IrcamCorpusExpressivity: Nonverbal Words and Restructurings

Grégory Beller, Christophe Veaux and Xavier Rodet

IRCAM
1. place Igor Stravinsky
75004 Paris, France
{beller, veaux, rodet}@ircam.fr

Résumé

In this paper, we present the various constituents of a spoken message which allow the observation of expressivity in speech. These constituents are joined into the perspective of the double coding of the speech, which distinguishes the linguistic channel of the paralinguistic channel in a spoken message. Among this last channel, several phenomena seem to participate in the demonstration of the expressivity: The prosody, naturally, but also the nonverbal sounds, as well as of possible restructurings. In a second part, we introduce the expressive French multi-speaker corpus: IrcamCorpusExpressivity. Several steps of labeling and analysis allow the examination of this corpus under the various angles corresponding to the constituents of the spoken message. These results can be used to improve the tasks of recognition, transformation and synthesis of the expressivity in the speech, and so contribute to the anthropomorphisation of the Human-machine interfaces.

1. Introduction

Human-machine interfaces based on voice processing allow to obtain good rates in neutral speech recognition and to supply an understandable quality with synthesis. The introduction of the treatment of the *expressivity* in these tasks bring the researchers today to analyze the *expressive* speech (Bulut et al., 2007) (Yamagishi et al., 2005). The term *expressivity* is, here, defines as a level of information in the communication (Beller, 2008b). This level groups together the external demonstrations, simulated or not, which are attributable to internal states. Among these internal states are included the emotions, the attitudes, the feelings, the humors as well as the other styles which compose the range of actor's performance style. Our goals is to transform the expressivity of a neutral given utterance (recorded or synthesized) (Beller and Rodet, 2007) and are intended for artistic purposes (musical composition, contemporary theater, dubbing of cinema, animation, avatars and robots).

To do it, studies on the prosodic variations that can be attributed to the changes of expressivity were led. In particular, the introduction of new paradigms of analysis allowed to estimate the influence of the expressivity on the speech rate (Beller et al., 2006) and on the degree of articulation (Beller et al., 2008a). These studies notably showed the importance of the breaths which are part of nonverbal sounds. So other phenomena as purely prosodic participate in the communication of the expressive information. Some of these phenomena are emphasized here, notably thanks to the observation of an expressive French multi-speaker corpus : IrcamCorpusExpressivity.

After a theoretical introduction of the various constituents of the speech, we describe in a exhaustive way the corpus IrcamCorpusExpressivity realized within the VIVOS[1] project. The various levels of manual labeling are detailed to supply dictionaries created in this occasion and with the aim of showing certain tendencies according to the expressivity. The study of the continuous parameters of the prosody, which is one of our major subjects usually, is voluntarily put aside in this paper, so as to leave more place with the presentation of the corpus, as well as on the examination of the various levels of labeling relative to the other constituents of the speech.

[1]VIVOS :`http://www.vivos.fr`

2. Constituents of the speech

To observe the influence of the expressivity on the speech, we describe in this part various phenomena which the verbal communication implies and called constituents of the speech. They are summed up in the figure 1 and connected together with the perspective of the double coding of the speech, proposed by Fónagy (Fónagy, 1983). This perspective differentiates the linguistic channel of the paralinguistic channel. The linguistic channel is carrier of the semantic information and can be transcoded, without loss of information, in a text. The paralinguistic channel vehicles other levels of information that those carried by the linguistic channel, as the speaker identity, the speaking style, the modality, the prominence and, indeed on, the expressivity (Beller, 2008b).

2.1. Linguistic channel : verbal words and syntax

The linguistic channel brings the verbal words and their syntactical relationships. From an acoustic point of view, it is supported by sequences of segments, called *phones*. These phones are realizations of phonemes, which constitute the symbolic closed dictionary of differentiable sounds of a language. A verbal word possesses a meaning and a linguistic transcription. It can be written by use of a term stemming from the dictionary of common and proper nouns of a language. It thus depends on the sociocultural standards, quite as the syntax which depends on the grammar and which is also a part of the linguistic channel.

