Evaluating an Authentic Audio-Visual Expressive Speech Corpus

Rilliard Albert, Aubergé Véronique & Audibert Nicolas

Institut de la Communication Parlée, Grenoble, France
E-mail: {rilliard, auberge, audibert}@icp.inpg.fr

ABSTRACT

This paper presents an evaluation of the acted part of an audio-visual corpus of emotional speech. This corpus is intended to collect both spontaneous and acted emotions, and then the perceptive efficiency of stimuli to carry emotional expression has to be rated. The evaluation of acted speech is presented here, and will give us a scale to measure the spontaneous expressions.

INTRODUCTION

When one aims at describing and analysing emotional speech, he rapidly faces the problem of data: how to find corpora that share either a large panel of emotions, with the smallest linguistic variability, and a high recording quality? Most corpora of emotional speech are based on acted or elicited emotions (see e.g. Douglas-Cowie et al., 2003, for a review). Therefore, we have tried to build a corpus (for French) that contains both acted and spontaneous emotions, with very basic and repetitive linguistic variations, by using a protocol inspired from the wizard of Oz technique. The corpus and the methodology are described in Aubergé et al. (2003). The collected data is aimed at studying the expression of emotions in speech, in relation with face.

The study presented here consists both in evaluating the corpus and in perceptively testing some hypotheses held on the prosodic morphology that encode the expressions. It follows a general idea that we proposed for the evaluation of prosody (Rilliard & Aubergé, 2003): the evaluation of prosody must be « modularized » into the different functions encoded by prosody (the direct emotion encoding being one) and must be related to the cognitive representation of prosody. In a previous work on the acoustic analysis of the corpus, we proposed to extend a model of prosody based on the integration of superposed Gestalts contours and gradient tuning (Aubergé, 2002) on these contours of direct expressions of emotions.

A Corpus of Emotional Speech

One of the major singularities of these data is that emotions are spontaneously produced by the speakers in a first time, and then acted by the same speakers (who are professional actors). The induction of the emotional variations was not expected by the speakers, who were performing a Wizard of Oz task, held by a devoted man-machine scenario (cf. the "Sound Teacher" scenario in Aubergé et al. 2003). For this paper, we only work on the productions of one speaker.

One of the greatest advantages of such a method consists in the strict control made over the linguistic variations: the same sentences are repeatedly produced by the speakers with all emotional variations. This is very important in order to analyse the acoustic parameters linked to the expression of emotions: all the non-emotional variations are counterbalanced.

Some other constraints were imposed to our corpus:

- The collected emotions had to be spontaneously experienced and expressed before to be acted.
- Speech was recorded in a soundproof room in order to ensure a very high sound quality.

Both the speech signal and a facial video of the speaker were recorded synchronously during the experiment. Physiological data (skin conductance, skin temperature, hearth rate, respiration rate and EMG) were also recorded, in order to validate the speaker's actual physiological changes during the recording.

The sentences used during the experiment were based on:

- Five monosyllabic French words referring to different colours. They have been chosen in order to propose a set of vowels dispersed amongst the vocalic triangle: [u o a ɛ i], in the words "*rouge, jaune, sable, vert, brique*".
- A longer stimulus of three syllables was also recorded: "*page suivante*" (next page).

At the end of the "Sound Teacher" scenario, the speaker was interviewed in order to write down a list of all different emotions they had experienced. Then, he had to repeat the same utterances (the colour words and "*page suivante*"), plus ten sentences from three to seven syllables. Each stimulus is produced with the complete set of acted emotions based on: the "big six" emotions (i.e. happiness, sadness, fear, disgust, anger, surprise) plus the emotions he think he has experienced during the first phase (i.e. neutral, anxiety, deception, amusement, worried, resignation, satisfaction, expectancy).

Validation of the Corpus

In order to validate the emotional expression collected through such a paradigm, a perceptive validation has to be carried out. It has to first validate the acted emotions: the "big six", and the emotions reported by the listener himself. The results of this test give a first map of what listeners can efficiently perceive, and what kind of emotions cannot be differentiated. Then, the spontaneous data can be evaluated on a pre-tuned set of emotional category. This paper presents the results of the first step of the evaluation: the analysis of acted emotional expressions.

