

# The role of intonation and voice quality in the affective speech perception

*Ioulia Grichkovtsova<sup>1</sup>, Anne Lacheret<sup>2</sup>, Michel Morel<sup>1</sup>*

<sup>1</sup>CRISCO, Université de Caen, France

<sup>2</sup>MoDyCO, Université Paris X, Nanterre, France

`ioulia.grichkovtsova@unicaen.fr`

## Abstract

The perception value of intonation and voice quality is investigated for six affective states: anger, sadness, happiness, obviousness, doubt and irony. The main research question is whether the role of intonation and voice quality is equally important in the perception of studied affective states or whether one of them may be privileged. Six affective states were tested on utterances with natural lexical meaning. The transplantation paradigm was used in designing audio stimuli. Perception results show that each studied affective state has its own usage of prosody and voice quality. Some differences were found in the identification of emotions and attitudes. New questions risen from the present study and further directions of work are presented.

**Index Terms:** affective prosody, voice quality, perception, speech synthesis

## 1. Introduction

Affective speech research [1, 2] suggests that both intonation and voice quality participate in the vocal encoding of affective states. Voice quality is attested as an important acoustic correlate of affective expressions by many researchers [2, 3, 4, 5]. It was shown [5, 6] that voice quality alone may evoke affective associations. Nevertheless, these associations do not exist on one-to-one basis, as specific voice qualities may be associated with a group of related, or even unrelated affective states. The search for the associations between specific intonation contours and affective states was not very successful [7], but it does not mean that intonation is not involved in the realization of vocal affective states. Some authors suggest [8] that intonation contours may primarily serve linguistic functions, such as realizations of different sentence types, while pitch range is one of the primary parameters covering phonetic variability and serving paralinguistic functions.

Recent studies [9, 10] examined how voice quality and intonation contours combine in the signaling of affective states. Their results suggest that stimuli involving voice quality variation are much more consistently associated with affect, than stimuli based on  $F_0$  variation. The methodological approach in the two mentioned studies was to use utterances carrying no lexical meaning for the listeners. The question is raised if this approach of working with non-sense utterances could bias the identification of affective meaning from stimuli with  $F_0$  variation. Characteristic features of intonation contour, namely pitch accents and prosodic phrase boundaries, are linked to chunks, meaningful components of the utterance. As non-sense texts are formed out of syllables, listeners may be unable to break the utterance into chunks, and hence to fully interpret meaning of the prosodic features, used in the utterance. The usage of natural utterances as audio stimuli may help to test our hypothesis.

This study is aimed to investigate the perception value of intonation and voice quality for six affective states (anger, sadness, happiness, obviousness, doubt and irony) based on the corpus of utterances with natural lexical meaning. Two main research questions are addressed: 1) whether the role of intonation and voice quality is equally important in the perception of studied affective states or whether one of them may be privileged; 2) whether emotions may be differentiated from attitudes based on the perceptive value of intonation and voice quality.

## 2. Method

### 2.1. Prosody transplantation method

A prosody transplantation method has been chosen for the design of audio stimuli. It involves extraction and exchange of prosody between two utterances with the same segmental content. The prosodic transplantation involves the extraction of the intonation contour, rhythm and temporal parameters, but it does not modify the voice quality characteristics of the utterance. This method has already been used in the research on speech perception [11].

### 2.2. Stimuli description

Six utterances were pronounced by an actor and an actress for each studied affective state (anger, sadness, happiness, obviousness, irony and doubt), thus the total of 72 utterances was used in the experiment. The lexical meaning of the utterances was not neutral, they were designed to carry natural lexical meaning, appropriate to particular affective states. See examples in Appendix. The same utterances were synthesized with Kali, a French-speaking text-to-speech diphone synthesis system [12], these synthesized utterances did not have any vocal affective meaning.

In the process of prosody transplantation, the prosody of Kali was mapped on the utterances encoded by actors. As a result, the new version of the utterance adapted the prosody of Kali, but it preserved the original voice quality used by the actor. In the same way, the prosody of the actor was mapped on the utterance synthesized by Kali. Thus, four versions of each utterance were developed: version 1 - "natural" (natural prosody and voice quality of the encoded affective state by the actor), version 2 - "voice quality" (natural voice quality and prosody from Kali), version 3 - "prosody" (natural prosody and voice quality of Kali), version 4 - "lexical" (voice quality and prosody of Kali, the only indication on the affective state of the utterance may come from the lexical meaning). Thus, 288 stimuli were prepared for the perception experiment. Graphical examples of the stimuli, used for the transplantation method are given in Figures 1 and 2.