2.2. Paralinguistic channel

Among the paralinguistic channel, we discern the prosody, the nonverbal words and certain restructurings. All these elements are carriers of information others than linguistic.

2.2.1. Nonverbal words

The nonverbal words are sounds deprived of linguistic functions. By opposition to the verbal word, a nonverbal word does not possess usual transcription. However, it is not rare to find phonetic-spelling transcriptions of these sounds as ”ah ah ah” or ”laughter” to describe the presence of a laughter in a text (from the comic-strip to the novel, by way of the script of a play). It is because of this semantic dimension relative to the expressivity that we speak here of nonverbal words and not nonverbal sounds.

As the nonverbal words do not possess standardized transcription, they are with difficulty describable otherwise than by reproduction. In spite of a big variety, we distinguish among the nonverbal words, ”fillers” (laughter, scream, tear...), the breaths (inspirations, stops, expirations...) (Beller et al., 2006) and the other noises (guttural, nasal, of mouth...). It seems that these nonverbal words are of rich meaningful for the expressivity (Schroeder et al., 2006). The sadness can be only perceived by a tear and the fear, only by a scream, without the support of any verbal word. More finely, an informal perceptive experiment shows that the simple local addition of a breath in the middle of a neutral sentence, can change the perceived expressivity of the whole utterance[2]. The expressive power of the nonverbal words is such, that speech synthesizers begin to generate them (Beller, 2008a), so as to increase the naturalness and the expressivity of the synthesis. It requires, among others, the definition of standard for their transcriptions. The recent attempts base for the greater part on extensions of the SSML[3] language (Eide et al., 2004) (Blankinship and Beckwith, 2001).

2.2.2. Restructurings

The way are temporarily ordered the verbal and nonverbal words is informative. This is well known in the case of the verbal words temporal organization of which is defined by syntactical constraints. In the case of a spontaneous communication, these words can however not respect any more the order governed by the grammatical rules while preserving them syntactical functions. Indeed, the contiguity between nonverbal and verbal sounds force these last ones to possible temporal reorganizations called *restructurings*. So, although the syntax adjacent to the linguistic message organizes a priori the words and thus the sequences of phones, numerous not grammatical restructurings come into play, as the repetition of phones, syllables, whole word either even whole propositions (resetting). In spontaneous speech, the repetition which is frequent does not affect necessarily the understanding of the words and their syntactical relations. On the other hand, it can be a demonstrator of the hesitance or the confusion which are categories of expressivity. Other restructurings are carriers of sense for expressivity, while they are generally considered as *disfluencies* for the neutral speech (Piu and Bove, 2007) and concern the pronunciation : The *coarticulation*, the caesura, the connection and the elision are examples.

[2]Examples listenable to at : `http://www.ircam.fr/anasyn/beller`

[3]SSML : Speech Synthesis Markup Language : `http://www.w3.org/TR/`

2.2.3. Prosody

The stream of speech is thus a sequence of verbal and nonverbal words all organized by the conjugate action of the syntactical rules and the possible restructurings. At the same time, the acoustic realization of all these sound segments is ”modulated” by the prosody. If this is well known as regards the verbal words, it remains true for the nonverbal words as the laughter, for example (Beller, 2008a). The prosody includes suprasegmental phonological features the temporal span of which exceeds the boundaries of the phone (the syllable, the accentual group, the word, the clitic, the breath group, the prosodic group, the sentence...) and which do not annul the comprehensibility (that is that they do not deprive a phone of its membership in a phonetic category). Five characteristic features are generally quoted in the literature as the five dimensions of the prosody (Pfitzinger, 2006) :

– intonation : fundamental frequency, pitch
– intensity : energy, volume
– speech rate : flow, rhythm, speed of delivery
– degree of articulation : pronunciation, configurations of the vocal tract, dynamics of formants
– phonation : glottis signal, voice quality (pressed, normal, breathy), vibratory mode (fry, normal, falsetto), voicing frequency...

