RECOGNITION OF EMOTIONAL SPEECH FOR YOUNGER AND OLDER TALKERS: BEHAVIOURAL FINDINGS FROM THE TORONTO EMOTIONAL SPEECH SET

Kate Dupuis[1] and M. Kathleen Pichora-Fuller[1, 2]
[1]Dept. of Psychology, University of Toronto, 3359 Mississauga Rd North, Mississauga, Ontario, Canada
[2]Toronto Rehabilitation Institute, 550 University Ave, Ontario, Canada

1. INTRODUCTION

Spoken communication involves integrating what is said (lexical information) and how it is said (prosodic information). Emotion can be conveyed through both lexical and prosodic cues, and the ability to understand emotion in speech is important for effective communication. Researchers have examined how both healthy and clinical populations understand affective prosody using speech stimuli ranging from monosyllables to full sentences. Unlike in the visual domain, where a select number of well-validated sets of stimuli (e.g., Ekman faces [1], IAPS photos [2]) are used across many studies, in the auditory domain researchers have typically created their own sets of stimuli. As a result, numerous inconsistencies that are likely related to the semantic and lexical properties of the stimuli exist across experiments. One goal of the current programme of research was to create a set of stimuli with well-controlled lexical and semantic properties based on an existing test of speech intelligibility, the Northwestern University Auditory Test- Number 6 (NU-6 [3]), so that the lists of stimuli in the set are balanced for properties such as word frequency as well as word and syllable length.

Experiment 1 provides a description of the actors, recording process, and stimulus selection process used for the creation of the novel set of stimuli, the Toronto Emotional Speech Set (TESS). In Experiment 2 recognition rates for the emotions portrayed in these stimuli were determined for a group of healthy younger listeners.

2. EXPERIMENTS

2.1 Experiment 1

Recording methodology

Two female actors, one younger and one older, were recruited from the community. Respectively, they were 26 and 64 years of age. The actors consented to create voice recordings which would be used as stimuli for research purposes, in educational presentations at scientific or professional conferences or in public education or community presentations. Both actors spoke English as a first language and had clinically normal hearing thresholds in the speech range (see Table 1 for demographic characteristics of the actors). The recording stimuli were the 200 items from the NU-6 test. Each item begins with the same carrier phrase and terminates in a monosyllabic noun (e.g., "Say the word *bean*"). The actors recorded each item to portray seven different emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). These seven emotions were chosen because they were recently used by two groups of researchers to create sets of German and Portuguese sentences [4, 5]. In this way we could extend work on affective prosody understanding for these seven emotions to the English language.

Each actor recorded the stimuli individually in a sound-attenuating booth for approximately 20 hours. During the recording sessions, which typically lasted three to four hours, the majority of the time was spent creating the voice recordings, while approximately 10% of the time was devoted to practicing and fine-tuning each actor's portrayal of each of the emotions. Emotion was blocked such that the actor would finish recording all stimuli in one emotion before moving on to the next. At each session the actor recorded at least one emotion. For each emotion, the experimenter (KD) would request repetitions of specific items until that emotion was judged by the experimenter to have been appropriately conveyed. In total, 2607 stimuli were recorded by the younger actor while 3004 stimuli were recorded by the older actor. Three female undergraduate students (all aged 18 years) with normal hearing listened to the stimuli and identified, for each actor, which token of each NU-6 item they considered to be the most representative for each of the seven emotions. The experimenter used the same procedure to listen to each of the sound files. The sets of ratings were then compared. Rater agreement between the experimenter and at least one of the students as to which token was the best was 80% and 92% for sentences spoken by the younger and the older actor, respectively. In this way, a final set of 2800 TESS stimuli (200 NU-6 items x 7 emotions x 2 actors) was created. In order to facilitate the use of these stimuli in future experiments, they were made available online through the University of Toronto library [6].

