T-RES: TEST OF RATING OF EMOTIONS IN SPEECH: INTERACTION OF AFFECTIVE CUES EXPRESSED IN LEXICAL CONTENT AND PROSODY OF SPOKEN SENTENCES

Ben-David, B. M.[1,2,3], Multani, N.[1,2,4], Durham, N, A-M.[1,2,4], Rudzicz, F.[1,5] and Van Lieshout, P. H. H. M.[1,2,3,4,6]

[1]*Oral Dynamics Lab, Department of Speech-Language Pathology, University of Toronto, Canada*

[2]*Research, Toronto Rehabilitation Institute, Canada*

[3]*Department of Psychology, University of Toronto Mississauga, Canada*

[4]*Graduate Department of Rehabilitation Studies, University of Toronto, Canada*

[5]*Department of Occupational Science and Occupational Therapy, University of Toronto, Canada*

[5]*Institute of Biomaterials and Biomedical Engineering, University of Toronto, Canada.*

E-mail: boaz.ben.david@utoronto.ca; Homepage: www.boazbendavid.org

Abstract

When the sentence 'I am so happy' is spoken in an angry prosody (i.e., tone of voice), is it interpreted as happiness or anger? This portrays the complex interaction of the two dimensions that convey emotion in speech: lexical content and prosody. How is each dimension weighed perceptually? Is the rating of one impacted by the other? We present a comprehensive paradigm that examines the interplay between these two dimensions. Fifty lexical sentences, reliably associated with one of five emotions (anger, fear, sadness, happiness and neutral), were recorded spoken in five distinct prosodies. The stimulus-set was comprised of two samples for each combination of dimensions, creating congruent and incongruent pairings in addition to neutral spoken sentences. Sixty-four participants were asked to rate the degree to which each spoken sentence expressed each emotion in the lexical content, prosody or both. Results show that "processing strategies" for integrating lexical and prosodic dimensions were dependent on the nature of the rated emotion.

The identification of emotions in spoken language is part and parcel of spoken communication. To partake effectively in social dialogue, it is essential to identify, understand and respond appropriately to the emotion conveyed in speech. Emotions in speech are conveyed via two auditory dimensions: the lexical content of the words and the prosody (i.e., the tone of speech). Consider a talk show host saying 'I am so happy' on the radio with an angry tone. The sentence conveys an emotion of happiness in its lexical meaning, but the tone of voice conveys anger. What emotion is this person trying to convey to the listeners? How would the listener combine the two dimensions? Would the listeners' interpretation change if the speaker conveyed the same emotion in both dimensions (e.g., 'I hate you' in an angry prosody)? This example portrays the complex interaction of these two dimensions. Given its central role in communication, it is surprising to find only a limited number of studies examining this interplay between lexical and prosodic expression of emotions (see Zupan, Neumann, Babbage, & Willer, 2009).

This gap in the literature may be explained by the complexity of mimicking this process in the lab, while controlling for the possible influence of other factors. In a typical spoken language study, participants are asked to identify the emotion expressed in spoken sentences, or to judge whether two sentences convey the same emotion. However, the existing

tools (e.g., the Florida Affect Battery, Bowers, Blonder, & Heilman, 1999) are limited in their ability to reveal the full complexity of the interaction of emotions conveyed in both lexical content and prosody. First, the lexical sentences used in these tests are not equated for their linguistic characteristics (e.g., average word frequency, the number of syllables, phonologic neighborhood density) across the emotional categories and are not validated for their emotional content (i.e., they were not rated on their lexical content separately in a written format). These linguistic characteristics were found to impact the time-course of cognitive processes, as shown in the literature on the emotional Stroop paradigm (Larsen, Mercer & Balota, 2006; see a discussion in Ben-David, Van Lieshout, & Leszcz, 2011). Second, these tests present unequal combinations of emotions conveyed along the two dimensions of prosody and lexical content, presenting a different set size of emotions along each dimension. For example, the FAB presents only two different prosodies for each sentence, even though lexical sentences present more than two emotions. This limits the analysis of dimensional interaction to an analysis of congruent and incongruent stimuli. Moreover, unequal combination of dimensions was found to bias responses in an equivalent test (see a discussion in Melara & Algom, 2003). We further note that these tests lack a baseline condition (e.g., a neutral lexical sentence spoken with varying prosodies) that can serve to gauge performance on one dimension, without the possible interference of the other. Finally, the standard tests use a forced-choice response, where the listener is asked, in each trial, to choose which one of the possible emotions best describes the utterance. We maintain that this restrictive mode of response provides a limited vantage point, as no information can be obtained on the options rejected by the listener. For example, in a forced-choice paradigm the listener may respond "sad," whereas in a rating paradigm, she/he may rate other emotions as possible competitors.

