Implicit recognition of vocal emotions in native and non-native speech

Marc D. Pell

School of Communication Sciences and Disorders
McGill University, Montréal, CANADA
marc.pell@mcgill.ca

Abstract

There is evidence for both cultural-specificity and 'universality' in how listeners recognize vocal expressions of emotion from speech. This paper summarizes some of the early findings using the Facial Affect Decision Task [1, 2] which speak to the implicit processing of vocal emotions as inferred from "emotion priming" effects on a conjoined facial expression. We provide evidence that English listeners register the emotional meanings of prosody when processing sentences spoken by native (English) as well as non-native (Arabic) speakers who encoded vocal emotions in a culturally-appropriate manner. As well, we discuss the timecourse for activating emotion-related knowledge in a native and non-native language which may differ due to cultural influences on vocal emotion expression.

1. Introduction

Recent decades are witness to a growing literature on whether nonverbal emotion signals are recognized through 'universal' principles shared among human cultures, or rather, shaped in a distinct manner by cultural conventions. To date, significant evidence has emerged favouring the notion of universality as well as certain cultural variations in how emotions are understood when the facial channel of expression has been studied [3, 4]. However, similar cross-cultural evidence on the recognition of vocal expressions of emotion (i.e., emotional prosody) is comparatively scarce [5], and accordingly, implications of these data for understanding notions of universality and culture specificity in vocal emotion processing are less conclusive.

Nonetheless, results of several studies indicate that members of a given cultural background can infer emotional meanings from vocal cues encoded by members of a distinct cultural/language community with better than chance accuracy. For example, Scherer et al. [5] reported extensive data on the recognition of vocal expressions of emotion involving listeners recruited from nine countries in Europe, Asia, and America. They found that emotionally-inflected "pseudosentences" spoken by German actors to express *fear*, *joy*, *sadness*, *anger*, or neutral affect were identifiable by all cultural groups at levels beyond chance. Moreover, there was a high degree of similarity in emotion confusion patterns across cultures.

At the same time, the researchers noted that the emotional judgements of German (i.e., native) listeners were most reliable overall, and recognition accuracy tended to be higher for those cultural groups whose own language was more linguistically related to German. Collectively, results argue for the existence of universal inference rules which may be applied by listeners to infer the emotional meaning of prosody in a non-native language [5]. However, the importance of both culture-specific and culture-independent factors in the identification of vocal emotions from speech was highlighted by the overall findings and should be considered in future research in this area.

2. Implicit processing of vocal emotions from speech

Most if not all studies which have investigated cultural influences on vocal (and facial) emotion recognition have employed tasks in which subjects explicitly categorized the emotional meaning of the target expression, typically in a multiple-choice response format. It is possible that such "off-line" judgements of emotional prosody influence the way participants actually perceive the presented stimulus [4], and if employed exclusively, this behavioural approach prevents a complete view of what meaning specifications are likely generated by vocal-prosodic cues as speech is processed in real time (i.e., "on-line").

To address the paucity of on-line research on vocal emotion processing, the Facial Affect Decision Task (FADT) has proven successful in evaluating whether listeners detect the emotional meanings of prosody when listening to spoken utterances in a more implicit manner. On the assumption that both facial [6] and vocal [7] expressions of emotion activate categorical knowledge held in associative memory, the FADT gauges the influences of emotional prosody on decisions about a related or unrelated, conjoined facial expression of emotion through evidence of emotion priming effects. Subjects are presented an emotionally-inflected sentence followed by a static facial expression that represents an unambiguous exemplar of emotion or a facial "grimace" that does not represent an emotional state. Akin to executing a lexical decision, participants judge whether the face represents a "true" expression of emotion (YES/NO response) in the absence of any explicit requirements to identify emotional attributes of the prosody prime or the face target stimulus.

By manipulating the emotional relationship of the vocal and facial expressions across a series of trials, the implicit meanings activated by prosodic information when passively listening to prime sentences can be inferred by the presence of congruity effects involving a related versus an unrelated facial expression which shares emotion category membership with the vocal expression. This approach was employed in two recently-published experiments which investigated the ability of English listeners to detect vocal expressions of emotion produce by *native* speakers of the same language [1, 2].

