Emotional Speech Processing and Language Knowledge

Conor I. Frye (cifrye@cogsci.ucsd.edu)
Department of Cognitive Science, UC San Diego, 9500 Gilman Dr. M/S 0515
La Jolla, CA 92093 USA

Sarah C. Creel (creel@cogsci.ucsd.edu)
Department of Cognitive Science, UC San Diego, 9500 Gilman Dr. M/S 0515
La Jolla, CA 92093 USA

Abstract

How does language knowledge affect processing of paralinguistic information—vocal properties that are not directly related to understanding words? This study investigates links between a listener's native language, any other languages they may have experience in, and the ability to identify vocal emotional information in those languages. The study focuses on two particular classes of languages: those with lexical tone, such as Mandarin Chinese, which use pitch properties to distinguish otherwise-identical words; and those without lexical tone, such as English. English listeners and bilingual Mandarin-English listeners listened to sentences and categorized the emotional content of English and Mandarin sentences. Half of the sentences were presented normally; the other half were low-pass filtered to remove all but prosodic cues (pitch and timing). English listeners were better at identifying emotions in English sentences, while bilinguals were equally good at identifying emotions in both languages. This indicates better overall emotion recognition from prosody alone for listeners more familiar with a language. It may point to a connection between tone language experience and augmented paralinguistic processing capabilities.

Keywords: speech perception; paralinguistic perception; voice; language background; individual differences; bilingualism

Introduction

Spoken language as a medium is not just a symbol system of discrete speech sounds; it is also replete with cues to the talker's identity, region of origin, and emotional state. Although much research has been devoted to understanding how exposure to a language affects speech sound identification (Kuhl, 1994), almost no one has asked how language knowledge affects processing of *paralinguistic* information—vocal properties that are not directly related to understanding words like speech rate and pitch changes (see Thompson & Ballkwill, 2006, for an exception). Emotion in the voice is thought to be conveyed by these paralinguistic cues. Though differing languages seem to use similar vocal acoustic cues for the "basic emotions" in non-speech vocalizations such as laughter and crying, it is not clear how readily listeners perceive these emotional cues cross-linguistically when only presented with the auditory signal (Sauter et al., 2010).

Language-specific recognition of vocal affect

One likely set of cues that listeners use to identify vocal emotion is *prosody*: pitch and timing information. Happy speech, for instance, typically has more variable pitch and volume, higher overall pitch level, and a faster speaking rate, whereas sad speech sounds exhibit lower average pitch, attenuated loudness and pitch variation, and a slower pace of speech (Morton & Trehub, 2001). Previous work demonstrates that humans use paralinguistic cues during speech to alert co-communicators to their current emotional state (Kehrein, 2002). However, this ability to attribute certain paralinguistic cues to particular emotional states may not be fully present at birth, but may require learning through lengthy exposure to one's native language.

One indication of the learned nature of paralinguistic processing is that children experience difficulty in identifying vocal emotional cues (Morton & Trehub, 2001); for instance, 6-year-olds who hear "my mommy gave me a treat" with "sad" emotional prosody will report that the speaker sounded happy, suggesting that they are still learning the mapping between particular speech patterns and emotional states. Further research suggests that these learned aspects may be language-specific (Thompson & Balkwill, 2006), though those authors do not pinpoint particular cues that may be relevant, nor do they offer a hypothesis as to what level of fluency one needs to access the learned aspects of emotional speech. In a related area, speaker recognition shows some language specificity in infants (Johnson et al., 2011) and adults (Bregman & Creel, 2012). If encoding of vocal emotional information works similarly to encoding of voices, then good emotional recognition within a language may be dependent on lengthy language experience, and may not generalize to emotion recognition in other languages.

General ability to recognize vocal emotion

On the other hand, there is evidence that expertise in processing the cues that communicate vocal emotion may generalize widely across domains. This implies that better attention to or encoding of pitch for another purpose or in another domain may lead to better perception of vocal emotion. For example, certain types of languages have been claimed to boost pitch perception abilities: Speakers of *tone languages* such as Mandarin are better at making relative

pitch distinctions in musical stimuli (Pfordresher & Brown, 2009). Conversely, musicians are better than non-musicians at brain encoding of *linguistic* pitch changes (Wong et al, 2007). These studies suggest that facility with pitch processing generalizes even across domains. This implies that good linguistic pitch processing should facilitate pitch processing generally, which would thereby facilitate perception of pitch-related vocal-emotional information, even outside one's native language. The novel prediction for vocal affect detection is that, over and above language-specific knowledge, tone-language speakers may excel at perceiving vocal-emotional information due to their language background.

