four emotions.

5. Conclusion

Listeners can decode vocal expressions of emotion in unfamiliar languages, suggesting that some prosodic cues to emotion are universal. Evidence for an in-group advantage, however, points to cultural determinants of the production and/or perception of emotional prosody. Together, the findings suggest a blended model in which emotions are communicated prosodically through a combination of culturally determined and universal cues. According to Matsumoto (1989), emotions are universal but the ability to control and decode their expression is highly dependent on cultural factors. In particular, displays or explicit recognition of emotions affect social interactions. Social interactions can be disrupted either by displaying an emotion (Ekman 1972) or by recognizing an emotion (Matsumoto 1992). Thus, any cultural variation in the perceived significance of such interactions is reflected in norms for displaying and decoding emotions. Understanding how emotion judgments are guided by physical properties of stimuli and cultural norms has valuable implications for cross-cultural communication in many domains, including business, education, and conflict resolution.

References

- Albas, D. C., McCluskey, K. W., and Albas, C. A. (1976). Perception of the emotional content of speech: A comparison of two Canadian groups. *Journal of Cross-Cultural Psychology* 7, 481–490.
- Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology* 70 (3), 614–636.
- Barbee, A. P., Rowatt, T. L., and Cunningham, M. R. (1998). When a friend is in need: Feelings about seeking, giving, and receiving social support. In *Handbook of Communication and Emotion: Research, Theory, Applications, and Contexts*, Peter A. Anderson and Laura K. Guerrero (eds.), 281–301. San Diego, CA: Academic Press.

- Biehl, M., Matsumoto, D., Ekman, P., and Hearn, V. (1997). Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability data and cross-national differences. *Journal of Nonverbal Behavior* 21 (1), 3–21.
- Bezooijen, R. van., Oto, S. A., and 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.
- Boersma P. and Weenink D. (2001). Praat — a system for doing phonetics by computer. <http://www.fon.hum.uva.nl/praat/>
- Bolinger, D. L. (1978). Intonation across languages. In *Universals of Human Language*, vol. 2: Phonology, Joseph H. Greenberg (ed.), 471–524. Stanford: Stanford University Press.
- Brown, S. (2000). The 'musilanguage' model of music evolution. In *The Origins of Music*, N. L. Wallin, B. Merker, and S. Brown (eds.), 271–300. Cambridge, MA: MIT Press.
- Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments*. Berkeley: University of California Press.
- Darwin, C. (1872). *Expression of the Emotions in Man and Animals*. London: John Murray.
- Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In *Nebraska Symposium on Motivation*, J. Cole (ed.), 207–283. Lincoln: University of Nebraska Press.
- Ekman, P., Friesen, W., O'Sullivan, M., Chan, A., Diacyoyanni-Tarlatzis, I., Heider, K., et al. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology* 53, 712–717.
- Elfenbein, H. A. and Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin* 128 (2), 203–235.
- (2003). Cultural similarity's consequences: A distance perspective on cross-cultural differences in emotion recognition. *Journal of Cross Cultural Psychology* 34 (1), 92–109.
- Frick, R. W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin* 97, 412–429.
- Izard, C. E. (1971). *The Face of Emotion*. New York: Appleton-Century-Crofts.
- (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin* 115, 88–299.
- Johnstone, T. and Scherer, K. R. (2000). Vocal communication of emotion. In *Handbook of Emotion*, 2nd ed., Michael Lewis and Jeannette M. Haviland-Jones (eds.), 220–235. New York: Guilford Press.
- Juslin, P. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance* 26 (6), 1797–1813.
- Juslin, P. and Laukka, P. (2003). Communication of emotions in vocal emotion and music performance: Different channels, same code? *Psychological Bulletin* 129 (5), 770–814.
- Katsivriades, K. and Qureshi, T. (2003). Ten language families in detail. <http://www.Kryystal.com>
- Kitayama, S. and Ishii, K. (2002). Word and voice: Spontaneous attention to emotional utterances in two languages. *Cognition and Emotion* 16 (1), 29–59.
- Kramer, E. (1964). Elimination of verbal cues in judgments of emotion from voice. *Journal of Abnormal and Social Psychology* 68, 390–396.
- Matsumoto, D. (1989). Cultural influences on the perception of emotion. *Journal of Cross-Cultural Psychology* 20, 92–105.
- (1992). American-Japanese cultural differences in the recognition of universal facial expressions. *Journal of Cross-Cultural Psychology* 23, 72–84.
- Mesquita, B. and Frijda, N. H. (1992). Cultural variations in emotions: A review. *Psychological Bulletin* 112, 197–204.