Subjects

26 subjects have participated in this experiment, including 4 males and 22 females, from 19 to 45-years old, aged of 25 in average.

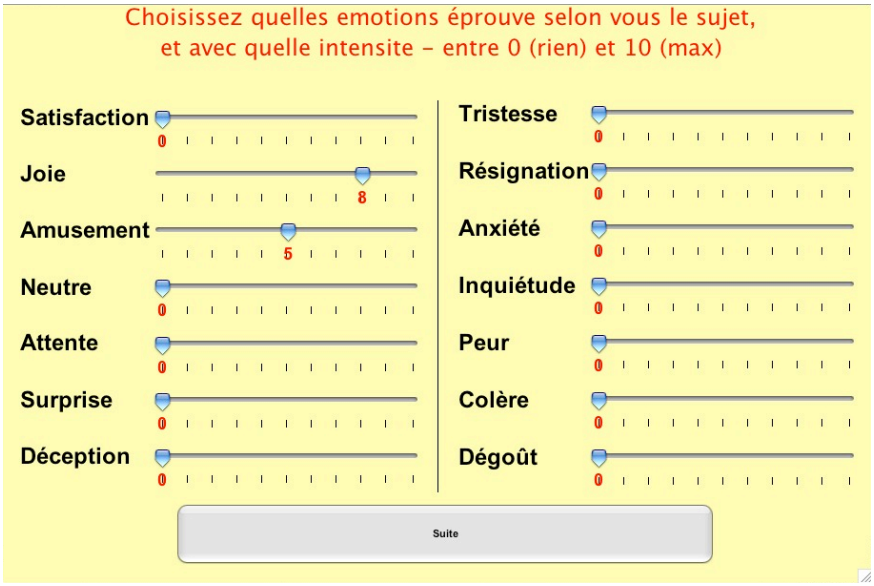

Figure 1: screenshot of the perception test answer page, showing the 14 emotional scales

Stimulus

The sentences proposed to listeners were extracted from the recording of one actor of the corpus described hereunder. There are two reasons for that: first the acoustic analyses made on the corpus (cf. Aubergé et al., 2004) are highly speaker-dependant, and the set of spontaneous emotions reported by this speaker is quite open.

Then, two experts listeners rated all his productions, in order to select only the best-acted performances, and to restrain the corpus for the listening test. In order to rate each stimulus, the judges listened to each stimulus in a random order, first in audio only, and then in the audio-video version, and gave each one a grade from 1 (very bad) to 4 (very good). Only the stimuli with a 3 or 4 grade were kept for the test. Then, a subset of these stimuli was extracted with the following criterions:

- Stimuli were selected in order to propose a systematic variation of their length. For each emotion, one stimulus is proposed with the following length: 1, 3, 5 and 7 syllables. This is made in order to test if the length influences the perception of emotional expressions.
- One stimulus was selected to represent all the emotional variation (the "*page suivante*" sentence), in order to test all emotional expression on exactly the same linguistic structure.
- The 14 acted emotional expressions were tested, either the "big six" ones, and the 8 reported at the end of the spontaneous phase, i.e.: amusement, anger, anxiety, deception, disgust, expectancy, fear, happiness, neutral, resignation, sadness, satisfaction, surprise, worried.

This gives 70 different stimuli, presented both through an audio and an audio-visual modality to listeners (resulting in 140 different stimuli).

Protocol

The perception test was carried out in a quiet room, using a computer to play the stimuli and to record the answers. Subjects listened to the stimuli via headphones, at a comfortable hearing level.

They heard in a first time the audio-only stimuli, mixed in a random order (controlled in order to avoid the successive presentation of the same sentence) different for each listener. Then, they perceived the audio-video stimuli, in a different random order.

They always heard the audio only stimuli and then the audio-video ones, because audio-video stimuli are used only as validation stimuli, to check if the audio expressions match the facial ones.