Table 1. Demographic information for the two actors in Experiment 1

	Younger actor	Older actor
Age	26 years	64 years
Education (years)	19	18
Vocabulary (out of 20)	12	15
Health (1-4)	4	2
Right ear (PTA dB)	0	6.7
Left ear (PTA dB)	3.3	6.7

2.2 Experiment 2

Methods

Fifty-six undergraduate students at the University of Toronto were tested in this experiment. All participants spoke English as a first language and had clinically normal hearing thresholds from 250 to 8000 Hz (see Table 2 for participant characteristics). Participants listened to stimuli spoken either by the younger or by the older talker. Each participant listened to an equal number of stimuli spoken in each of the seven emotions. The stimuli were presented through a loudspeaker in a sound-attenuating booth at an average presentation level of 70 dBA. In response to each stimulus they used a touch computer screen to indicate which emotion the talker was portraying.

	Younger adults (N=56)
Age (years)	19.7 (.38)
Education (years)	13.5 (.28)
Vocabulary (out of 20)	12.4 (.36)
Health (1-4)	3.5 (.08)
Right ear (PTA dB)	-.3 (.46)
Left ear (PTA dB)	1.0 (.45)

Table 2. Demographic information (Means and SEs) for the participants in Experiment 2

Results

The primary measure of interest was the percentage of correctly recognized emotions. The overall accuracy was 82% (*SD* = 11.08). Stimuli spoken to portray anger and sadness had the highest accuracy while stimuli spoken to portray disgust and pleasant surprise had the lowest accuracy. Results from an ANOVA indicated a significant main effect of emotion, $F(6, 360) = 13.22$, $p < .01$, but no significant main effect of talker, $F(1, 30) = 2.33$, $p = .14$, and no significant emotion by talker interaction, $F(6, 360) = 2.07$, $p = .08$. These null results suggest that both talkers portrayed the different emotions in a similarly effective manner, as can be seen in Figure 1.

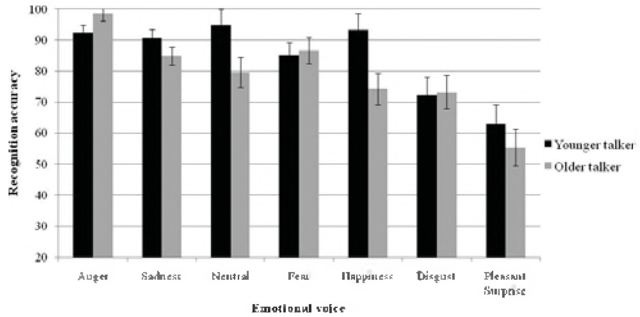

Figure 1. Mean recognition accuracy for participants in Experiment 2 plotted by talker and by the emotion portrayed

Discussion

The results from this experiment indicate that participants were able to recognize the emotions portrayed in the TESS stimuli with very good accuracy. The accuracy rate of 82% was almost six times greater than chance and higher than the 55-65% level described in recent reviews of studies in this field that used sentences with similar emotions [7, 8]. Furthermore, the lack of a main effect of talker indicates that the two actors created highly recognizable portrayals of the seven different emotions.

Although overall recognition of the emotions was high, there were significant differences in the accuracy with which some emotions were recognized. Consistent with previous findings [9, 10], angry and sad emotions had the highest recognition rates overall.

This is the first experiment to examine how well listeners can recognize the emotions portrayed in the Toronto Emotional Speech Set. Future studies should attempt to replicate these findings by using the TESS with healthy good-hearing younger listeners and extend this work to include older adults, participants with poor hearing, and those who suffer from cognitive difficulties.

REFERENCES

[1] Ekman, P., & Friesen, W. V. (1976). *Pictures of Facial Affect.* Palo Alto, CA: Consulting Psychologists Press.

[2] Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1999). *International affective picture system (IAPS): Instruction manual and affective ratings.* Technical Report A-4, The Centre for Research in Physiology. Gainesville, FL: University of Florida.

[3] Tillman, T. W., & Carhart, R. (1966). An expanded test for speech discrimination utilizing CNC monosyllabic words: Northwestern University Auditory Test No. 6. Technical report no. SAM-TR-66-135. San Antonio, TX: USAF School of Aerospace Medicine, Brooks Air Force Base.

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

[5] 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.

[6] Dupuis, K., & Pichora-Fuller, M. K. (2010). Toronto Emotional Speech Set (TESS). Available from https://tspace.library.utoronto.ca/handle/1807/24487

[7] Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40,* 227-256

[8] Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. In R. J. Davidson, K. R., Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 433-456). Oxford: Oxford University Press.

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

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