3. Vocal emotion recognition from the native language

In our initial experiments which tested the implicit recognition of vocal emotions produced by *native* speakers of the same

language (English), vocal primes consisted of "pseudo-utterances" (e.g., *Someone migged the pazing*) which had been emotionally-intoned by male and female speakers of English. All stimuli were carefully validated prior to the study to ensure that prosodic and facial materials reliably conveyed the intended emotional meanings when presented to an independent group of perceptual raters.

In an initial study [1], prosody and face stimuli that unambiguously conveyed *happiness*, *(pleasant) surprise*, *anger*, and *sadness* were exhaustively combined across emotions for presentation to a group of young adults. The resulting data established that participants were significantly more accurate and faster to render facial affect decisions when a target face was preceded by an emotionally-congruent rather than incongruent prosody. These findings were obtained in spite of instructions to attend strictly to the face, implying that the emotional significance of vocal-prosodic cues was registered when passively listening to pseudo-utterance primes, perhaps in an involuntary manner [8]. The observed congruity effects were attributed to the activation of conceptual knowledge about emotion categories which is partly shared by expressions of emotion in the prosody and face channels.

In a follow-up study which presented paired vocal and facial expressions conveying a *happy* or *sad* emotion to an independent group of listeners [2], it was again shown that facial affect decisions were systematically enhanced when prosodic primes were congruent rather than incongruent in emotion with the face target (i.e. emotion priming was observed). These results reinforced our earlier conclusion that listeners implicitly activate emotion-specific features of prosodic information when processing speech, at least when these vocal expressions are produced according to conventions appropriate to native speakers of the same language.

In addition to testing for the presence of congruity effects in emotion processing, this study sought to determine *how much* prosodic information may have been necessary to activate emotional meanings of the prosody which prime facial affect processing. To accomplish this, pseudo-utterances produced in a happy or sad prosody were cut into fragments lasting 300, 600, and 1000 milliseconds from the onset of the stimulus and then presented as prime stimuli in three separate conditions. Results indicated that emotional priming between the prosody and face was only evident in conditions which presented 600 and 1000 millisecond prosodic fragments (but not 300 ms primes); moreover, there was evidence of maximal priming of facial affect decision latencies when a congruent prosodic stimulus lasted approximately 600 as opposed to 1000 milliseconds. These patterns suggest that listeners require critical levels of exposure to vocal expressions signifying emotions such as happy or sad, possibly in the range of 600 milliseconds, to activate underlying emotional meanings when these are encoded by native speakers [2].

However, it is entirely unknown whether English listeners would recognize the emotional meanings of prosody encountered in a non-native (i.e., completely unknown) language, and if so, whether the activation of emotion-related knowledge in memory would reveal a distinct timecourse according to the cultural expression of vocal emotions in different languages. The FADT can be usefully extended to consider these questions by presenting prosodic primes which are emotionally-inflected by speakers of a foreign language to test whether English listeners recognize the emotions expressed, and if so, in which time conditions.

4. Vocal emotion recognition from an unknown language

In our most recent study [9], we employed the FADT to investigate whether English listeners infer emotional meanings from vocal expressions encoded by speakers of an unfamiliar language, Arabic.

The methods of Pell [2] were modified by replacing English-like pseudo-utterances produced in a *happy* or *sad* prosody with comparable utterances produced by two male and two female speakers of Arabic. The Arabic pseudo-utterances were then cut into prime fragments lasting 600 and 1000 milliseconds from the onset of the stimulus and paired with face targets in a similar manner to our English study [2]. Participants were informed that they would hear the voice of a "foreign sounding" speaker prior to each face, but as in previous administrations of the FADT, they were encouraged to judge the emotional status of the facial expression as accurately and as quickly as possible.

Based on data from cross-cultural studies summarized in section 1, we predicted that vocal expressions of emotion produced by Arabic speakers would be meaningful to English listeners, influencing the accuracy and/or response latency of facial affect judgments giving rise to emotion priming effects [5]. However, we also predicted that cultural variations in the expression of vocal emotion might delay recognition of emotional meanings when exposed to non-native compared to native speech samples [10], yielding significant priming only in conditions of a longer prosodic stimulus when listening to non-native speech.

