WHEN IS THE EMOTIONAL INFORMATION? A GATING EXPERIMENT FOR GRADIENT AND CONTOURS CUES

Nicolas Audibert[1], Véronique Aubergé[1,2], Albert Rilliard[3]

[1]Institut de la Communication Parlée - Gipsa Lab - CNRS UMR 5009
[2]UMAN Lab - Usages Marchés Attitudes Nanotech
[3]LIMSI CNRS, BP 133, 91403 Orsay Cedex, France

{Nicolas.Audibert, Veronique.Auberge}@icp.inpg.fr; rilliard@limsi.fr

ABSTRACT

The cognitive processing involved in the decoding of emotional expressions vs. attitudes in speech, as well as the modeling of emotional prosody as contours vs. gradual cues are debated questions. This work aims at measuring the anticipated perception of emotions on minimal linguistic units, to evaluate if the underlying processing is compatible with the hypothesis of gradient contours processing. Monosyllabic speech stimuli extracted from an expressive corpus and expressing anxiety, disappointment, disgust, disquiet, joy, resignation, sadness and satisfaction, were gradually presented in a gating experiment. Results show that identification along gates of most of expressions follow a linear pattern typical of a contour-like processing, while expressions of satisfaction present distinct gradient values that make possible an early identification of affective values.

Keywords: expressive speech; morphology; contours; gradient cues.

1. INTRODUCTION

Different kinds of affects are expressed in speech, related to voluntary communicative controls of the speaker (i.e. intentional values/attitudes and linguistic strategies/expressiveness) or to involuntary controls ("direct" expressions of emotions in voice). A discussion has been running for years, both in the fields of linguistics [8] and psychology [12], about the communication levels and/or the cognitive processing involved in the expression of emotions, moods, attitudes, mental states, feelings…

A central question arises when modeling the morphology of vocal expressions of affects, which is to understand how these different kinds of processing (summarized by the push vs. pull effect in Scherer's model [12]) can be implemented in the same acoustic material. Is the retrieval of emotional prosody vs. intentional affective prosody based on a separation of acoustic parameters or on morphological processing? It has been proposed by many authors, and developed by Bänziger and Scherer [6], that gradient processing would be the more relevant for emotional expressions, while contour processing would be reserved to linguistic prosody. We hypothesize [3] that for every kinds of affect, both processes are used together in a gradient contour processing, i.e. the affective information is carried by both the shape of the contour and global values that parameterize this contour. In this view, the relative weights of those processes can vary according to the type of affect expressed and to expressive strategies of the speaker.

If one considers that a gradient processing is involved in the decoding of emotional expressions, an interesting question would be to determine to what extend it makes possible an anticipated identification of emotion values, and the location of gradient cues. Though, to our knowledge, no studies investigating the perception of emotions in speech on such small units have been conducted, it can be expected that perceptually relevant variations appear on units smaller than the syllable. Indeed, Kohler [10] found in German different communicative values for the same stimuli, according to the late, medial or late position of the F0 peak. Consequently, a frame for studying the emotional values perceived on smaller units could be to divide vowels in three equal parts and compare the information carried by different parts.

The gating paradigm [9], classically used for testing prediction capabilities of judges on lexical access or phonemic identification tasks, consists in gradually presenting auditory stimuli. In such experiments each stimulus is cut at several fixed

points called gates, defined according to either absolute duration values or linguistic units such as the syllable. This paradigm has been applied to various purposes, including the analysis of the perception of French attitudes in speech [2], which showed an anticipated perception of attitudes from the 2nd syllable, using salient cues identified early in a Gestalt processing. In addition of that, current works carried out at the lab apply the gating paradigm to English and Japanese attitudes.

The aim of this paper is to evaluate whether salient features or gradient processing may also lead to an anticipated perception of emotional expressions or not, as well as the possible differences in anticipated perception for different emotional labels, by applying the gating paradigm to monosyllabic emotional expressions.