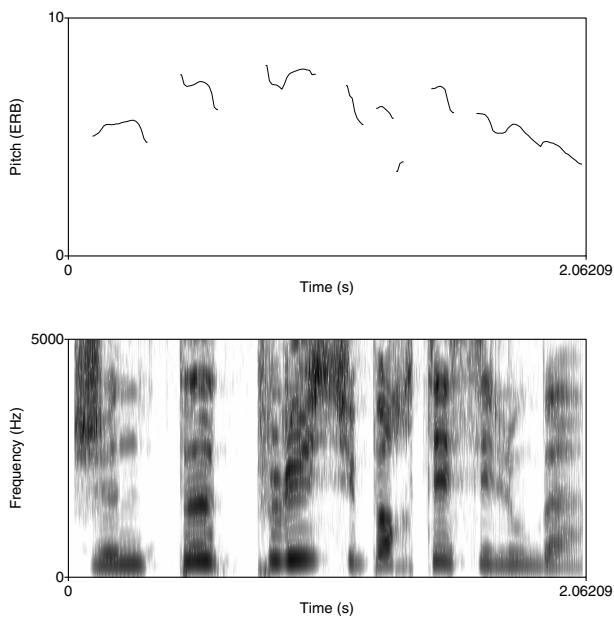


Figure 1: The realization of the utterance with anger by the actress. “Je ne peux plus supporter ces loubards./I can’t stand these rogues any more.”

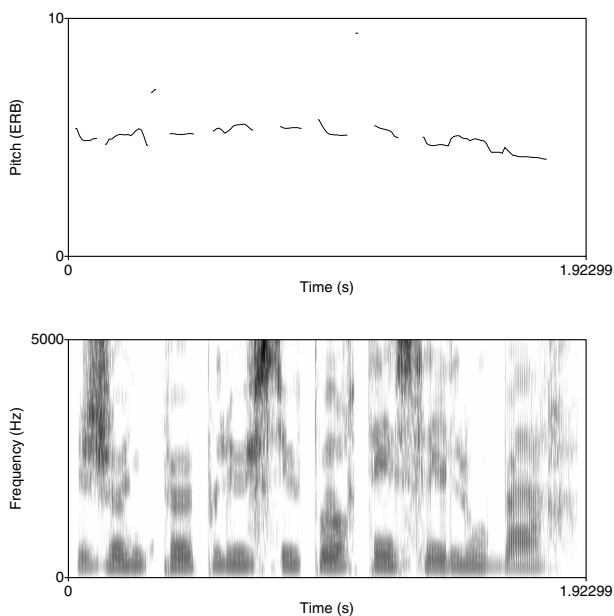


Figure 2: The realization of the utterance by Kali. “Je ne peux plus supporter ces loubards./I can’t stand these rogues any more.”

### 2.3. Participants

12 subjects were recruited in University of Caen (5 females and 7 males). They were students and professionals working in the university. Average subject age was 31 years old, with standard deviation of 13.2. All subjects were native speakers of French and none reported having any hearing difficulty.

### 2.4. Procedure

Special software for psycholinguistic perception tests *Perceval* was used for the experiment design. *Perceval* was developed in the Laboratoire Parole et Langage, Aix-en-Provence, France, and is free for academic purposes [13]. Stimuli were presented in a randomized order. The experiment was run on a computer in a quiet laboratory room. The whole experiment was run by the *Perceval* software, responses and response times were automatically recorded in the data file. In average, the experiment took about 40 minutes.

### 2.5. Results

The results of the perception test, displayed in Table 1, show the level of identification for the four versions of audio stimuli in percentage. Results for “natural” stimuli show the best identification of the encoded affective state, as both natural voice quality and prosody of the actor are present. Results for “lexical” stimuli have the lowest identification levels, as the lexical meaning of these utterances is not supported by voice quality and prosody specific to the studied affective state. At the same time, it is possible to observe variability in the level of identification of affective states for stimuli with only lexical meaning. See Figure 3 for the values with confidence intervals. The identification of “voice quality only” and “prosody only” stimuli vary according to the analyzed affective states.

Table 1: Identification of affective speech stimuli. The correct responses are expressed in percentage.

	Natural speech	Voice quality	Prosody	Lexical meaning
Anger	95%	59 %	84%	44%
Doubt	85%	35%	78%	15%
Obviousness	78%	66%	66%	66%
Happiness	87%	65%	63%	34%
Irony	74%	49%	63%	47%
Sadness	86%	86%	68%	66%

#### 2.5.1. Neutralization of the lexical meaning

The effect of the lexical meaning is present not only in the “lexical” version of the stimuli, but in all the used stimuli. In order to abstract from the role of the lexical meaning in the identification of affective states, a neutralization procedure was applied. The value of the identification for “lexical” stimuli was considered as the role of lexical meaning in the identification of the affective states. This value was extracted from the identification values received for “natural”, “voice quality” and “prosody” stimuli of the corresponding affective state. For example, “natural” - “lexical” (for anger: 95% -44%). Thus, the received value was judged as the contribution of voice quality and/or prosody in the affective speech perception. These results are displayed in Figure 4. The confidence intervals of the values are also given, they allow to see if the differences between groups of stimuli are statistically significant.

