

Individual Variability in the Discrimination of Audiovisual Spontaneous vs. Acted Expressive Speech

Nicolas Audibert¹, Véronique Aubergé^{1,2} & Albert Rilliard³

¹ Gipsa-lab Speech & Cognition Dept (formerly ICP), CNRS UMR 5216/INPG/UJF/Stendhal, Grenoble, France

² LIG/GETALP, CNRS UMR 5217, Grenoble, France

³ LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France

{Nicolas.Audibert, Veronique.Auberge}@gipsa-lab.inpg.fr;

Albert.Rilliard@limsi.fr

INTRODUCTION

Though a very large majority of studies focused on expressive speech have used acting as a convenient method for obtaining utterances with the same phonetic contents expressing various affects, the reliability of such material for the modeling of vocal expressions of spontaneous expressive speech has come into debate during the last decade (Campbell, 2000). Such reservations have incited a growing number of speech and affects researchers to look for spontaneous corpora (see for instance Campbell, 2004). On the other hand, collection of acted corpora has gone deeper into the acting process with elicitation-based approaches designed to obtain more natural acted speech (Enos & Hirschberg, 2006).

However very few studies have focused on differences in perception or production of acted vs. spontaneous expressive speech. The ability of naïve listeners to discriminate simulated vs. spontaneous audiovisual amusement was shown with a large inter-listeners variability independently of the actors' skills (Aubergé & Cathiard, 2003). More recently, a higher perceived emotional intensity was found in expressions acted by naïve Dutch speakers than in expressions spontaneously uttered by the same speakers after emotional induction (Wilting *et al.*, 2006).

We hypothesize (Aubergé, 2002) that expressions of affects are cognitively distinguished according to the way they are controlled by the speaker: voluntarily vs. involuntarily. The main questions asked in this study are whether utterances expressing the same emotional values through acting (voluntary) vs. spontaneous expressions (involuntary) can be discriminated by human listeners, and whether all listeners have similar competences for accessing such cues.

A CORPUS OF ACTED AND SPONTANEOUS EXPRESSIVE SPEECH

The spontaneous part of the multimodal expressive corpus E-Wiz/Sound Teacher (Aubergé *et al.*, 2004) was recorded using a Wizard of Oz technique, and self-labeled for affective contents by the subjects themselves. 17 subjects were recruited with the pretext of participating in the evaluation of a voice-recognition-based language-learning software before its commercialization, in which subjects' performances were manipulated to induce both positive and negative affects. 7 subjects out of 17 were also actors used to practice improvisation theater and/or street acting, and trained to rely on past emotional episodes when acting. Those actors were requested immediately after the Wizard of Oz task to produce again the affects they reported to have felt during the experiment on the same utterances as well as the most frequently studied emotions (sadness, anger, fear, disgust, surprise and joy), using their own acting methods. The experimenters insisted that the actors should express the affects felt in the experiment the same way they had been feeling and expressing them.

193 acted and spontaneous utterances expressing affects broadly related to happiness, fear and anger and produced by 6 actors (3 males, 3 females) were validated and rated for emotional intensity in audio-only condition by 15 naïve French-speaking listeners (Laukka *et al.*, 2007). This evaluation showed that emotional intensity was perceived higher for acted utterances vs. spontaneous ones, in

line with Wilting *et al.* (2006). 24 pairs of stimuli produced by 4 of the actors (2 male, 2 female) were retained for the present discrimination task.

PERCEPTIVE EVALUATION AND STATISTICAL ANALYSIS

The 24 selected pairs of stimuli were presented with a latency of 1.5 seconds between both, with 3 presentation conditions: audio only condition (A), visual only condition (V) and audiovisual condition (AV). Stimuli were presented grouped by condition and randomly sorted within each condition, A and V conditions being alternatively chosen as first condition, with AV as last. Each pair was presented in both orders in each condition. After each presentation of a pair the subject was requested to indicate which stimulus was the spontaneous one, using a slider ranging from 'certainly the first one' to 'certainly the second one' and intended to capture both identification and confidence level, similarly to the procedure used by Bänziger (2004). 33 native French subjects (15 male, 18 female, mean age 33.1) took the listening test after having been informed of the context of the corpus recording.