For a half a century of study of the neutral speech, the prosody was often reduced to the intonation. The intonation so benefited from a lot of attention and modelling, because, easy to observe, it allowed it only, bringing to the foreground functions of the prosody (modality, emphasis). The case of the expressive speech seems to require more strongly the observation of the other dimensions (Campbell and P.Mokhtari, 2003). Finally, of part its continuous character in the time, the prosody accompanies the production of verbal and nonverbal sounds and also interacts with the syntax and the restructurings.

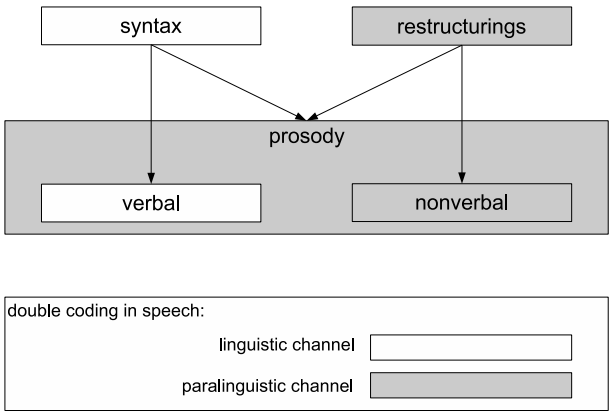

FIG. 1: Representations of the constituents of the speech. The verbal words and the syntax establish the linguistic channel. The nonverbal words, the prosody and the restructurings are the vectors of the paralinguistic channel.

A carrier vocal message of several levels of information is established by a sequence of verbal sounds and nonverbal sounds, all modulated by the prosody and organized by the conjugate action of the syntax and the restructurings (see fi-

gure 1). These paralinguistic phenomena are especially observable in the spontaneous speech, even more in the spontaneous dialogue and of advantage still in the case of the expressive speech as shows it the examination of the expressive corpus IrcamCorpusExpressivity.

3. IrcamCorpusExpressivity

The corpus IrcamCorpusExpressivity consists of recordings of four actors : Jacques, male, storyteller/comedian ($\sim$40 years), Philippe, male, comedian dubber ($\sim$40 years), Olivia, female, comedian dubber ($\sim$25 years) and Danielle, female, comedian dubber ($\sim$50 years). Every recording was guided thanks to a computing interface allowing a simplification of the recording process. This interface possesses a screen presenting the sentence, the expressivity and the intensity to be realized. The comedian starts and ends the recording thanks to a pedal. This interface also facilitates the post-production because it allows the synchronization, the labeling and the segmentation of the corpus as this one is recorded. So the actor can make a mistake or begin again without that it entails gaps. The comedians were recorded in the same conditions and in the environment which they know because it is their workroom. The studio of dubbing presents the advantage of an appropriate acoustics, being enough reverberating. So the actors feel less vocal fatigue than in an anechoic chamber, which possesses an unusual and particularly dry acoustics. A static microphone wearing an anti-pop filter allowed the acquisition of the data in ADAT[4] quality. Data stemming from an Electro-Glotto-Graph (EGG) are also available on certain parts of the corpus. In the end, more than 500 utterances were taken in by actor, forming a corpus of total duration about 12 hours of expressive speech.

3.1. Recited Text

The recited text were extracted from a French corpus of twenty sets of ten sentences. Every set is phonetically balanced (Combescure, 1981) :

1. C'est un soldat à cheveux gris.
2. Alfred pris la tête de l'expédition.
3. Il ne pourra pas me voir si j'éteins la lampe.
4. Il entre avec sa chandelle, dans la vielle chambre.
5. Le nez du personnage s'abaisse, au-dessus de sa moustache.
6. Vous êtes vraiment obéissant !
7. En attendant, c'est moi qui vais ouvrir.
8. Je ne pourrai jamais, me plier à son autorité.
9. Tout le monde sait que tu es la meilleure.
10. Je me demande, où se trouve cet endroit ?