2. STIMULI GENERATION

27 speech stimuli extracted from a perceptively validated part [11] of the E-Wiz expressive corpus [1], on which no effect of the utterance was found, were selected as a basis for the generation of gated stimuli. The selected set was replayed by a male actor immediately after having been tricked in a Wizard of Oz experiment. Those stimuli express anxiety, disappointment, disgust, disquiet, joy, resignation, sadness and satisfaction on the French monosyllabic color names [ʒon], [ʁuʒ] and [vɛʁ], as well as neutral expressions on each of these words. The set of represented emotions was chosen as matching the one used in experiments of dimensional projection of prosodic contours [4, 5] in order to make possible a direct comparison of results. Partially unvoiced stimuli expressing the same emotions on the words [bʁik] and [sabl] would not have been suitable for such a gating experiment and were therefore discarded. Mean durations of selected stimuli range from 396 ms to 941 ms, with a mean value of 587 ms.

6 gates were defined relatively to hand-labeled phoneme boundaries with an increment of 1/3 phoneme, gate 1 being set at the 1st third of the vowel, gate 3 at the end of the vowel and gate 6 at the end of the final consonant.

Gated stimuli were generated using Praat [6], by extracting the part of signal ranging from the beginning of the stimulus to the gate value. White noise of variable duration was added at the end of the signal in order to normalize the total duration of all generated stimuli to 1250 ms.

3. PERCEPTIVE EVALUATION

The 162 generated stimuli were perceptively evaluated at the lab by 20 native French judges (7 male, 13 female, aged 30.6 in average), in a quiet environment with high quality headphones. Stimuli were presented sorted by ascending gates in a random order different for each judge within each gate length, the same stimulus being not presented twice consecutively. The test was automated using a graphical interface: judges had to select either an emotional label within the 8 proposed (anxiety, disappointment, disgust, disquiet, joy, resignation, sadness or satisfaction) or a "no emotion" label.

4. RESULTS

The first level of analysis consisted in extracting the confusion matrix for stimuli cut at gate 6 (i.e. full stimuli). Though differences in experimental design do not allow testing statistical significance, most of identification scores and confusions between labels appear as very close to those observed in [5]. However, since more stimuli were used for each emotion, and given that utterances of [sabl] had to be discarded, a few discrepancies could be observed. Expressions of sadness were notably less identified, and largely confused with joy. These confusions can be explained by a jitter, much more important on these expressions than on the expression of sadness on [sabl], and that might have been interpreted as laughter. Moreover, the expression of joy on [vɛʁ] was less identified than other expressions of joy. As a matter of fact, this expression has a flatter F0 contour than the ones of expressions of joy on [ʒon] and [ʁuʒ]. Results obtained from expressions of sadness and from the expression of joy on [vɛʁ] were therefore discarded from gates analysis. Though the confusions patterns do not differ to a large extend from those previously observed, the identification level on the expressions of disgust on [ʒon] and [ʁuʒ] is lower. A possible explanation for that is that, despite their similar morphologies, several judges reported the identification of disgust as more difficult on the vowels [o] and [u] than on [ɛ] and [a].

Mutual confusions over chance level remain the same as in [5]. The same clustering was thus applied to the analysis of answers: anxiety and disquiet were grouped together, as well as resignation, disappointment and sadness, joy was

clustered with satisfaction, while disgust and neutral remained separate categories. However, if prosodic morphologies are quite the same within each label, excepting discarded expressions of sadness and joy, they differ between emotional labels of a same cluster. Data were consequently analyzed separately for each label, which was not the case in previous studies.

Table 1 presents the confusion matrix a gate 6, in which answers are analyzed by cluster. All correct identification scores except those of sadness and resignation are significantly over chance level (paired t-tests on recoded data, $p<0.01$). Chance level, indicated on the first line, depends on the number of labels in a cluster and is therefore not the same on every column.

Table 1: Confusion matrix at gate 6. Rates of correct identification appear in bold characters. Chance level for each cluster is indicated on the first line.

