How we are not equally competent for discriminating acted from spontaneous expressive speech

Nicolas Audibert[1], Véronique Aubergé[1] and Albert Rilliard[2]

[1] Gipsa-lab Speech & Cognition Dept (Institut de la Communication Parlée),
CNRS UMR 5216/Université Stendhal, 38040 Grenoble Cedex 9, France
[2] LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
{Nicolas.Audibert, Veronique.Auberge}@gipsa-lab.inpg.fr; Albert.Rilliard@limsi.fr

Abstract

This paper reports how acted vs. spontaneous expressive speech can be discriminated by human listeners, with various performances depending on the listener (in line with preliminary results for amusement by [3]). The perceptive material was taken from the Sound Teacher/E-Wiz corpus [1], for 4 French-speaking actors trapped in spontaneous expressive monoword utterances, and then acting immediately after, in an acting protocol supposed to be a very convenient for them. Pairs of acted vs. spontaneous stimuli, expressing affective states related to anxiety, irritation and satisfaction, were rated by 33 native French listeners in audio-only, visual-only and audiovisual conditions. In visual-only condition, 70% of listeners were able to identify acted vs. spontaneous pairs over chance level, for 78% in audio-only condition and up to 85% in audio-visual condition. Globally, a highly significant subject effect confirms the hypothesis of a varied affective competence for separating involuntary vs. simulated affects [2]. One feature used by listeners in the acoustic task of discrimination can be the perceived emotional intensity, in accordance with the measurement of this intensity level for the same stimuli from a previous perception experiment by Laukka and al. [9].

1. Introduction

Although the question of the validity of corpora of acted emotional speech for the modeling of affective speech has been debated (see for instance [5]), leading to an increase of the research effort directed towards spontaneous emotional speech, few studies have been comparing the performances of acted vs. spontaneous speech to our knowledge. Aubergé et al. [3] proposed that acted vs. spontaneous amusement could be discriminated, judges discrimination competences being highly variable independently of the speaker's acting skills. More recently, Wilting et al. [12] recorded naïve Dutch participants without particular acting skills while inducting positive and negative moods using the Velten procedure, prior to asking them to produce the same utterances while simulating similar moods. Though acted and spontaneous utterances were not directly compared, a perception experiment in visual condition showed that acted expressions were rated as more intense than spontaneous ones.

Such findings are coherent with a strong hypothesis that we proposed for the processing of affects (after Fonagy [8] and Scherer [11]): affects are cognitively distinguished following two ways of control by the speaker: voluntary vs. involuntary, and not as a function of the affective information carried by the expressions. In this view [2], authentic emotions are performed by involuntary control (the "push effect" in Scherer's model [11]). The speaker is also able to reproduce the expressions of his experienced emotions outside the body loop described by Damasio [7] and through a voluntary control, that are the social affects [2]. That means than a same value of affect can be processed voluntarily or not. The voluntary performance, that informs the speech acts generated by the speaker, can be very spontaneous and sincere, but it does not reflect the same processing than the authentic emotion one. We claim [2] that this competence is central in the communication processing and is the more productive into expressivity in interactions, and is not implied in the same timing processing (voluntary/social expressions would be anchored in the time of the language organization). Cultures and languages have developed large specific scales of such voluntary controlled affect, over the basic reproduction of emotions, that we call attitudes.

We make the hypothesis that this competence of reproducing expressions is used by the actors, on speech acts given by an author, especially when they belong to an acting field (1) devoted to simulate to be very authentic in the given acting story context (2) based on method using the memory of previously felt emotions. That is precisely the field of the actors participating to this experiment.

The main question asked in the presented experiment is whether expressions with same emotional values through voluntary vs. involuntary control, i.e. in this case through a simulation by acting vs. involuntary felt emotions, can be discriminated by human listeners, and whether all humans have a similar competence for accessing these cues.

Moreover Aubergé et al [3] showed that the acoustic information is integrated to the visual decoding of affective values, even when the face carries strong affective information. More generally it has been shown that emotional expressions must be considered as multi-modal processing [10]. The study presented in this paper thus focuses on multimodal expressions of affective speech, trying to separate the information carried by different modalities, even though the face also carries information about speech and this information consequently cannot be considered as additive across modalities.