There were strong indications in our new data that English listeners engaged in a meaningful analysis of emotional prosody and that they accessed relevant meanings of the vocal expressions in critical conditions, despite the fact that prosodic cues were expressed in an unknown and typologically distinct language such as Arabic. Similar to our findings when English listeners were presented native speech input [1,2], the accuracy of facial affect decisions was systematically biassed by the emotional relationship of vocal cues produced by Arabic speakers, yielding reliable emotion congruity effects in certain conditions (i.e., the 1000 ms prosody duration condition). These findings suggest that vocal expressions of emotion possess certain universally recognisable characteristics which were detected by our English listeners [5] despite their complete lack of familiarity with Arabic language and cultural expression.

However, it was also true that congruity effects in our cross-cultural data appeared to be more selective or "fragile" than in our studies which presented native vocal expressions of emotion as primes. Most notably, for non-native primes, emotion congruity effects on decision accuracy occurred only when 1000 milliseconds of prosodic information preceded the face target and showed virtually no differentiation in the prosody condition that restricted vocal information to 600 milliseconds in duration. These patterns contrast with our previous finding that when the prosody was produced by native speakers of English, maximal congruity effects were strongly associated with prosodic primes lasting 600 milliseconds in duration. These cross-cultural comparisons imply that when non-native prosody was processed, underlying knowledge about vocal expressions of emotion was not sufficiently activated in the 600 ms condition, precluding priming effects in this environment when speech input was not culturally appropriate to the listener.

When our data are compared in contexts of native versus non-native speech processing, the observed patterns are consistent with the idea that listeners apply 'universal inference rules' to gain meaning from emotional prosody in an unknown language [5]. However, perhaps owing to cultural differences in emotion expression in the vocal channel, it would appear that the recognition of emotional prosody from foreign (in this case Arabic) speech requires a longer level of exposure to spectro-temporal patterns in speech than the processing of emotion expressed through the decoders' native language [10]. We are now conducting a series of follow-up experiments to test the validity of these claims by presenting vocal expressions of emotion produced by our English and Arabic speakers to unilingual speakers of Argentinian Spanish; these data are still forthcoming.

5. Summary and future directions

To date, experiments of both off-line and now on-line speech processing reinforce the notion that there are certain universally recognisable characteristics of emotional prosody, similar to what has been demonstrated in the much larger literature on the facial channel. Our initial findings serve to reinforce this claim, using a novel experimental context which looked at automatic priming effects of emotional prosody on an emotional face, where we found that English listeners processed the emotional meanings of prosody in their native and a completely unfamiliar language.

However, the influences of cultural shaping variables on how emotion is expressed and understood through prosody were also emphasized by our findings and need to be explored further. In particular, it will be critical to pursue continued studies of how such factors as cultural distance and exposure duration (among others) play a role in how "universal principles" are presumably applied to understand vocal expressions from native and non-native speech.

The Facial Affect Decision Task and other "on-line" behavioural or neuroinvestigative paradigms seem well suited to advance this literature in a focussed and sensitive manner.

6. References

[1] Pell, M.D. (2005). Nonverbal emotion priming: Evidence from the 'facial affect decision task'. *Journal of Nonverbal Behavior, 29*(1), 45-73.

[2] Pell, M.D. (2005). Prosody-face interactions in emotional processing as revealed by the facial affect decision task. *Journal of Nonverbal Behavior, 29*(4), 193-215.

[3] Ekman, P., Friesen, W., O'Sullivan, M., Chan, A., Diacoyanni-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*(4), 712-717.

[4] Russell, J.A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin, 115*(1), 102-141.

[5] Scherer, K.R., Banse, R., & Wallbott, H. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology, 32*(1), 76-92.

[6] Etcoff, N.L., & Magee, J.L. (1992). Categorical perception of facial expressions. *Cognition, 44*, 227-240.

[7] Laukka, P. (2005). Categorical perception of vocal emotion expressions. *Emotion, 5*(3), 277-295.

[8] de Gelder, B. & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition & Emotion, 14*(3), 289-311.

[9] Skorup, V. & Pell, M.D. (In review). On-line sensitivity to vocal expressions of emotion in an unknown language.

[10] Beier, E., & Zautra, A. (1972). Identification of vocal communication of emotions across cultures. *Journal of Consulting and Clinical Psychology, 39*(1), 166.