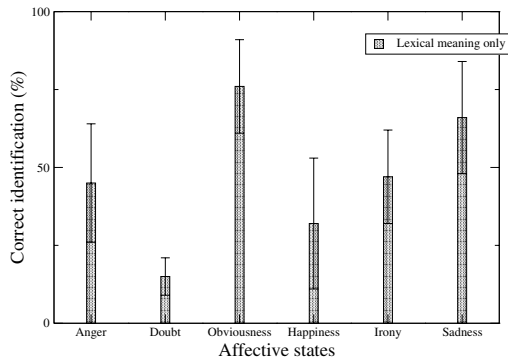


Figure 3: Identification of stimuli with lexical meaning only. Values are presented in percentage.

Results for anger show that both prosody and voice quality are used in the process of identification. Nevertheless, “prosody” stimuli were identified much more successfully than “voice quality” stimuli, and this difference is statistically significant. Doubt results show that prosody is the main acoustic correlate in the identification of this affective state. No significant results are observed for the identification of obviousness, apparently this affective state was identified exclusively on the basis of lexical meaning. Voice quality and prosody play an equally important role in the identification of happiness. Lexical meaning is apparently very important in the perception of irony, but prosody contributes to its better identification. The importance of lexical meaning for irony may be explained by the choice of utterances. Ironic utterances contained words of opposing literal meaning. The results for sadness show that voice quality plays the privileged role in the identification. Nevertheless, it is important to note that its identification in “lexical” stimuli is very high. It may be explained not by the importance of the lexical meaning for sadness identification, but by a particular prosody and voice quality of the synthesized voice Kali. The proper voice of Kali sounds relatively sad, so it could potentially facilitate the identification of sadness in the “lexical” stimuli and bias the results on the role of prosody and voice quality. These results are summarized in Table 2.

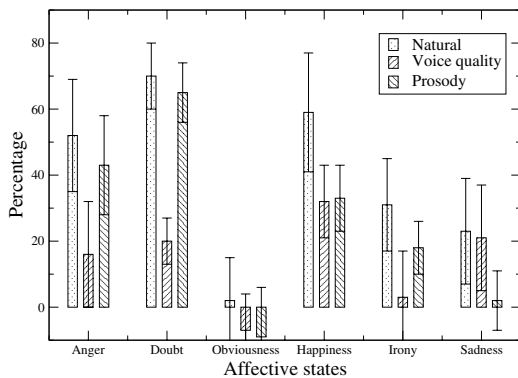


Figure 4: Neutralized values of successful identification.

Table 2: Role of prosody and voice quality in affective speech.

Affective state	Roles
Anger	prosody > voice quality
Doubt	prosody
Obviousness	lexical meaning
Happiness	prosody and voice quality
Irony	lexical meaning > prosody
Sadness	voice quality (and prosody?)

### 3. Discussion

This study investigated the perception value of voice quality and prosody for six affective states: anger, doubt, obviousness, happiness, irony and sadness. Previous studies suggested that voice quality has stronger and more consistent associations with affective meaning than prosody. Please, note that these conclusions were driven from the perception studies with non-sense utterances.

Our choice to use natural utterances was motivated by the hypothesis that listeners have difficulty to interpret affective meaning of prosodic features without access to chunks, meaningful components of the utterance. Our results showed the importance of prosody in the identification of affective states, even if its level of importance is variable. The presence of natural prosody may be even crucial for the successful identification of some affective states, like doubt.

Moreover, our approach to use natural utterances allowed to shed light on the role of lexical meaning in the identification of different affective states based on stimuli with the lexical meaning only, without the corresponding support of voice quality and prosody. The identification of evidence was based almost entirely on the lexical meaning. Listeners could not categorize this affective state on the basis of prosody and/or voice quality information. Results for irony and anger show that lexical meaning may be important. Happiness may be identified to a lesser extent from the lexical meaning. Finally it is almost impossible to identify doubt from the lexical meaning only. The presence of natural lexical meaning does not influence the identification of the affective states in the same way.

It is not possible to separate completely emotions from attitudes based on the perceptive value of prosody and voice quality. Nevertheless, some interesting observations were done: lexical meaning played an important role for two out of three studied attitudes (evidence and irony), prosody also was involved in the identification of doubt and irony. Voice quality was not identified as important in the three studied attitudes. The analysis of the results for emotions showed that both prosody and voice quality carry an important perceptive value. Lexical meaning is generally less important for emotions than for attitudes. While for happiness, both prosody and voice quality are equally important, anger privileges prosody.