The current study

The current study focuses on two hypotheses about processing of vocal affect. First, the *language-specificity hypothesis* suggests that listeners are best at identifying vocal affect in their native language, due to lengthy perceptual learning of vocal correlates of emotional states specific only to that language. This also assumes any second language that the speakers are fluent or near-fluent at will also experience this emotional state comprehension. Crucially, a listener who is not a fluent speaker of a language will have difficulty identifying emotion in that language relative to fluent speakers. Second, the *tone-language benefit hypothesis* posits that listeners with a history of speaking tone languages will show good identification of vocal affect even in non-native languages because tone languages generally facilitate listeners' processing of pitch information, and pitch information is one important cue to affect.

To investigate how language background affects emotional speech processing, we asked speakers of Mandarin Chinese (a tonal language) who also spoke English, and speakers of American English (a non-tonal language) to identify emotion in utterances produced in Mandarin and English. Participants' abilities to identify emotions in their own language were compared to their ability to identify emotions in languages unfamiliar to them.

Methods

Participants

Thirty-six undergraduates from the University of California, San Diego participated in this study for class credit. Eighteen of the participants were native English speakers who did not speak Mandarin Chinese, and the remaining eighteen were native speakers of Mandarin Chinese who also spoke English fluently as a second language. English was a second language for each of the 18 Mandarin speaking participants, who acquired English at a mean age of 8 years (range: 0-17).

Stimuli

Eight speakers recorded 96 sentences each in their native language. Four speakers (2 male, 2 female) were native speakers of English, and the other four (2 male, 2 female) were native speakers of Mandarin Chinese. The speakers' ages ranged from 19 to 26, with a mean age of 21.75.

Sentences spanned six different emotional states: anger, disgust, fear, happiness, sadness, and surprise. Sentence semantic content was created to elicit the intended emotion, to make the task of emotional speech production more naturalistic for our speakers. The 16 sentences for each emotion were originally written in English and translated to Mandarin Chinese. Each sentence contained five syllables in the English version. The translation was retranslated separately by all four Mandarin speakers to ensure a good content match with the original English sentences.

Sentences were recorded in a sound-attenuated booth and saved as .wav files. Files were edited so that each sentence had its own sound file. Two types of sound files were created for each sentence. One type of stimulus (Figure 1a) reflected the original recording, complete with naturalistic emotional sentence semantic content. The other (Figure 1b) was low-pass filtered at 500 Hz using Praat software (Boersma & Weenink, 2011). Low-pass filtering removes high-frequency information including cues to consonants and vowels, while retaining low-frequency information in the speech signal. The result is a muffled, unintelligible sound that preserves fundamental frequency variability, including lexical tone (Mandarin only), prosody, and speech rate. This manipulation allowed for the measurement of prosody recognition without the confound of language-specific semantic content. Each file was set to an average loudness of 60 decibels.

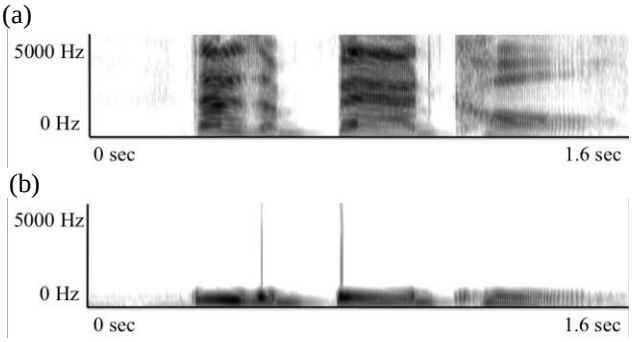

Figure 1. Spectrogram of an (a) unfiltered and (2) low-pass filtered angry sentence.