This set was especially chosen because it contains neutral sentences with regard to the expressivity. That is that these sentences make sense, with every the expressivity with which they are pronounced. The prominence of some syllables was indicated to the actors by the punctuation and by the usage of uppercase characters. It allowed to vary the places of prominence and thus the resultant prosody, and to "congeal" the accentuation of the sentence to fix the semantic contents from a repetition to the other one.

3.2. Expressivity

The range of the wanted expressive categories was defined at the starting point of the VIVOS project, taking account the needs of a dubbing studio, of a commercial TTS synthesizer company and of an embedded video games company :

– Neutral
– *introvert anger* : contained or cold anger
– *extrovert anger* : explosive or warm anger
– *introvert happiness* : sweet or maternal happiness
– *extrovert happiness* : explosive or enthusiastic happiness
– *introvert fear* : contained or tetanic fear
– *extrovert fear* : explosive or alarming fear
– *introvert sadness* : contained sadness
– *extrovert sadness* : explosive or tearful sadness
– discretion
– disgust
– confusion
– positive surprise : the speaker is pleasantly surprised
– negative surprise : the speaker is unpleasantly surprised
– excitement

So as to be able to represent the recorded expressivities in a dimensional space axes of which are the valence (positive vs negative), the intensity (degree of intensity of the expressivity) and the activation (introversion vs extraversion) (Schroeder, 2003), we asked the actors to express the primary emotions (Ekman, 1999) with several degrees of intensity and according to two versions relative to the introversion and to the extraversion. For the last expressivities (in normal character), the comedians directly said all the text with the level of intensity the strongest possible. For the expressivities in italics, the degree of intensity was varied according to five levels. The progress of the recording is described by the following procedure. For a given expressivity, the speaker utters the first sentence in a neutral way. Then she/he repeats five times this sentence, with the wished expressivity, by increasing her/his degree of intensity. Then she/he moves to the following sentence and begins again this progress. Finally, she/he repeats this plan with the other expressivities. This procedure notably allows to obtain an intensification of the expressivity without that the speaker is to read again the text every time. From an intensity to the other one, neither the sentence, nor its accentuation changes, letting seem only the variations attributable in the intensity of the expressivity. The actors had for explicit order not to vary the pronunciation of their realizations corresponding to a sentence. This so as to minimize the variations due to the phenomena of restructuring which complicate the comparison of an expressive utterance with its neutral version and the building of conversion prosodic model (Tao et al., 2006) (Hsia et al., 2007). Interestingly and as it will be shown further, some restructurings appear despite this order. Nonverbal sounds were recorded then separately at the end. The following fillers was collected : "ah", "oh", "laughter", "tear", "fear", "panic", "enjoyment", "euh", "interrogation", "argh", "effort", "running", "hhh", "fff", with several realizations according to the ex-

[4] ADAT : Alesis Digital Audio Tape : 16bit, 48KHz

pressivity, for some of them.

3.3. Collected data

The data collected during the recording consist of audio files for every sentence and corresponding XML[5] files, containing the labeling of the expressivity (category and intensity), of the recited text, and of the information relative to the identity of the speaker (age, sex, name). These starting data have been manually labeled : phonetic segmentation, paralinguistic labeling and prominence labeling. Then symbolic analyses derived from these labels and, finally, acoustic analyses of the prosody have been processed. All the data which we are afterward going to describe, are stored and made accessible by IrcamCorpusTools, a database management system involving a powerful language of request (Beller et al., 2008b).