When the listener heard one stimulus, he had to rate the perceived intensity of the emotional expression for each of the fourteen labels proposed, on a scale from 0 (the emotion was not perceived) to 10 (the emotion is very intense). In order to give his answer, he had to use a set of 14 sliders corresponding to each emotion (cf. fig. 1). Stimuli can only be heard once, and listeners were told to give their answer as spontaneously as they could.

Results

The results of this perception test were compared amongst the 26 listeners, in order to ensure the coherence of their answers. The correlation between all their answers for each stimulus, for all pairs of listeners was calculated: all are significantly correlated with $p<.05$. Once the inter-listener coherence is checked, the overall dispersion matrices for the audio only and audiovisual condition were calculated (cf. fig 2 & 3).

General analysis

These two dispersion matrices reflect the first (and expected) result of this perception test: results in the audiovisual condition are always equal or better than those in the audio only condition; but the two conditions are quite coherent. A first analysis of the differences between the two conditions shows that:

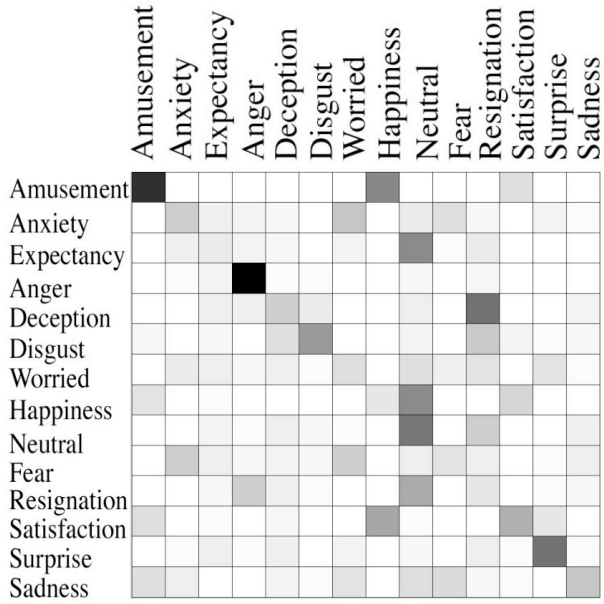

Figure 2: Dispersion matrix of the audio-only condition. The rows show input emotions, and the columns the mean answer of listeners. The intensity of the grayscale filling each square reflects the perceived intensity of each emotion.

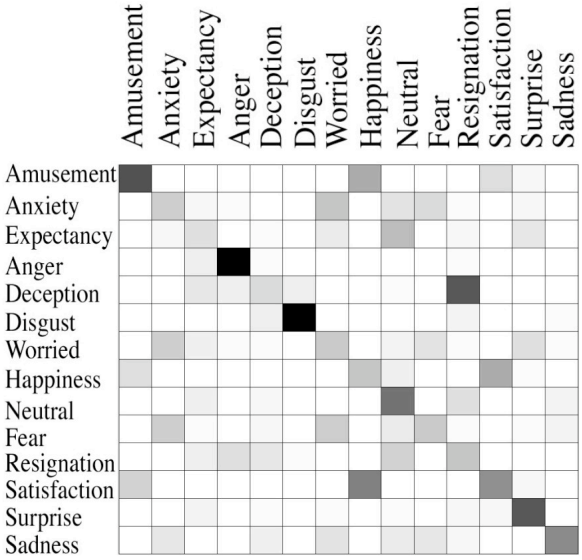

Figure 3: Dispersion matrix of the audiovisual condition. The rows show input emotions, and the columns the mean answer of listeners. The intensity of the grayscale filling each square reflects the perceived intensity of each emotion.