Procedure

Each participant listened to 384 total sentences, counterbalanced across two conditions so no participant heard the same sentence from the same speaker or in the same language or in the same filter condition twice. Of the 384 sentences, half each were English and Mandarin; crossed with this, half were unfiltered and half were filtered. Participants were asked to identify the emotional state of each sentence as it was presented through Sennheiser HD 280 Pro headphones. The computer monitor displayed the six possible emotions and a number that corresponded to each emotion. Participants pressed the number key that they

thought matched the emotion of the sentence. Perceived emotional responses and reaction times were recorded in Matlab using the Psychtoolbox3 (Brainard, 1997; Pelli, 1997).

Results

For the current experiment, the language-specificity hypothesis and the tone-language hypothesis make similar predictions regarding the participants due to the fact that our tone language speakers were fluent bilinguals. If listeners showed native language specificity of emotion recognition, then accuracy should be higher for English listeners in English. Mandarin-English bilinguals should perform equally well in both languages—either due to knowledge of both languages, or due to enhanced abilities as tone-language speakers. This may vary by *degree* of language fluency, as assessed by age of English acquisition. Importantly, this pattern should still hold for filtered speech, which crucially does not contain any semantic language information. That is, if listeners are using language-specific acoustic cues to vocal emotion, they should still be more accurate at recognizing emotion in a familiar language even when lexical cues are removed.

To test this hypothesis, we performed a mixed ANOVA on recognition accuracy with Listener Language (English, Mandarin) as a between-participants variable, and Stimulus Language (English, Mandarin) and Intelligibility (unfiltered, filtered) as within-participants variables. The three-way interaction between these variables was significant ($F(1,34)=121.70$, $p<.0001$). This interaction qualified all lower-level effects and interactions.

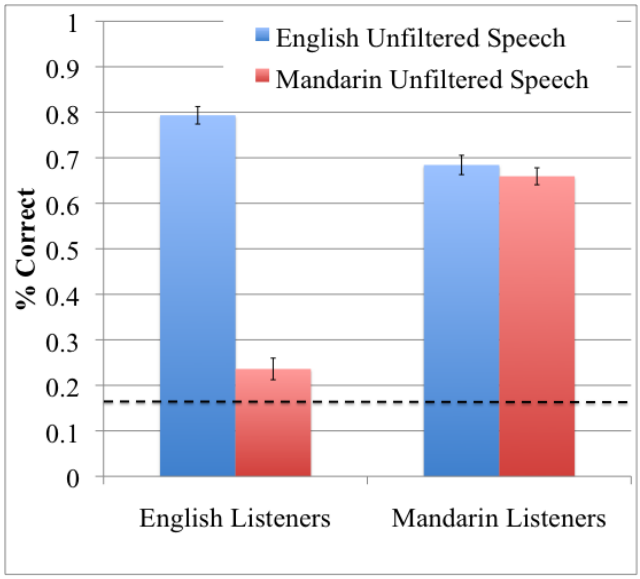

Figure 2. Accuracy for unfiltered sentences with standard errors. The dotted line represents chance performance

Therefore, we broke the data out into filtered and unfiltered data to better highlight interactions at that level. For unfiltered sentences, with full naturalistic verbal content, there was a significant interaction between Listener Language and Stimulus Language ($F(1,34)=325.58$, $p<.0001$), indicating that participants were better able to comprehend emotional affect in languages they spoke highly fluently, which is supportive of our language-specificity hypothesis. This result is also important as a control for the stimuli used, and demonstrated that the sentences provided ample emotional content clues. When presented with unfiltered speech, English speakers were significantly more accurate with English speech ($t(17)=63.706$, $p>.0001$) whereas Mandarin speakers were equally proficient at identifying emotional affect in both languages ($t(17)=.891$, $p=.385$) as would be expected from their language background.

Considering the filtered stimuli, which contained only prosodic cues, we again found an interaction of Listener Language and Stimulus Language ($F(1,34)=46.278$, $p<.0001$), indicating that even when there was a lack of verbal information, participants were significantly more capable to parse emotional affect when presented with languages they spoke fluently. Considering each listener language group individually, English speakers were significantly more accurate in identifying the intended emotion when given filtered English speech than filtered Mandarin speech ($t(17)=9.949$, $p<.001$). Mandarin speakers, however, showed good performance on filtered speech in both languages with no significant differences in accuracy ($t(17)=1.923$, $p=.0714$).