3.4. Phonetic segmentation

The phonetic segmentation of this corpus is, actually, a semi-phonetic segmentation (thus more precise). Indeed, a phone consists of two semiphones whose borders also allow to establish diphones (for the analysis and the synthesis). This segmentation was initialized by an automatic method (Lanchantin et al., 2008). This one leans on of multiple phonetizations of the text (Bechet, 2001) and implies a neutral French multi-speaker corpus (Lamel et al.,). This tool allows a fast and automatic segmentation which is not regrettably sufficient in the case of expressive speech. This *bootstrap* segmentation was thus manually checked and corrected by a phonetician. Then this correction was verified by another one. Not only the borders were moved but labels were also changed when it turned out necessary. The used code is the XSampa, which is an ASCII version of the IPA[6] chart.

Correcting the *bootstrap* segmentation, the phoneticians noticed numerous differences with the predictions of the machine (trained on neutral speech). For certain expressivities, expected phones was so differently realized that they were relabelized by other phonemes (opened /E/ moved to closed /œ/, for instance). Furthermore, some disappeared whereas other, unexpected, appeared. That is why, although all the expressivities were supported by the same text, we expected disparities in the posterior distributions of labeled phones, possibly attributable to the expressivity. However, all actors included, the average proportions of appearance of phones grouped together into phonological classes (see figure 2), do not show significant differences, function of the expressivity. Only confusion shows significant differences, cause by the numerous repetitions, as shown by the analysis of paralinguistic labels (see next section). On the other hand, direct local comparisons (and not statistical, as shown here) of the phonetic labels of the ideal phonemic sequence deduced from the text, and of the labeling of the realized phonetic sequence can allow to deepen this study.

3.5. Paralinguistic segmentation

Simultaneously in the operation of manual correction of the phonetic segmentation, a layer of supplementary labels was

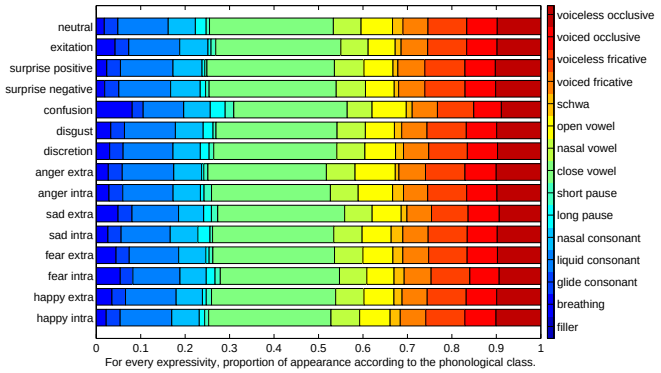

FIG. 2: All actors included. Average proportions of appearance of phonological classes by expressivity.

produced so as to supply paralinguistic information. This information notably describes the used nonverbal sounds, the possible restructurings and the diverse particular phonatory or prosodic phenomena. This stage of labeling required, once the stage of phonetic segmentation ended, a second pass to homogenize the labels. Indeed, because no dictionary of paralinguistic labels for this type of phenomena was defined a priori, the vocabulary employed by the annotators evolved according to the task and thus, from a corpus to the other one. The dictionary which we subject here, was thus the object of several inter-annotators discussions (and intra) and seems to gather the most important labels :

- Nonverbal sounds, breaths and voicing of the phonation :
  - [˚] : inspiration
  - [˚˚] : expiration
  - [nz] : nasal breath
  - [bx] : non vocal noise annoying signal analysis
  - [bb] : mouth noises
  - [ch] : whisper, instability of the voicing during the phonation, partial devoicing
  - [nv] : not voiced : total absence of vocal folds vibration
  - [ph] : transition : label indicating a nonverbal zone in continuity with a verbal zone (often short, but crucial for expressivity). In most of the cases, we meet this phenomenon either just before the first semiphone of a breath group, or just after the last semiphone of a breath group
- Pitch and guttural effects :
  - [fp] : pitch effects : sudden pitch variation often upward, concerning mostly only one semiphone, sudden change of vibratory mode ”normal” ↔ ”falsetto”
  - [fg] : guttural effect : audible glottal stops or starts, cutting of glottis excitement, sudden change of vibratory mode ”normal” ↔ ”fry”
  - [fi] : other effects than guttural or of pitch
  - [nt] : not transcribable : label put compared to the phonetic segmentation allowing to mean the doubt as for the attribution of the phone in a phonemic category
- Restructurings :
  - [lg] : length : phone abnormally long
  - [cu] : caesura : label applied to a silent phone by place (jerky phonation)