In the current study, we present the Test for Recognition of Emotions in Speech (T-RES), a novel tool to assess listeners' ability to identify emotion in spoken language conveyed either in prosody, lexical content, or both. This test provides a wholesale approach in amending the shortcomings of the previous tests, making for a tool that can uncover the complex interaction of the emotional dimensions in speech.

Method

Participants

Sixty-four young adults (mean age = 19 years, *SD* = 1.8 years) recruited from the undergraduate population at the University of Toronto Mississauga participated in this study. They received either a course credit or $10/hour for their participation. All participants were native English speakers as assessed by a self-report and achieved a minimum score of 9/20 on the Mill Hill Vocabulary Test, corresponding to normal vocabulary levels for native-English speakers (*M* =12.3/20, *SD* = 2.0/20). All participants had pure-tone air-conduction thresholds within clinically normal limits from 0.25 to 3 kHz in both ears (≤ 20 dB HL).

Stimuli

We used a set of 50 linguistically equated sentences, each of which were reliably associated with one particular emotion only (10 in each of the following categories: happiness, anger, sadness and fear) or no emotion at all (10 neutral sentences) taken from Ben-David et al. (2011). An actress (native English speaker) recorded these sentences in each of the five tested prosodies. In a pretest, a small panel of listeners selected the top 50 recorded sentences based on the perceived quality of their prosodic information. Thus, in the T-RES each combination of emotions in the two dimensions is equally represented. Note, to shorten

the test, we have removed one of the trials in which a neutral sentence was recorded in a neutral prosody.

Procedure

Each participant took part in three different rating tasks: combined-dimensions rating, prosody rating, and lexical-content rating. In each task, we used a subset of 25 spoken sentences (comprising every combination of the two dimensions) - each one presented four times in four experimental blocks. Half of the 50 spoken sentences were used in the combined-dimension rating (= presented four times) and the other half were used in both the prosody and lexical-content rating tasks (= presented eight times). Spoken sentences were presented via headphones to a participant seated in a sound attenuated booth. In each block, the participant was asked to rate each spoken sentence on four 6-point Likert scales, relating to how much they agree that speaker was ____ (happy/ angry/ sad/ fearful). *Combined-dimensions (prosodic and lexical) rating.* In this task, the participant was asked to rate the sentence as a whole using the information in both dimensions, as if he/she were listening to a person over the phone. *Prosody rating.* The participant is asked to selectively attend only to the emotions conveyed by the prosody, ignoring the information provided by the lexical content. Conversely, in the *Lexical-content rating* task, the participant was asked to focus exclusively on the lexical content, ignoring the prosody.

To control for the possible effects of practice, participants were randomly assigned to one of eight experimental groups (eight participants in each). For all groups, the experiment started with the combined-dimensions task. For half of the participants, the second task was prosody rating and the third was the lexical-content rating, whereas for the other half this order was reversed. The order of the four emotional rating scales was counterbalanced across participants using a Latin square procedure. This counterbalancing procedure, when combined with the counterbalancing of the order of the tasks, generated the eight groups.

Results

The 12 panels of Figure 1 present the average rating of each combination of dimensions in all four rating scales, in the three rating tasks. Columns present the different prosodies used and rows present the different lexical contents. Cells are color-coded to facilitate reading, with different shades of grey varying from a filled (black) cell, indicating the maximum score of 6 (the highest agreement that the spoken sentence conveyed a given emotion) to an empty (white) cell, indicating the minimum score of 1 (the lowest agreement that the spoken sentence conveyed a given emotion). Emotional rating scales are indicated by an italic font for sentences that conveyed the rated emotion in either the lexical content or the prosody. For example, when the rated emotion was Anger (Panels A, B and C), the cells in the column corresponding to angry prosody and in the row corresponding to an angry lexical content are italicized. We use a white font for averages that indicate a score that is higher than (or equal to) the mid-range (3.5).

The scope of this paper does not allow us to present the full analysis of the results. We therefore focus on two specific cases. First, we discuss the generic (and surprising) failures of selective attention in the separate ratings of the prosody or the lexical dimensions. Second, we compare two rating scales, anger and fear, showcasing two separate strategies for forming combined-dimension rating. In the first combined-dimension ratings were mainly based on one dimension (prosody), whereas in the latter a more integrative form of weighing dimensions was used.