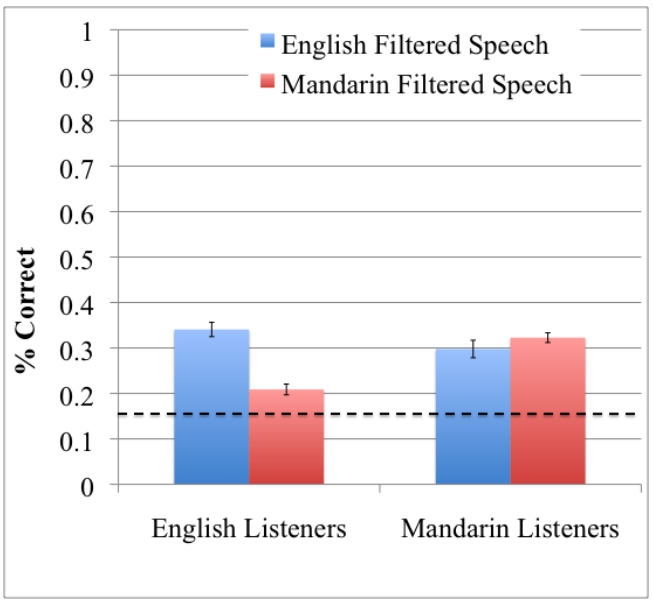

Figure 3. Accuracy for filtered (prosody only) sentences with standard errors. The dotted line represents chance performance

Finally, we assessed whether English performance was affected by degree of English fluency by calculating the correlation between bilinguals' English accuracy on filtered speech and their age of English acquisition. This correlation was not significant ($r(16) = -.2176$, $p = .3856$), suggesting that somewhat surprisingly, age of second language acquisition was unrelated to ability to process semantic-information-free speech.

Discussion

Previous work has demonstrated a link between tonal language background and enhancements of abilities in other perceptual domains such as music (Pfordresher & Brown, 2009), but the link between language background and vocal emotion had been underexplored. The current study explored the relationship between language background and vocal affect identification, focusing on hypotheses of a *language-specific benefit* and a *tone-language benefit* in vocal affect identification.

In support of these hypotheses, we showed that listeners are better at discerning emotional content in speech for all languages they have achieved fluency in, even when high-frequency lexical cues are removed. This is important, because the lack of high frequency information removes any clues as to the specific language being presented. This means that all performances on filtered speech represent the participant's processing of the low frequency emotional pitches and tonal changes without the influence of any clues to the actual language. This implies that any responses are based entirely on pitch processing, and any benefits can only come from validation of one of the two hypotheses presented. The data demonstrate, specifically, that English monolinguals identified emotions more accurately in English than in Mandarin, whereas Mandarin-English bilinguals showed equivalent performance in both languages. This is consistent with the *language-specificity hypothesis:* that listeners are better at discerning cues to emotion in their native language. However, due to the design of the current study, it is also consistent with the *tone-language facilitation hypothesis*: that tone-language speakers, due to lengthy experience attending to fine-grained pitch characteristics of language, have a general advantage at recognizing vocal emotion. That is, Mandarin listeners performed at above-chance levels in both languages because they are tone-language speakers, not only because they also speak English. This would predict that Mandarin listeners would also be superior at emotion recognition in a completely unfamiliar language, a hypothesis we are currently testing. Further study is currently being performed to address these design limitations, and will alleviate the current confound of the tonal language speaking subjects being fluent in the languages of all the presented stimuli

Nevertheless, in the current study, discussion of the implications of both hypotheses is warranted by the data. In regards to the first hypothesis, the language-specific hypothesis, our data are consistent with the expectation that monolingual listeners perform better in identifying vocal emotion from prosodic cues in their native language. Listeners familiar with two languages (Mandarin and English) performed comparably in both languages. It is possible more subtle effects were also present in the data regarding the bilingual speakers' performances. For instance, if the emotional-speech processing capabilities are affected by the "sensitive period" demonstrated for phonology, then early-exposed speakers of a second language should show native-level abilities in emotional speech processing, and those who acquired their second language later should not. When a Pearson's product-moment correlation was run on the data, however, the correlation suggested no link between a subject's age of acquisition and performance on filtered speech processing. If age of acquisition does not play a correlational role, some other aspect of the subjects' background must be dictating their abilities.