- Disgust seems difficult to recognize in the audio-only condition, whereas the audio-visual one is extremely efficient. These findings are completely similar to the conclusions of Scherer (2003) and Juslin & Laukka (2003) about disgust. However, we noticed an important difference between the accuracy of listeners to acoustically perceive disgust: about one half of them rated acoustic disgust as efficiently as audio-visual one, whereas the other half did not perceive acoustic disgust.
- Listeners did not use some categories: "*expectancy*" is recognized as "*neutral*" and "*deception*" is recognized as "*resignation*". For these two emotions, the face does not give more efficient information.
- Listeners in the audio-only condition mainly use the neutral category, when they don't "understand" the emotional expression. This happens for emotion with a low activation, such as "*expectancy*", "*happiness*" (the happiness played by this listener is a very low activated one), and "*resignation*".
- "A*nxiety*", "*worried*" and "*fear*" are mixed together, and listeners could hardly made differences, in both conditions. There are also some confusion between "*amusement*", "*happiness*" and "*satisfaction*", but not systematically: "amusement" is discriminated in the audiovisual condition, whereas "*satisfaction*" is mixed between "*happiness*" and "*satisfaction*". As it was already said, "*happiness*" is reported as "*neutral*" in the audio-only condition, but it is distributed between "*amusement*", "*happiness*" and "*satisfaction*" for the audiovisual one.
- The better recognized acoustic emotional expressions are "*amusement*" (even if it is mixed with happiness), "*anxiety*" (mixed with "worried" and "fear"), "*anger*", "*neutral*", "*satisfaction*" (mixed with happiness) and "*surprise*"

In order to more precisely analyse the results of this experiment, we will group the results of the different emotions that were not distinguished by listeners, in order to extract the cognitively pertinent classes of vocal expression of emotion. Then will be presented the results concerning the influence of the stimuli length on the emotional expressions.

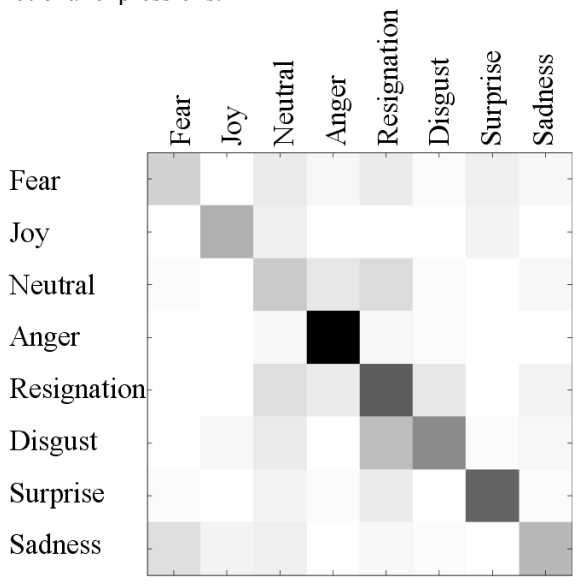

Figure 4: dispersion matrix for the 8 new categories obtained after grouping together the non-pertinent ones. The rows show input emotions, and the columns the mean answer of listeners. The intensity of the grayscale filling each square reflects the perceived intensity of each emotion.