PROSODY PROSODY PROSODY

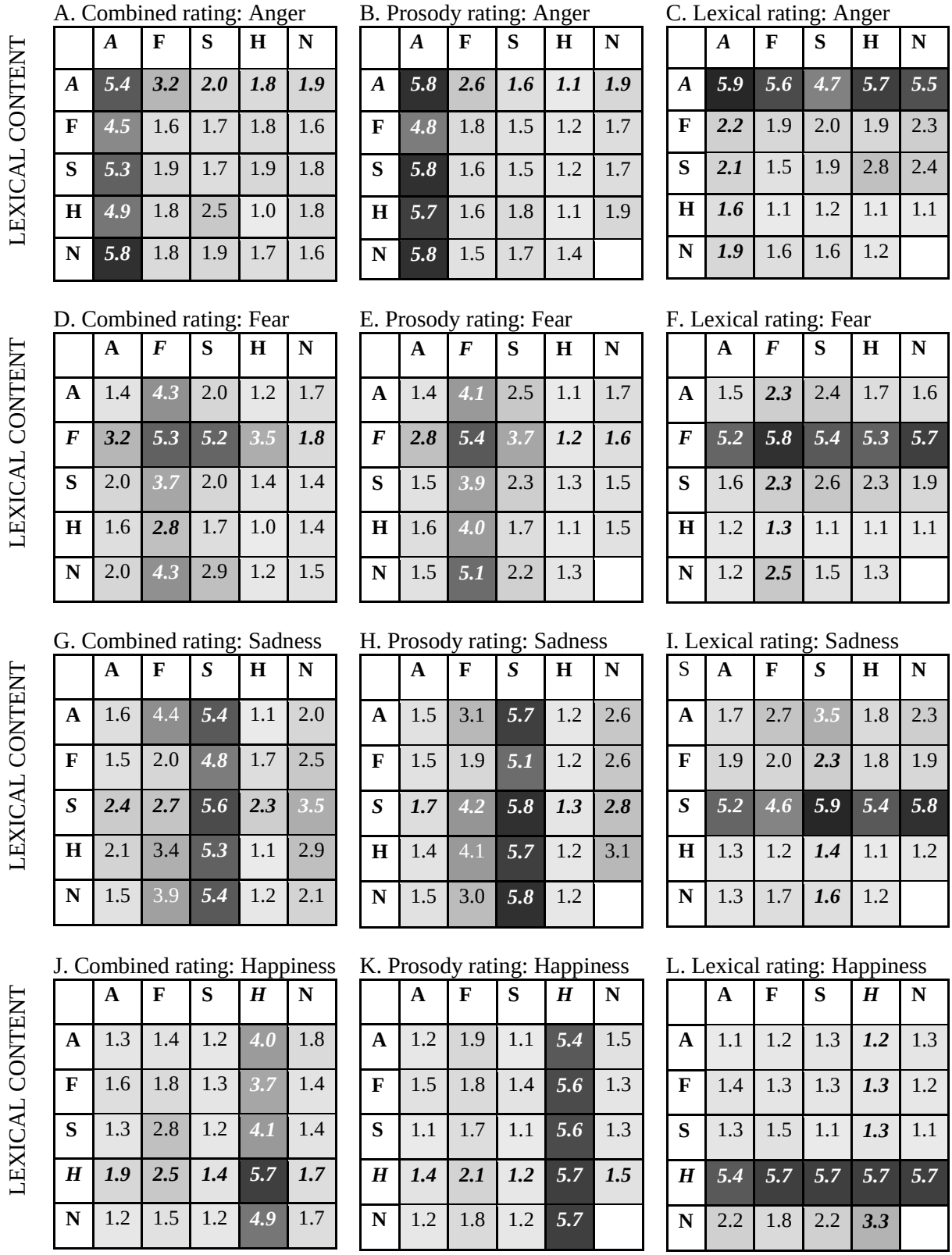

Figure 1. Average ratings of emotions. In each panel, a row presents sentences that share the same emotional lexical content, and a column presents sentences that were spoken in the same emotional prosody. Emotions are: A – anger, F – fear, S – sad, H – happy and N – neutral. Combined-dimension ratings, prosody ratings and lexical ratings are displayed in the left-most, middle and right-most panels, respectively.

To test the impact of the prosodic information on lexical ratings, we conducted a confusion matrix analysis. We aggregated the scores in each of the five columns for each of the matrices related to lexical rating (right-most panels). Next, we conducted 16 paired t-tests comparing the lexical ratings of a particular emotion for sentences spoken in the respective prosody, with ratings of this emotion for sentences spoken in the other four emotions. For example, in Panel F, we compared the ratings of the extent of fear expressed in the lexical content of sentences spoken in a fearful prosody with sentences spoken in a sad, angry, happy and neutral prosody. In a similar fashion (averaging rows instead of columns), we tested the impact of the lexical information on prosody ratings in the middle Panels. For example, in Panel E, presenting prosody rating of fearful emotion, we compared the average rating of sentences with a lexical content conveying fear with sentences whose lexical content conveyed a sad, angry, happy and a neutral emotion.