This leaves the possibility that a general benefit for tone-language speakers better accounts for the results. The current results, although consistent with a language specific (or familiar-language) benefit, are also consistent with the tone-language benefit hypothesis; speakers of tonal languages performed equally well at identifying vocal affect in both languages presented. However, the data described here cannot distinguish whether bilinguals' good performance in both languages resulted from their tone-language background, or simply being fluent in both languages present in the stimuli. Further study will present these subject groups with tonal and non-tonal languages with which they are not familiar.

If there is a general tone-language advantage, Mandarin speakers should outperform English speakers on all *un*familiar languages. This would indicate that a tonal language background affords the speaker with a type of emotional prosody processing training, and that there are aspects of tonal languages that improve the processing capabilities of individuals in emotional speech. This might be true if there are universal pitch characteristics that can be found in all languages, but are very subtle and difficult to pick up on if the listener does not have sufficient training in either that specific language or excellent awareness in pitch perception in general.

There could even be a specific tone-language advantage, such that tone-language speakers would outperform non-tone language speakers only for unfamiliar *tone* languages. This data pattern might hold if tone languages use devices to convey vocal emotion that are similar across a range of tone languages, but different than non-tone languages. We are currently testing these possibilities regarding the tone-language advantage with English speakers and Mandarin-English bilinguals and four languages (English; Mandarin; Dutch [unfamiliar non-tone language]; Vietnamese [unfamiliar tone language]).

The present data demonstrate that there is emotional affect information present even without higher frequency information. It also provides evidence that speakers are

capable of picking up on this information without relying on distinct linguistic information in languages that they are familiar with. The current study provides a tantalizing peek into the emotional affect processing provided by language background, and with further study already in process, moves our understanding of vocal emotional affect processing forward.

Acknowledgements

CIF was supported by the Center for Research in Language at UC San Diego, and SCC was supported by an NSF CAREER Award (BCS-1057080). We would like to acknowledge the help of undergraduate research assistants Sheryl Soo, Christina Hwang, and Ben Howard, who were instrumental in developing language stimuli and collecting data on participant's language background.

References

Balkwill, L., & Thompson, W. (1999). A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music perception*, *17*(1), 43–64.

Bregman, M. R., & Creel, S. C. (2012). Learning to recognize unfamiliar voices: the role of language familiarity and music experience. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, 5, 1002-1011.

Kehrein, R. (2002). The prosody of authentic emotions. *Proc. Speech Prosody Conf* (pp. 423–426).

Kuhl, P. K. (1994). Learning and representation in speech and language. *Current Opinion in Neurobiology*, 4, 812-822.

Ley, R. G., & Bryden, M. P. (1982). A dissociation of right and left hemispheric effects for recognizing emotional tone and verbal content. *Brain and cognition*, *1*(1), 3–9.

Morton, J. B., & Trehub, S. E. (2001). Children's Understanding of Emotion in Speech. *Child Development*, 72(5), 834-843.

Pfordresher, P. Q., & Brown, S. (2009). Enhanced production and perception of musical pitch in tone language speakers. *Attention, Perception, & Psychophysics*, *71*(7), 1385–1398. Springer.

Sauter, D. A, Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(7), 2408-12.

Slevc, L. R., & Miyake, A. (2006). Individual differences in second-language proficiency: does musical ability matter? *Psychological Science*, *17*(8), 675-81.

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

Thompson, W. F., & Matsunaga, R. I. E. (2004). Recognition of emotion in Japanese, Western, and Hindustani music by Japanese listeners 1, *46*(4), 337–349.

Thompson, W., Schellenberg, E. G., & Husain, G. (2006). Perceiving Prosody in Speech. *Annals of the New York Academy of Sciences*, *999*, 530–532.

Thompson, W. F., Schellenberg, E. G., & Husain, G. (2004). Decoding speech prosody: do music lessons help? *Emotion (Washington, D.C.)*, *4*(1), 46–64.