[5] XML : eXtensible Markup Language

[6] IPA : International Phonetic Alphabet

– [rp] : repetition : indexation of repeated phone or group of phones (up to 9 repetitions have been observed in the corpus : rp1, rp2, ..., rp9)

Finally, these various labels are composables thanks to the usage of the symbol [/] who allows to elaborate complex pattern from the given basic labels. For example, a voiced inspiration in falsetto mode starting a vowel is annotated by [˚/fp/ch] (seen in the extrovert fear case, for example), whereas a nasal expiration will be represented by [˚˚/nz]. Finally the interaction with the layer containing the phonetic segmentation is strong, because the same label put compared to a silence or to a phone will not mean the same thing.

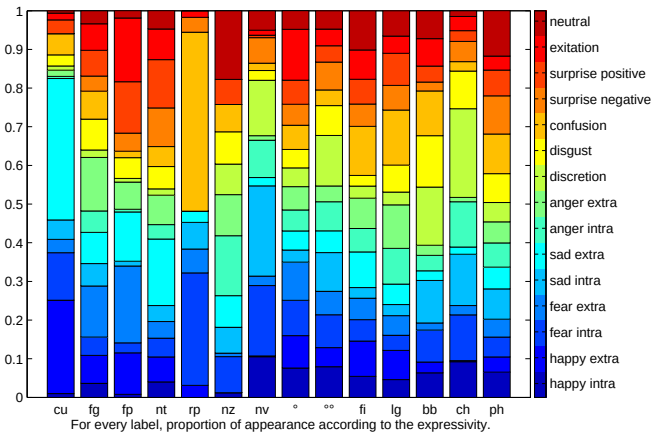

FIG. 3: All actors included. Average proportions of used paralinguistic labels according to the expressivity.

The figure 3 presents for each of the paralinguistic labels, the proportion of every expressivity. The more an expressivity contains a label in a recurring way (with regard to the others), the more the height of its associated rectangle is big. So, we observe that the caesura [cu] was strongly employed on the labeling of utterances expressed with extrovert sadness (jerky phonation). This expressivity contains so a lot of pitch effects [fp] and of no transcribable phones [nt]. In fact, several sentences are almost unintelligible because of a too weak degree of articulation (Beller et al., 2008a). The extrovert anger is marked by the presence of guttural effects [fg] as well as of nasal expirations [nz] (like the introvert anger and the disgust). Repetitions [rp] appear mainly for the introvert fear, and the confusion. This last expressivity also distances itself by numerous phonemes abnormally long [lg], like the ”angers” and the ”happinesses” (Beller et al., 2006). The discretion and the introvert sadness contain numerous markers of weak voicing ([nv], [ch] and [ph]). Furthermore, the inspirations [˚] are less labeled (perceived) compared to expirations [˚˚] in the case of the discretion. Finally the negative surprise seems less voiced that the positive surprise and presents less inspirations than expirations with regard to that last expressivity. Other numerous interpretations are possible and can, there also to be supported by more detailed local examinations.

3.6. Prominence labeling

From the phonetic segmentation, a rule based syllabifier (Veaux et al., 2008) produces a segmentation in syllables. The syllable plays a particular role in the prosody, notably because it is the smallest pronounceable prosodic group. Of a perceptive point of view, syllables distance themselves according to their levels of prominence. The prominence reflects an acoustic contrast (culminance, distinction, demarcation), performing several functions. First of all, it shows the accentuation of certain syllables which can be defined linguistically (Lacheret-Dujour and Beaugendre, 1999). The prominence plays sometimes also a role in the disambiguation of the sense (pragmatic accent). Finally, the prominence serves for emphasizing certain elements of the utterance (accent of focus, of emphasis, of insistence). A single annotator labeled the degree of prominence of the syllables of the whole corpus. The used scale consists of four levels :