- Miceli, M. and Castelfranchi, C. (2003). Crying: Discussing its basic reasons and uses. *New Ideas in Psychology* 21 (3), 247–273.
- Montepare, J. M. and Dobish, H. (2003). The contribution of emotion perceptions and their overgeneralizations to trait impressions. *Journal of Nonverbal Behavior* 27 (4), 237–254.
- Murray, I. R. and Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America* 93 (2), 1097–1108.
- Ohman, A., Lundqvist, D., and Esteves, F. (2001). The face in the crowd revised: A threat advantage with schematic stimuli. *Journal of Personality and Social Psychology* 80, 381–396.
- Parr, L. A. (2003). Case study 10A. Emotional recognition by chimpanzees. In *Animal Social Complexity: Intelligence, Culture, and Individualized Societies*, Frans B. M. de Waal and Peter L. Tyack (eds.), 288–292. Cambridge, MA: Harvard University Press.
- Pittam, J. and Scherer, K. R. (1993). Vocal expression and communication of emotion. In *Handbook of Emotions*, Michael Lewis and Jeannette M. Haviland (eds.), 185–197. New York: Guilford Press.
- Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of cross-cultural studies. *Psychological Bulletin* 115, 102–141.
- Sadoff, R. L. (1996). On the nature of crying and weeping. *Psychiatric Quarterly* 40, 490–503.
- Sarbin, T. R. (1989). Emotions as narrative emplotments. In *Entering the Circle: Hermeneutic Investigation in Psychology*, M. J. Packer and R. B. Addison (eds.), 185–201. Albany, NY: State University of New York Press.
- Scherer, K. R. (1979). Nonlinguistic vocal indicators of emotion and psychopathology. In *Emotions in Personality and Psychopathology*, C. E. Izard (ed.), 137–53. Plenum: New York.
- (1982). Methods of research on vocal communication: Paradigms and parameters. In *Handbook of Methods in Nonverbal Behavior Research*, K. R. Scherer and P. Ekman (eds.), 136–198. Cambridge: Cambridge University Press.
- (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin* 99, 143–165.
- (1989). Vocal correlates of emotional arousal and affective disturbance. In *Handbook of Social Psychophysiology*, H. Wagner and A. Manstead (eds.), 165–197. New York: Wiley.
- (1997). The role of culture in emotion-antecedent appraisal. *Journal of Personality and Social Psychology* 73, 902–922.
- Scherer, K. R., Banse, R., and Wallbott, H. G. (2001). Emotion inferences from vocal expressions across languages and cultures. *Journal of Cross-Cultural Psychology* 32 (1), 76–92.
- Standke, R. (1992). *Methods of Digital Speech Analysis in Research on Vocal Communication*. Frankfurt: Peter Lang.
- Sundita, C. (2003). Tagalog language. <http://www.wikipedia.org>.
- Tooby, J. and Cosmides, L. (1990). *Evolutionary Psychology and the Emotions*. In *Handbook of Emotions*, 2nd ed., M. Lewis and J. M. Haviland-Jones (eds.), 91–115. New York: Guilford Press.

William Forde Thompson (b. 1957) is Director of Communication, Culture and Information Technology at the University of Toronto at Mississauga <b.thompson@utoronto.ca>. His interests include music, gesture, and cognition. His recent publications include ‘Listening to Music’ (with G. Schellenberg, 2006); ‘A comparison of acoustic cues in music and speech for

three dimensions of affect' (with G. Ilie, 2006); and 'The subjective size of melodic intervals over a two-octave range' (with F. A. Russo, 2006).

Laura-Lee Balkwill (b. 1965) is a Postdoctoral Fellow at Queen's University <balkwill@post.queensu.ca>. Her research interests include the expression and recognition of emotion in music and speech across cultures, and the effect of emotive tone of voice on reasoning. Her recent major publications include 'Expectancies generated by recent exposures to melodic sequences' (with B. Thompson and R. Vernescu, 2000); 'Recognition of emotion in Japanese, North Indian, and Western music by Japanese listeners' (with B. Thompson and R. Matsunaga, 2004); and 'Neuropsychological assessment of musical difficulties: A study of "tone deafness" among university students' (with L. L. Cuddy, I. Peretz, and R. Holden, 2005).