2. Acting vs. spontaneous speech collection

The French expressive corpus E-Wiz [1] was recorded using the Wizard of Oz technique, in which the subject is convinced to be interacting with a complex person-machine interface while the apparent behavior of the application is remote-controlled by the wizard. Subjects were asked to participate in the testing prior to its commercialization of a so-called voice-recognition-based language-learning software. In this task the subjects had to interact with the system using a command language composed of the French monosyllabic color names [bʁik], [ʒon], [ʁuʒ], [sabl] and [vɛʁ] and the command [paʒsɥivɑ̃t] (*next page*). The performances of the 17 subjects participating in the experiment were manipulated to induce positive then negative emotions, and the affects expressed were labeled by the subjects themselves from the video recording, as a first labeling step before perceptive validation. A particular protocol was set up for the 7 subjects who were also actors: those subjects 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 acting methods. The experimenters insisted that the actors should express the affects felt in the experiment the same way they had been feeling them just before. The actors recruited for this task were practicing improvisation theater and/or street acting, and used past felt emotions as a basis for expressing emotions, as described in [6]. All of them reported the experimental set-up as optimal for being in good acting conditions with regards to their acting habits.

A first experiment using both acted and spontaneous utterances from the E-Wiz corpus in audio-only condition, and focusing on the typicality of vocal expressions of emotion, was conducted by Laukka et al [9]. In this experiment, 193 acted and spontaneous utterances produced by 6 actors (3 males, 3 females) were validated and rated for emotional intensity in a pre-test, showing a higher perceived emotional intensity for acted utterances vs. spontaneous ones, in line with results obtained by Wilting et al. [12] in visual-only condition. We present in this paper the results of a discrimination task between acted and spontaneous utterances, based on productions of the speakers evaluated in this pre-test.

3. Experimental protocol

The 24 selected pairs of stimuli were presented to subjects with a latency of 1.5 seconds between both, with 3 presentation conditions: audio only (A), visual only (V) and audiovisual (AV). Stimuli were presented grouped by condition and randomly sorted within each condition, AV condition being always the last one while A and V conditions were alternatively chosen as first condition. Each pair was presented twice in each condition with the spontaneous utterance in first or second position to compensate for a possible effect of the presentation order.

After each presentation of a pair the subject was requested to indicate which stimulus he considered to be the spontaneous one, using a slider ranging from 'certainly the first one' to 'certainly the second one', which initial position was set to the middle. This slider was intended to capture both identification and confidence level, similarly to the procedure used in [4]. Answers could be validated only after the slider had been moved. The presentation of stimuli and the recording of subjects' answers were automated through a user interface developed on purpose with the Revolution software. Subjects were explained the main goals of the experiment as well as the context of the corpus recording prior to the actual beginning of the task. 33 native French subjects (15 male, 18 female, mean age 33.1) without known hearing problems took the listening test, which lasted 25 minutes in average.

4. Analysis

4.1. Statistical method

Slider position values were converted into identification scores (right vs. wrong answers) according to the direction in which the slider had been moved, and into a confidence level according to the distance from the slider position to the initial position. The mean identification score for each pair was only moderately correlated to the confidence level (r=.408 in audio-only condition, r=.690 in visual-only condition, r=.583 in audiovisual condition, r=.622 overall). Identification scores for different presentation conditions and speakers are summarized in figure 1.

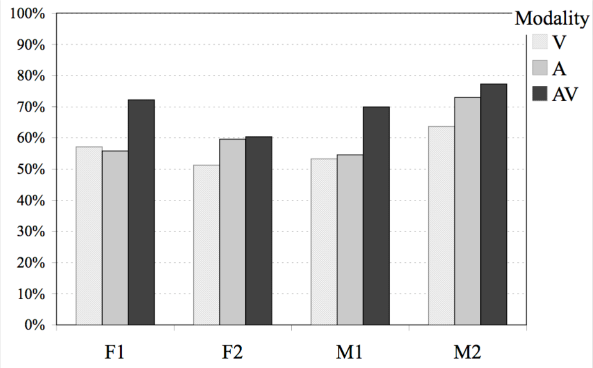

Figure 1: *Overall results according to speakers and presentation conditions*

Identification scores and confidence levels were analyzed using repeated measures analyses of variance (ANOVA) with listener, speaker, emotion class, presentation condition, utterance length and presentation order as fixed factors. As most of significant effects were found to be the same on identification and confidence scores, only a few remarkable effects on confidence are reported here.