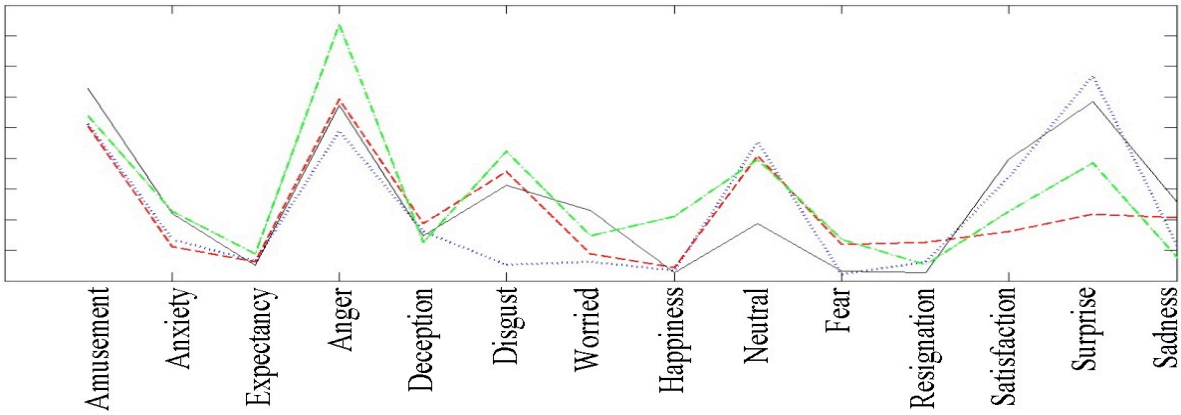

Figure 5: mean intensity given by listener to each emotional expression for each group of length. Each curve corresponds to a particular stimuli length: The doted one to the one-syllable long stimuli, the dashed one to the 3-syllable long ones, the plain line to the 5-syllable long ones, and the dashed-dotted line to the 7-syllable long ones.

On the cognitive pertinence of the different emotional expressions

In order to have a more precise view of the relevant emotional expressions produced by this speaker, we have grouped together 9 emotional labels into 3 new and more general labels, and exchanged the answers given in two categories ("Deception" and "Resignation"):

- "*Fear*", "*Anxiety*" and "*Worried*" are grouped together in a general "***Fear***" category
- "*Amusement*", "*Happiness*" and "*Satisfaction*" are regrouped inside the "***Joy***" category.
- "*Neutral*", "*Expectancy*" and "*Deception*" are grouped in a global "***Neutral***" category.

This results in a new dispersion matrix, with 8 emotional categories (cf. figure 4). The perceptive distinction for these categories is quite good. Thus, this set of label, and the grouping made to obtain them is very important in the perspective of evaluating the spontaneous data.

Influence of the stimulus' length on the perception of emotion

The last index that has to be analysed concerns the effect of the stimuli length on the perception of emotional expression. In order to obtain this information, we have grouped together the answers obtained for each stimuli of a given length. This results in 4 groups of length, for the 1, 3, 5 and 7-syllables stimuli. For each group of length, the average intensity given to each of the 14 emotional labels was calculated, in order to test if the listeners' answers differ from one length to another (cf. figure 5).

The correlation between the results of the four stimuli length was calculated, and all correlations are significant ($p<0.05$), indicating that the length of stimuli does not change the answer, for each emotional label.

Conclusion

These results are conceived as a first sorting of the collected data, as the expression of emotion raised a lot of very basic question, such as (1) the ability of human to act an emotion, or to perceive the difference between acted and spontaneous speech; (2) the cognitive pertinence of each emotions' label, in one language, several language, or even different culture; or (3) the relation between one emotion and its expression in speech (e.g. what is the intelligibility of acoustic contours for each emotional function).

This experiment deals with the second question, by pointing out labels' grouping, and by rating the relative efficiency of labels and acted productions. It could also bring some information to question 3, by comparing the acoustical analysis and the listeners' answers. Moreover, these first results underline that the length of stimuli does not change the ability of listeners to rate the emotional expressions.

References

Aubergé V. (2002). A Gestalt morphology of prosody directed by functions: the example of a step by step model developed at ICP, Proc of 1st Int Conf on Speech Prosody, Aix-en-Provence, 151-155

Auberge, Audibert, Rilliard (2003). Why and how to control the authentic emotional speech corpora? Eurospeech proceedings p 185-188.

Aubergé, V., Audibert, N. & Rilliard, A. (2004, to be published). Acoustic Morphology of Expressive Speech: What about Contours? Speech Prosody.

Douglas-Cowie, E., Campbell, N., Cowie, R. & Roach, P. (2003). Emotional speech: towards a new generation of databases. Speech Communication special issue on Speech and Emotion, .

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? Psychological Bulletin, 129(5), 770-814.

Rilliard A & Aubergé V (2003), Prosody evaluation as a diagnostic process: Subjective vs. objective measurement ». International Journal of Speech Technology, Kluwer Academic Publishers, p 409-418.

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

Acknowledgement

This work is part of the "Expressive Speech Project", held by the CREST/Japan Science and Technology and directed by Nick Campbell.