### 4. Conclusions

The usage of the transplantation paradigm allowed to investigate the role of prosody and voice quality in the identification of emotions and attitudes. The hypothesis that voice quality is more important in the identification of emotions, and that prosody is privileged in the identification of attitudes was not supported. Each affective state showed its particular usage of prosody and voice quality. This study is based on the corpus of utterances encoded by only two French speakers. Recent

research draws the attention to the high variability existing in affective speech. The question stands if other speakers may use other strategies in the encoding of affective states. If yes, different distribution of voice quality and prosody importance for the same affective state may be possible. Another question is if emotions may be realized with more different strategies than attitudes. These questions will be addressed in our ongoing study, which is based on the work with a multi-speaker corpus of affective speech.

At the present, the new corpus is being developed for the future perception tests with the transplantation paradigm. A large number of speakers will be recorded encoding several affective states on the same neutral utterance. It is another way to understand the roles of prosody and voice quality in the identification of affective states without the influence of the lexical meaning. This method does not interfere in our approach to work on the perception of prosody in meaningful utterances.

## 5. Appendix

**Sadness:** J'ai finalement compris que je ne la reverrais plus. (I finally understood that I would never see her again.)

**Happiness:** Il a appelé sa mère pour lui annoncer la bonne nouvelle. (He called his mother to tell her the good news.)

**Doubt:** J'ai roulé sans phares en pleine nuit? (I was driving with lights off at night?)

**Anger:** J'ai encore retrouvé ma voiture neuve toute rayée, c'est inadmissible! (I have found my new car scratched, it is unacceptable!)

**Obviousness:** Il a mis un manteau pour sortir! (He put on his coat to go out!)

**Irony:** J'ai réussi ma chute en pleine rue brillamment. (I managed to fall in the middle of the street very nicely.)

## 6. Acknowledgements

This work was supported by a grant of the Basse-Normandie Regional Council. The authors would like to thank all the participants who kindly agreed to do the perception tests.

## 7. References

- [1] Johnstone, T. and Scherer, K. R. (2000) "Vocal communication of emotion." In Lewis, M. and Haviland, J. M. (editors) "Handbook of emotions." Guilford, New York, 220-235.
- [2] Murray, I. R. and Arnott, J. L. (1993) "Towards the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion." *Journal of Acoustic Society of America*, 93 (2), 1097-1108.
- [3] Laver, J. (1994) "Principles of phonetics." Cambridge University Press.
- [4] Gendrot, C. (2004). "Influence de la qualité de la voix sur la perception de quatre émotions simulées: une étude perceptive et physiologique." *Parole*, 13 (1), 1-18.
- [5] Ladd, D.R., Silverman, K., Tolkmitt, F., Bergmann, G. and Scherer, K.R. (1985). "Evidence for the independent function of intonation contour type, voice quality, and f0 range in signalling speaker affect." *Journal of the Acoustical Society of America*, 78 (2), 435-444.
- [6] Gobl, C. and Ni Chasaide, A. (2003) "The role of voice quality in communicating emotion, mood and attitude." *Speech Communication*, 40 (1-2), 189-212.
- [7] Stibbard, R. M., "Vocal Expression of emotions in Non-Laboratory Speech: An Investigation of the Reading/Leeds Emotion in Speech Project Annotation Data." (2001) PhD thesis. University of Reading.
- [8] Mozziconacci, S. (2002) "Prosody and Emotions." Proceedings of the International Conference Speech Prosody, Aix-en-Provence, France.
- [9] Yanushevskaya, I., Gobl, C. and Ni Chasaide, A. (2006) "Mapping voice to affect: Japanese listeners." Proceedings of Speech Prosody 2006, Dresden, Germany.
- [10] Morel, M. and Banziger, T. (2004) "Le rôle de l'intonation dans la communication vocale des émotions : test par la synthèse." *Cahiers de l'Institut de Linguistique de Louvain (CILL)*, 30 (1-3), 207-232.
- [11] Garcia, M.-N., d'Alessandro, C., Bailly, G., Boula de Mareuil, P. and Morel, M. (2006) "A joint prosody evaluation of French text-to-speech systems", Fifth International Conference on Language Resources and Evaluation (LREC), Gènes, 307-310.
- [12] Morel, M. and Lacheret-Dujour, A. (2001) Kali, synthèse vocale à partir du texte : de la conception à la mise en oeuvre." *Traitement Automatique des Langues*, 42, 1-29.
- [13] Ghio, A., André, C., Teston, B. and Cavé C. (2003) "PERCEVAL: une station automatisée de tests de PERCEPTION et d'EVALUATION auditive et visuelle", TIPA, vol. 22, Aix-en-Provence, France, 115-133.