4.2. Differences in listeners' discrimination competences

A strong listener effect was found on identification of spontaneous utterances (F(1,31)=801.58, p<.001), in line with results of Aubergé et al. [3] on amusement. As a matter of fact, identification scores of different listeners range from 32.7% to 80.6% of correctly classified pairs. Though the listener effect on the rated confidence level was also highly significant (F(1,31)=220.23, p<.001), strong conclusions should not be drawn from this result as it might more reflect different strategies in the use of the slider than differences in subjects' abilities.

Although listeners' competences for discriminating acted vs. spontaneous expressions were highly variable, they did not appear to show preferences for particular speakers independently of their acting skills. As a matter of fact, Cronbach's alpha value for individual discrimination performances on different speakers' production was quite high (alpha=0.8671), indicating that listeners competences were consistent across different speakers.

Figure 2 presents the distribution of listeners' identification scores for each presentation condition and overall. As it can be observed from this chart, 70% of listeners were able to correctly discriminate more than half of the presented pairs in visual-only condition, while 79% did in audio–only condition and 85% did in audiovisual condition (85% overall).

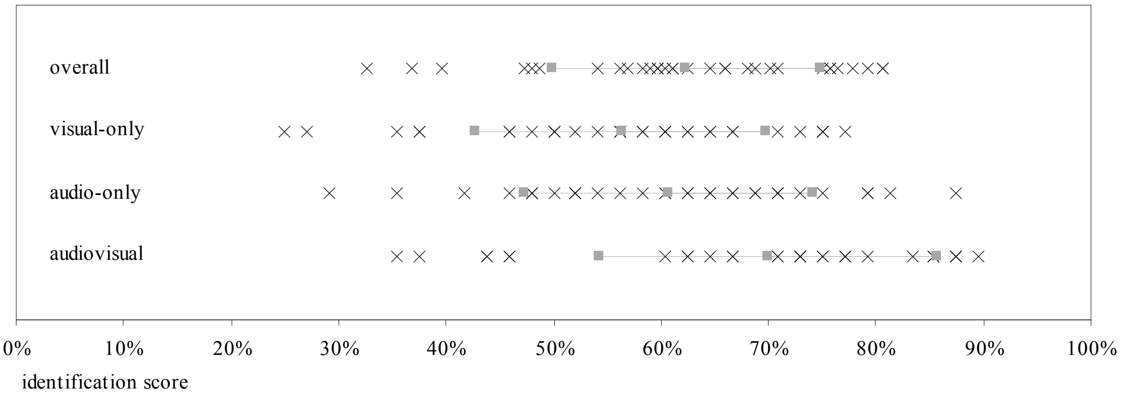

Figure 2: *Distribution of overall and per presentation condition listeners' identification scores. Grey boxes and solid lines indicate mean and standard deviation for each condition.*

4.3. Discrimination across modalities

A significant main effect of the presentation condition on identification scores was found (F(2,62)=21.33, p<.001). Contrasts between conditions show a significant gain of discrimination for the audiovisual condition when compared to the audio-only and visual-only (p<.001 for both contrasts) conditions, while the difference in identification scores between audio-only and visual only conditions was non-significant. However this advantage of audiovisual condition against audio-only and visual-only conditions was not constant across different speakers, as illustrated by figure 1: the effect of condition was indeed non-significant for speaker F2, and the identification increase from audio-only to audiovisual condition was non-significant for speaker M2.