– [UN] : indefinite or silence/pause (considered here as a syllable)
– [NA] : not prominent
– [AS] : secondary prominent
– [AI] : prominence with emphasis
– [AF] : final prominence (regular in French)

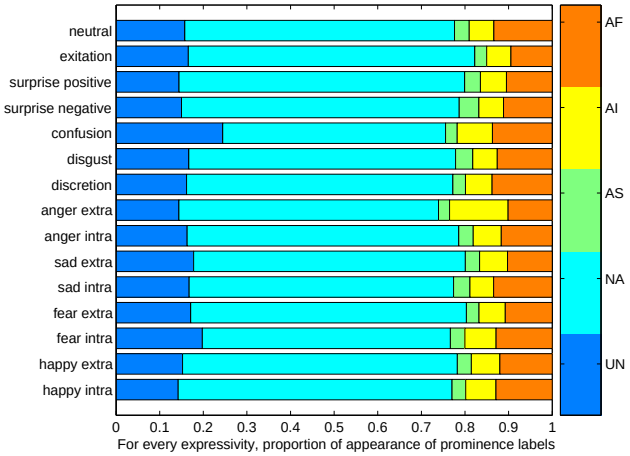

FIG. 4: All actors included. Average proportions of used prominence labels according to the expressivity.

A similar study to the previously presented ones concerning the distribution of these labels according to the expressivity does not show tremendous significant variations (see figure 4. Only the extrovert anger seems to distinguish itself from the other expressivities by a bigger proportion of [AI] labels. Indeed, syllables expressed with extrovert anger seems perceived more often as prominent. It can be explain, partially, by a hyperarticulation (Beller et al., 2008a) which provokes a detachment of consecutive syllables that become all prominent since they are all demarcated.

4. conclusion

In this paper, we presented, at first, a theoretical point of view allowing the observation of the expressivity in the speech. This point of view was included in the perspective of the double coding of the speech, which distinguishes the linguistic channel of the paralinguistic channel in a spoken message. Among this last channel, several phenomena seem to participate in the communication of the expressivity : the prosody, naturally, but also the nonverbal sounds,

as well as of possible restructurings. In a second part, we introduced the expressive French multi-speaker corpus : IrcamCorpusExpressivity. Several labelings and analyses allowed the examination of this corpus under the angle of the various phenomena belonging to the paralinguistic channel. Few differences attributable to the expressivity are visible in the phonological distributions, as well as in the distributions of the levels of prominence. However, these results are to be minimized because the actors had exactly for explicit order, not to make vary these constituents, but only the prosody. On the other hand, if they also had for order to avoid the usage of nonverbal sounds and restructurings, nevertheless this constituents appears frequently in the corpus. Their examination allowed to bring to light that certain expressivities distinguish themselves by strong apparences of some of these constituents. As if some of them required the use of nonverbal sounds and restructurings besides the prosodic variations to be expressed. To validate these findings, a similar study on corpus not basing on such orders is to be made. Nevertheless, these various results can be already used to improve the tasks of recognition, transformation and synthesis of the expressivity in the speech, and so, contribute to the anthropomorphisation of the Human-machine interfaces.

5. Acknowledgments

This work was partially funded by the French RIAM network project VIVOS. The authors wish to thank the actors involved in this study for their performances, as well as the various annotators having participated in the labeling of the data.

6. References

Frederic Bechet. 2001. Liaphon : un système complet de phonétisation de textes. In *Traitement Automatique des Langues - TAL*, number 1, pages 47–67.

Grégory Beller and Xavier Rodet. 2007. Content-based transformation of the expressivity in speech. In *ICPhS*, Saarbr ucken, August.