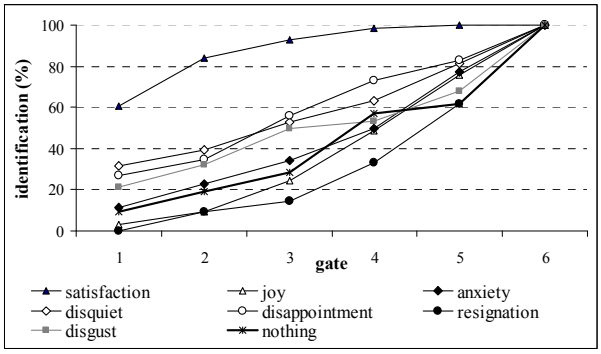
	joy satisf	disapp-res-sad	anx disq.	disgust	nothing
chance	*22.2%*	*33.3%*	*22.2%*	*11.1%*	*11.1%*
joy	**60%**	10%	18.4%	0%	11.7%
satisf.	**98.4%**	1.7%	0%	0%	0%
disapp.	1.7%	**88.3%**	0%	8.3%	1.7%
resign.	8.3%	**35%**	28.3%	1.7%	26.7%
sadness	58.3%	**15%**	23.3%	0%	3.3%
anxiety	3.3%	11.7%	**73.3%**	1.7%	10%
disquiet	15%	11.7%	**65%**	0%	8.3%
disgust	15%	33.3%	0%	**51.7%**	0%
nothing	1.7%	50%	8.3%	0%	**40%**

The stable identification point was defined for each judge*stimulus pair as the index of the gate from which one of the emotional labels of the correct cluster is chosen, without a change of answer in any of the following gates. As all judges did not manage to correctly identify stimuli at the last gate, a stable identification point could not be calculated for every judge*stimulus pair. Such records were thus discarded from analysis.

Pearson's chi-square tests were performed to compare distributions of stable identification points as a function of emotional label and utterance, showing a significant effect of emotional label ($p<0.01$), but no effect of the utterance. Indeed, the only noticeable differences found in identification rates as a function of the utterance appear for expressions of disgust and for the expression of joy on [vɛʁ].

Figure 1 presents the evolution of correct identification derived from stable gates values, for each emotional label. In order to make possible a comparison of patterns between labels, all curves were normalized to a final identification score of 100% at gate 6. While patterns for all other expressions are quite linear with a correct identification rate at gate 1 below chance level, showing a progressive identification with no salient cues, the expressions of satisfaction show a different pattern, with an correct identification rate of 60.7% at gate 1. All other tested expressions are identified below chance level at gate 1. Disquiet and disgust are identified over chance level from gate 2, anxiety and disappointment from gate 3, and joy and nothing from gate 4, while resignation only reaches chance level at gate 6.

Figure 1: correct identification as a function of the gate for each emotion, normalized to 100% at gate 6.



5. DISCUSSION

The affective information on expressions of joy and satisfaction was found to be mainly carried by F0 contours [4], while it was mainly carried by voice quality and duration in the case of negative emotional expressions. F0 contours of these expressions can therefore be considered as a reliable cue to the affective information presented at different gates. As the expression of joy on [vɛʁ] was excluded, contours of expressions of joy and satisfaction on [ʒon] and [ʁuʒ] were compared.

F0 values were converted to semitones, where the reference value (0 semitones) is set to the mean F0 value of the speaker in the whole corpus (96.8 Hz). Interestingly, expressions of joy and satisfaction share similar F0 contour shape, but have very different anchor values (on [ʁuʒ]: mean = 3.8 semitones for joy, 7.8 for satisfaction; range = 6.1 semitones for joy, 14.9 for satisfaction; on [ʒon]: mean = 3.9 semitones for joy, 7.0 for satisfaction; range = 8 semitones for joy, 16.5 for satisfaction). Those 2 realizations can therefore be considered, in a parallel with allophones, as "alloemotems". The

comparative analysis of F0 contours reveals that these gradient values are already known at 1st gate, making possible an early identification of satisfaction. On the other hand, as anchor values of joy are not different enough from those of other tested expressions, these values do not make possible an identification of the emotional value before the contour shape is known.