4.4. Speaker effect

A significant main effect of the speaker (F(3,93)=16.05, p<.001) was also observed. Only spontaneous utterances of speaker M2 were significantly better identified than those of all other speakers (p<.001 for all 3 contrasts), indicating that this speaker was less successful than the other actors in pretending that he was actually expressing spontaneous affects. On the other hand, all 3 other speakers' productions were discriminated with similar scores, although a large part of listeners reported to have considered the discrimination task as more difficult for speakers F2 and M1 than for the two other speakers. This intuition of listeners was illustrated by the fact that, whereas speaker M2 also received the highest confidence ratings, confidence ratings attributed to utterances of speaker F1 were significantly higher than those of speakers F2 and M1 (p<.001 for both contrasts). This speaker effect was stronger in audio-only and audiovisual conditions (p<.001 for both) than in visual-only condition (p<.05).

4.5. Other effects

No overall effect of the emotion class was found, suggesting that subjects have similar abilities in identifying a spontaneous vs. an acted expression whatever the emotion expressed. The effect of length was just significant, with utterances of [paʒsɥivãt] slightly better discriminated than monosyllabic utterances (F(1,1)=6.33, p<.05).

The order of presentation of the stimuli in the pairs (spontaneous then acted vs. acted then spontaneous) was globally significant (F(1,1)=8.32, p<.01), with a higher discrimination score for pairs in which the spontaneous stimulus was presented first. This effect is however only significant in the visual only condition (p<.001), and related to the speakers production: the effect is indeed significant (p<.01) for only two of them (speaker M2 who was the best recognized, i.e. the less good actor, and speaker F2 for whom the identification was the lowest, though not significantly different from other actors), which are also those for whom audiovisual identification scores were not significantly better when compared to audio-only. We did not yet proceed to a complete objective analysis of the visual expressions, but observable gestures amplitudes of those two actors are obviously larger than those of the two others. Moreover both of them moved almost systematically the head downwards in spontaneous speech. A possible explanation for this presentation order effect could be ecological strongly informative reference given by the spontaneous stimulus presented first, for a better discriminative comparison with the second, acted, stimulus.

Though both identification and confidence were higher for female than for male listeners, especially in audio-only condition, those differences were found to be non-significant. Although statistical significance cannot be calculated in that case, a particular result is worth being noted: for two male subjects performing better than the average in audio-only condition but worse than the average on visual condition, the visual information seem to have largely lowered discrimination performances in audiovisual condition. The identification score of those two subjects was indeed more than 20% lower in audiovisual condition than in audio-only condition.

Correlations between duration differences of the spontaneous vs. acted stimuli presented in each pair (ranging from -480 to 760 ms) and averaged identification and confidence scores were calculated in order to look for a possible account of this difference in the discrimination performances. However those correlations were very low (r=.047 for identification, r=.037 for confidence), suggesting that this information was not used as a cue for discrimination.

4.6. Perceived emotional intensities and discrimination

The partial correlations for different presentation conditions between identification scores or confidence ratings, and differences of perceived emotional intensities in the pair from [9] are presented in table 2 with the overall correlations. The number of pairs for which those correlations can be calculated does not allow to draw strong conclusions from these values. However the stronger correlation in audio-only condition, especially between identification scores and perceived intensity differences, suggests that the difference in perceived emotional intensity may at least partly account for the discrimination between spontaneous and acted utterances. As emotional intensity ratings were given from presentations in audio-only condition, it is not a surprising result to find higher correlations for this condition.

Table 1: *Correlations between difference in perceived emotional intensity in the pair (extracted from [9]) and identification scores or confidence level*

condition	A	V	AV	overall
identification	r=.745	r=.131	r=.335	r=.415
confidence	r=.402	r=.147	r=.283	r=.250

Although pairs with the highest difference in perceived emotional intensity appear to be among the best discriminated pairs, suggesting that perceived emotional intensity might be a strong cue for discrimination when usable, listeners did not rely only on that feature for discriminating spontaneous vs. acted emotions: as a matter of fact, among the 3 pairs for which the difference in emotional intensity between spontaneous and acted was the weakest, ranging from -4.6% to 4,6%, only the expressions of irritation produced by speaker M1 on monosyllabic utterances were poorly discriminated (correctly discriminated by only 42,4% of listeners in audio-only condition) while the expressions of irritation of speaker F1 on [paʒsɥivɑ̃t], evaluated with the same emotional intensity, were correctly discriminated by 62,1% of listeners. On the other hand expressions of irritation on [paʒsɥivɑ̃t] by speaker F2, for which the acted expression had been evaluated as 16.6% more intense than the spontaneous one, were among the most poorly discriminated in audio-only condition (correctly discriminated by only 40,9% of listeners).