Grégory Beller, Diemo Schwarz, Thomas Hueber, and Xavier Rodet. 2006. Speech rates in french expressive speech. In *Speech Prosody*, Dresden, may. SproSig, ISCA.

Grégory Beller, Nicolas Obin, and Xavier Rodet. 2008a. Articulation degree as a prosodic dimension of expressive speech. In *Speech Prosody 2008*, Campinas, May.

Grégory Beller, Christophe Veaux, Gilles Degottex, Nicolas Obin, Pierre Lanchantin, and Xavier Rodet. 2008b. Ircam corpus tools : Système de gestion de corpus de parole. *TAL, to appear*.

Grégory Beller. 2008a. Semi-parametric synthesis of speaker-like laughter. *to appear*.

Grégory Beller. 2008b. Transformation of expressivity in speech. In Peter Lang, editor, *The Role of Prosody in the Expression of Emotions in English and in French*. Peter Lang.

Erik Blankinship and Richard Beckwith, 2001. *UIST '01 : Proceedings of the 14th annual ACM symposium on User interface software and technology*, chapter Tools for expressive text-to-speech markup, pages 159–160. ACM, New York, NY, USA.

Murtaza Bulut, Sungbok Lee, and Shrikanth Narayanan. 2007. a statistical approach for modeling prosody features using pos tags for emotional speech synthesis. In *ICASSP*.

N. Campbell and P.Mokhtari. 2003. Voice quality : the 4th prosodic dimension. In *XVth Int. Congress of Phonetic Sciences*, volume 3, pages 2417–2420, Barcelona.

Pierre Combescure. 1981. 20 listes de dix phrases phonétiquement équilibrées. *Revue d'Acoustique*, 56 :34–38.

E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli. 2004. A corpus-based approach to <ahem/> expressive speech synthesis. In *5th ISCA Speech Synthesis Workshop*.

P Ekman, 1999. *The Handbook of Cognition and Emotion*, chapter Basic Emotions. John Wiley & Sons, Ltd.

I. Fónagy. 1983. *La vive voix : essais de psychophonétique*.

Chi-Chun Hsia, Chung-Hsien Wu, , and Jian-Qi Wu. 2007. conversion function clustering and selection for expressive voice conversion. In *ICASSP*.

A. Lacheret-Dujour and F. Beaugendre. 1999. *La prosodie du Francais*. CNRS langage.

L. Lamel, J. Gauvain, and M. Eskenazi. Bref, a large vocabulary spoken corpus for french.

Pierre Lanchantin, Andrew C. Morris, Xavier Rodet, and Christophe Veaux. 2008. Automatic phoneme segmentation with relaxed textual constraints. In *Language Resources and Evaluation Conference (LREC2008)*, volume ND, Marrakech, Maroc, Mai.

H.R. Pfitzinger. 2006. Five dimensions of prosody : Intensity, intonation, timing, voice quality, and degree of reduction. In H Hoffmann, R. ; Mixdorff, editor, *Speech Prosody*, number 40 in Abstract Book, pages 6–9, Dresden.

Marie Piu and Rémi Bove. 2007. Annotation des disfluences dans les corpus oraux. In *RECITAL*.

M. Schroeder, D. Heylen, and I. Poggi. 2006. Perception of non-verbal emotional listener feedback. In *Speech Prosody 2006*, Dresden, Germany.

Marc Schroeder. 2003. *Speech and Emotion Research : An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis*. Ph.D. thesis, University of Saarland.

Jianhua Tao, Yongguo Kang, and Aijun Li. 2006. Prosody conversion from neutral speech to emotional speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4) :1145 – 1154, July.

Christophe Veaux, Grégory Beller, and Xavier Rodet. 2008. Ircamcorpustools : an extensible platform for speech corpora exploitation. In *LREC*, Marrakech, Maroc, Mai.

J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi. 2005. Acoustic modeling of speaking styles and emotional expressions in hmm-based speech synthesis. In *IEICE Trans. on Inf. & Syst.*, volume E88-D, pages 503–509, March.