Slider position values were converted into discrimination scores and confidence level. Overall discrimination score was 60.7% in A, 56.3% in V and 69.9% in AV, while mean confidence was 53.7% in A, 60.4% in V and 63.6% in AV. Identification scores and confidence levels were analyzed using repeated measures analyses of variance with listener, speaker, emotion class, presentation condition, utterance length and presentation order as fixed factors. In spite of the quite weak correlation between discrimination and confidence ($r=.408$ in A, $.690$ in V, $.583$ in AV, $.622$ overall) most of the significant effects were found to be the same on identification and confidence scores.

The most important statistical effect observed, in line with the results obtained by Aubergé and Cathiard (2003) on amusement, was a strong listener effect on the discrimination of acted vs. spontaneous utterances ($p<.001$): discrimination skills appear to be highly variable, with individual scores ranging from 32.7% to 80.6% of correctly classified pairs in the present task. 70% of listeners were able to correctly discriminate more than half of the presented pairs in V, while 79% did in A and 85% did in AV condition (85% overall). Although the productions of one of the male actors were significantly better discriminated than those of the 3 others, individual discrimination performances were found to be consistent across productions of different speakers ($\alpha=.867$).

Discrimination was significantly better in AV when compared to other conditions ($p<.001$), while the gain between V and A was non-significant. However there was a large inter-speaker variability, without a significant gain from A to AV for 2 actors out of 4 (1 male, 1 female).

No significant effect of the emotional category nor of the utterance length were found, while the effect of the presentation order was significant only in V for the production of 2 speakers ($p<.01$). Though both discrimination and confidence were better for female listeners, no significant effect of the listeners' gender was found.

DISCRIMINATION PERFORMANCES VS. PERCEIVED EMOTIONAL INTENSITY

Partial correlations between perceived emotional intensities rated in Laukka *et al.* (2007) and discrimination scores (respectively confidence) were calculated for each presentation condition. Although correlation for other conditions remain low for both discrimination and confidence (the highest being $r=.335$ in AV for discrimination), it reaches $r=.745$ for the discrimination in A. As ratings of perceived emotional intensity were obtained in audio-only condition, it is not surprising to find higher correlations in this condition. A closer look at results for individual pairs reveals that, although pairs with the largest difference in perceived intensity are among the best discriminated, pointing this feature as a strong cue to such discrimination, pairs with similar ratings of perceived intensity can also be well discriminated. The perceived emotional intensity should therefore not be considered as the sole cue to discrimination.

CONCLUSION

Those results suggest that naïve listeners are able to discriminate acted vs. spontaneous multimodal expressions without an effect of the emotion, with large variability across listeners. Although the actors who participated in the corpus recording were certainly not among the best ones, 3 of them out of 4 performed well enough to trick the less competent listeners.

This variability in listeners' ability to discriminate acted vs. spontaneous expressive speech

A complementary perceptive evaluation focused on the perceived emotional intensity in each presentation conditions is currently being carried out in order to study to what extent this feature can account for discrimination of acted vs. spontaneous expressions in other modalities.

REFERENCES

- Aubergé, V., 2002. A Gestalt morphology of prosody directed by functions: the example of a step by step model developed at ICP. *1st International Conference on Speech Prosody*, Aix-en-Provence, France, 151-155.
- Aubergé, V. & Cathiard, M., 2003. Can we hear the prosody of smile? *Speech Communication* 40 (2), 87-97.
- Aubergé, V., Audibert, N. & Rilliard, A., 2004. E-Wiz: A Trapper Protocol for Hunting the Expressive Speech Corpora in Lab. *4th LREC*, Lisbon, Portugal, 179-182.
- Bänziger, T., 2004. *Communication vocale des émotions. Perception de l'expression vocale et attributions émotionnelles*. PhD thesis, University of Geneva.
- Campbell, N., 2000. Databases of Emotional Speech. *ISCA Workshop on Speech and Emotions*, Newcastle, North Ireland, 34-38.
- Campbell N., 2004. Speech & Expression; the value of a longitudinal corpus, *4th LREC*, Lisbonne, Portugal.
- Enos, F. & Hirschberg, J., 2006. A Framework for Eliciting Emotional Speech: Capitalizing on the Actor's Process. *1st International Workshop on Corpora for Research on Emotion and Affect*, Genoa, Italy, 6-10.
- Laukka, P., Audibert, N. & Aubergé, V., 2007. Graded structure in vocal expression of emotion: What is meant by "prototypical expressions"? *1st International Workshop on Paralinguistic Speech*, Saarbrücken, Germany, 1-4.
- Wilting, J., Kraemer, E. & Swerts, M., 2006. Real vs. acted emotional speech. *INTERSPEECH 2006-ICSLP* (CD-ROM proceedings).