Bänziger and Scherer [6] claim that the F0 mean level and range vary strongly with the activation of emotional expressions and can account for most of the perceptively measured variations, while contour shape carry much fewer information. Since activation, though not perceptually rated, is clearly higher on tested expressions of satisfaction vs. joy, our results on these expressions support this hypothesis. However the claim that contour shapes play a minor role in the decoding of emotional values should be revisited, as most of tested expressions appear to be identified using contour shapes characteristics.

6. CONCLUSION

Since the study presented in this paper was conducted on a very restrained set of emotional expressions, results need to be replicated before being generalized. However it brings evidence for a gradient contour processing of emotional expressions, which we consider as extending the hypothesis of gradient only processing proposed by many authors. Although gradient processing appears as more able to yield earlier identification than contour shape when gradient values are specific enough to discriminate from other expressions, a contour-like processing appears as predominant for most of tested expressions.

As the stimuli used in this study were produced by an actor replaying involuntary emotions, we cannot ensure that these expressions are similar to genuine involuntary expressions, and thus cannot interpret our results in terms of social vs. involuntary affects. However studies under progress on spontaneous data and attitudes bring indices in favor of a generalized gradient contour processing.

Perspectives for future work are multiple. In order to evaluate if the perception of affective information depends more on the repartition of information (i.e. contour shapes) or on the overall quantity of information (for instance if the length of a syllable can be sufficient for the perception of affects), a gating experiment on longer sentences has to be conducted. Moreover gating experiments carried out on French attitudes are being extended to gates smaller than the syllable to ensure that differences in the processing of emotions vs. attitudes can be generalized to a broader set of attitudinal expressions.

On the other hand, as it has been shown that no prosodic dimension carries the whole affective information alone [4, 5], another gating experiment will be conducted on dimensions separated by resynthesis, to evaluate how gradience and contours are distributed. on prosodic dimensions.

7. ACKNOWLEDGMENTS

The work presented in this paper was supported by the "Expressive communication" project of the Pegasus PPF (universities of Grenoble, France), and by the R&D division of France Telecom.

8. REFERENCES

[1] Aubergé, V., Audibert, N., Rilliard, A. 2004. E-Wiz: A Trapper Protocol for Hunting the Expressive Speech Corpora in Lab. *Proc.4th LREC*, Lisbon, 179-182.

[2] Aubergé, V., Grépillat, T., Rilliard, A. 1997. Can we perceive attitudes before the end of sentences? The gating paradigm for prosodic contours, *Proc. Eurospeech*, Rhodes, 871-874.

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

[4] Audibert, N., Vincent, D., Aubergé, V., Rosec, O. 2006. Expressive Speech Synthesis: Evaluation of a Voice Quality Centered Coder on the Different Acoustic Dimensions. *Proc. Speech Prosody*.

[5] Audibert, N., Aubergé, V., Rilliard, A. 2005. The prosodic dimensions of emotion in speech: the relative weights of parameters. *Proc. Interspeech*, Lisbon, 525-528.

[6] Bänziger, T., Scherer, K.R., 2005. The role of intonation in emotional expressions. *Speech Communication,* 46, 252-267.

[7] Boersma, P., Weenink, D. 1992-2007. Praat: doing phonetics by computer. http://www.fon.hum.uva.nl/praat

[8] Fonagy, I., 1986. Les langages de l'émotion. *Quaterni di semantica*, 7(2), M. Alinei (ed.), Bologna, 305-318.

[9] Grosjean, F., 1980. Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*. (28), 267-283.

[10] Kohler, K.J., 2005. Timing and Communicative Functions of Pitch Contours. *Phonetica,* 62, 88-105.

[11] Rilliard, A., Aubergé, V. and Audibert, N., 2004. Evaluating an Authentic Audio-Visual Expressive Speech Corpus. *4th LREC*, Lisbon, Portugal, 175-178.

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