5. Conclusion

The presented results suggest that listeners are globally able to discriminate acted vs. spontaneous multimodal expressions, without an effect of the emotion (the only three kinds of emotional information evaluated are indeed quite balanced in terms of activation and valence), and with a strong listener effect. The perceived emotional intensity, previously measured between acted and spontaneous in auditory condition [9] (and in visual condition by others authors on other data [12]), might be an artifact explaining part of the discrimination scores. Though surely a major bias as pointed out by Aubergé et al. [2], differences in perceived emotional intensity can definitely not account for the whole variability.

Even if the chosen actors were certainly not among the ones recognized as the best, three of them out of four could trick the less competent listeners, and are typically the kind of actors who are chosen for recording emotional databases. The acted speech, commonly used as a reference for studies on involuntary emotions, could be considered carefully knowing the human discrimination ability. It is an open and exciting question to verify if there is a variable competence of human for the identification (and not only discrimination) of simulation (that may be related to the affective quotient), that is, in our proposals, the modal processing used in interaction by a speaker expressing his attitude, whatever his sincerity (including the reproduction of involuntary emotions).

Since the cognitive processing of acted speech cannot be directly related to the cognition of voluntary expressed emotions, i.e. the social affects, a further experiment will be to catch and perceptively compare spontaneous emotions vs. spontaneous attitudes (reduced to emotion values).

The acoustic and visual analysis of the stimuli, according to the perceptive results, is under progress, on the basis of a strong hypothesis on the difference of timing anchoring between involuntary and voluntary (social) emotions [2].

6. Acknowledgment

Many thanks to Christophe Savariaux, Petri Laukka, to the actors, and to the listeners who kindly joined the experiment.

7. References

[1] 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.

[2] Aubergé V. 2002. A Gestalt morphology of prosody directed by functions: the example of a step by step model developed at ICP. *1st Speech Prosody*, Aix-en-Provence, France, 151-155.

[3] Aubergé, V.; Cathiard, M., 2003. Can we hear the prosody of smile? *Speech Communication* 40, *Special issue on Emotional Speech,* 87-97.

[4] Bänziger, T., 2004. *Communication vocale des émotions. Perception de l'expression vocale et attributions émotionnelles*. PhD thesis, University of Geneva.

[5] Campbell, N., 2000. Databases of Emotional Speech. *ISCA Workshop on Speech and Emotions*, Newcastle, North Ireland, 34-38.

[6] Enos, F.; Hirschberg, J., 2006. A Framework for Eliciting Emotional Speech: Capitalizing on the Actor's Process. *WS Corpora for Research on Emotion and Affect*, Genova, Italy, 6-10.

[7] Damasio A. R. 2003. *Looking for Spinoza. Joy, Sorrow an the Feeling Brain.* Orlando:FL/Harcourt.

[8] Fonagy, I. 1983. *La vive voix*. Paris:Payot.

[9] Laukka, P.; Audibert, N.; Aubergé, V., 2007. Graded structure in vocal expression of emotion: What is meant by "prototypical expressions"? *WS Paralinguistic Speech - between models and data*, Saarbrücken, Germany, 1-4.

[10] Scherer, KR & Ellgring, H. 2007. Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns? *Emotion*, 7(1), 158-171.

[11] Scherer K. R. 2001. Appraisal considered as a process of multi-level sequential checking. In *Appraisal processes in emotion: Theory, Methods, Research*, Scherer K. R., Schorr A., & Johnstone T. (eds.), Oxford University Press, 92-120.

[12] Wilting, J.; Krahmer, E.; Swerts, M., 2006. Real vs. acted emotional speech. *9th INTERSPEECH*, Pittsburgh, PA, USA (CD-ROM proceedings).