The tests were conclusive -- we found a significant ($p < .05$) impact of the irrelevant dimension in 15/16 of the tests for the impact of lexical content on prosody ratings, and in 13/16 of the tests for the impact of prosody on lexical rating. These results indicate that listeners were not immune to the information presented in the to-be-ignored dimension, whether it was the lexical content or the prosody.

Integration of Dimensions

First, we replicated the confusion matrix analysis described above for the combined-dimensions rating task. All tests (16/16) were found significant ($p < .0001$) indicating that participants were indeed using both the lexical and the prosodic dimension when they rated the spoken sentences as a whole. Next, we compared the combined-dimensions rating of the anger emotion, Panel A, with combined-dimension ratings of fear emotion, Panel D. It appears that anger ratings were mostly based on the prosodic information (compare Panel A with Panel B), whereas combined-dimensions ratings of fear were impacted to some extent by both the lexical (Panel F) and prosodic information (Panel E). To test this statistically, we deducted the marginal averages of the fear row (lexical content) from the fear column (prosody) for fear combined-dimensions ratings (excluding the congruent cell). Similarly, we deducted the marginal averages of the anger row (lexical content) from the anger column (prosody) for anger combined-dimensions ratings.

The prosody advantage for anger ratings ($M = 2.90$) was found to be significantly larger than the prosody advantage for the fear ratings ($M = .34$, $p < 1*10^{-19}$). We note that for both ratings, the average for prosody sentences was higher than for the lexical sentences. Yet while this effect was highly significant for anger ratings ($p < 1*10^{-28}$), significance was relatively marginal for fear ratings ($p = .057$).

Discussion

We presented a novel comprehensive tool, the T-RES that can identify the separate and combined impact of prosody and lexical content on identification of emotions in speech. With a large group of healthy young control listeners, T-RES revealed the complexity of the interaction of the two dimensions, lexical content and prosody, in spoken language. Rating of some emotions reflect processing of the prosody as the main source of information (anger), whereas others rely on the combination of the processing of both lexical content and prosody of the spoken sentences (fear). Failures of selective-attention appear to impact ratings, even when the listener was specifically asked to attend to one dimension, whether it is the prosody or the lexical content of the spoken sentence.

Since difficulty in identifying emotions in speech has significant impacts on the rehabilitation of several patient populations (e.g., persons with brain injury; Bornhofen et al, 2008), we believe this test can be useful as a tool to improve the reliability of the assessment and the rehabilitation of communication skills in these populations. Indeed in an ongoing study conducted at the Toronto Rehabilitation Institute, we already identified idiosyncratic patterns of emotion-rating for individual patients that are highly distinct from the averages presented here.

Acknowledgments

B. M. Ben-David was partially supported by a grant from the Ontario Neurotrauma Foundation (2008-ABI-PDF-659). This research was undertaken, in part, thanks to funding from the Canada Research Chairs program (303712CRC) awarded to P. H. H. M. van Lieshout. We wish to thank the following students for their assistance in constructing the stimuli and collecting the data: Kinza Ali, Bilal Athar and Lihn (LeTruc) Nguyen.

References

Ben-David, B. M., van Lieshout, P. H. H. M., & Leszcz, T. (2011). A resource of validated affective and neutral sentences to assess identification of emotion in spoken language after a brain injury. *Brain injury, 25*(2), 206-220. doi: 10.3109/02699052.2010.536197

Bornhofen, C., & Mcdonald, S. (2008). Emotion perception deficits following traumatic brain injury: A review of the evidence and rationale for intervention. *Journal of the International Neuropsychological Society, 14*(4), 511-525. doi: 10.1017/S1355617708080703

Bowers, D., Blonder, L. X., & Heilman, K. M. (1999). *The Florida affect battery manual-revised*. Retrieved April 2010 from University of Florida, Center for Neuropsychological Studies, Cognitive Neuroscience Laboratory. http://www.neurology.ufl.edu/forms/fab_manua

Larsen, R., Mercer, K., & Balota, D. (2006). Lexical characteristics of words used in emotional Stroop experiments. *Emotion, 6*(1), 62-72. doi: 10.1037/1528-3542.6.1.62

Melara, R. D., & Algom, D. (2003) Driven by information: A tectonic theory of Stroop effects. *Psychological Review, 110*(3), 422-471.

Zupan, B., Neumann, D., Babbage, D. R., & Willer, B. (2009). The importance of vocal affect to bimodal processing of emotion: Implications for individuals with traumatic brain injury. *Journal of Communication Disorders 42*(1), 1-17. doi: 10.1016/j.jcomdis.2008.06.001