Prosodic Parameters and Prosodic Structures of French Emotional Data

Katarina Bartkova [1], *Denis Jouvet* [2] *and Elisabeth Delais-Roussarie* [3]

[1] Université de Lorraine & CNRS, ATILF, UMR 7118
[2] Speech Group, LORIA, UMR 7503, INRIA, CNRS & Université de Lorraine
[3] UMR 7110-LLF (Laboratoire de Linguistique Formelle), Université Paris-Diderot
katarina.bartkova@atilf.fr, denis.jouvet@inria.fr, elisabeth.roussarie@wanadoo.fr

Abstract

The detection and modelling of emotions in speech remains a challenging issue in speech processing. The aim of the study presented here is to analyze and compare the use of several prosodic parameters in emotional speech in French. The data set used for the study contains utterances recorded in six emotional types: anger, fear, sadness, disgust, surprise and joy. The sentences of the emotional data are also recorded by the same speaker in a neutral reading style allowing a comparison between emotional and neutral speech. The prosodic analysis focuses to the main prosodic parameters such as vowel duration, energy and F0 level, and pause occurrences. The values of prosodic parameters are compared among the various emotional styles, as well as between emotional style and neutral style utterances. Moreover, the structuration of the sentences, in the various emotional styles, is particularly studied here through a detailed analysis of pause occurrences and their length, and of the length of prosodic groups.

Index Terms: emotional speech, emotions, neutral style, prosodic parameters

1. Introduction

Over the last 30 years, many works showed that prosody conveys information on the linguistic content of the message, but also on speaker's attitudes [1] and emotional states (see, among others, [2], [3] and [4]), leading to a distinction between linguistic prosody and emotional prosody. Note however that apprehending and distinguishing both prosodic domains is not an easy task, emotional and linguistic prosody being expressed by the same phonetic features (loudness, pitch, and speech tempo) and sometime interacting in a complex way [5]. As for perception, the robustness of the various prosodic parameters in the interpretation of speaker attitudes and emotions is still at issues in a given language as well as cross-linguistically (see, among others, [6]). Knowing how the different prosodic parameters are used to convey a particular emotion and how they interact with linguistic functions is crucial to provide a model for emotional prosody that can be used in speech technology. It requests however the development of methodological approaches.

Speech data bases that exhibit emotion styles are collected in various ways including in natural spontaneous speech, elicited emotional speech and simulated or acted speech [7], [8], [9]. Besides the traditional classification of emotions (anger, fear, disgust, etc.) multi-dimensional representations are also used, as for example arousal (relaxed vs. aroused), valence (pleasant vs unpleasant), etc. (see [10]). In [11] findings of previous research works relating prosodic and spectral features to the emotions are summarized.

Various sets of prosodic and spectral features as well as various classifiers have been used for emotion recognition [12], [13]. Most of the approaches rely on large set of features that are represented by their statistics (min, max, mean, standard deviation, skewness, etc.) [14]; and best performance is achieved using statistics computed on the entire utterance segment rather than on word or syllable segments [15].

In [16] and [17] various approaches that have been studied for expressive speech synthesis are described. Both prosodic and spectral aspects play a significant role for a correct perception of the emotions [18] and [19], although the most critical one depends on the type of emotion. Concatenative speech synthesis has been used for emotional speech synthesis thanks to the recording of an emotion speech corpus for each of the desired emotions [20], [21]; prosodic-phonology approaches have also been investigated for predicting fundamental frequency and duration through statistical models associated to linguistic units of prosody [21]. Statistical speech synthesis (HMM-based) is also used for expressive speech synthesis. Performance (perception of emotion by listeners) is similar whether separate HMM-based models are developed for each emotion or whether the emotion information is included in the detailed context label (along with usual phonetic and linguistic information) [22]. When developing the speech synthesis models, using the emotion information corresponding to the intended emotion acted by the speaker or the emotion type as perceived by listeners leads to similar performance [23]. Also, differences between neutral speech and emotion speech has been studied in order to produce emotion speech through "difference prosody models" [24], or through conversion procedures to transform neutral speech into a target emotion speech [25], [26].

This study analyses the behavior of prosodic parameters on French speech data in six different emotion styles: anger, fear, sadness, disgust, surprise and joy. After a presentation of the speech data and associated features in Section 2, Section 3 focuses on the analysis of the main prosodic parameters (vowel duration, vowel energy and fundamental frequency), with comparisons between each emotion style and the neutral style. Section 4 focuses on the analysis of the pauses and of the prosodic structures in the various emotion styles, a topic which is not investigated in the literature. A short discussion in Section 5 summarizes the results.

2. Corpus and features

The speech corpus used contains French sentences in various emotion styles, as well as some neutral style speech material.

All the data are recorded from the same male speaker. About 50 utterances are recorded for each of the 6 emotion styles: anger, fear, sadness, disgust, surprise and joy. The same sentences are also recorded in a neutral reading style. Hence comparison between each emotion style and the neutral style can be performed on the same set of sentences.

For each sentence, the speech material is aligned with its corresponding text using an automatic speech-text forced alignment procedure (available at `http://astali.loria.fr`). The approach relies on a statistical modeling of the speech sounds (Markov models). It provides the segmentation of each utterance into phone and silence (pause) units.

Prosodic features are then computed for each utterance. Segment duration is derived from the automatic phone level segmentation; whereas energy and fundamental frequency are extracted with the ETSI/AURORA front end [27].

3. Prosodic characterization

This section analyses for each vocalic segment duration, energy, and fundamental frequency in the six emotion styles. The behavior of these parameters is compared between emotion styles and also with respect to the neutral style.

3.1. Vowel duration

Vowel durations are measured in word non-final syllables as well as in word final syllables. Syllables in word non-final positions are in a non-stressed position, whereas those in word final position can be stressed and therefore lengthened. For every emotion style as well as for the neutral style, there is a significant difference for the mean duration of the vowels between word non-final and word final positions (Fig. 1). With respect to word non-final positions, the lowest vowel mean duration is observed in the fear emotion, and the longest one is found in the joy emotion. For word final positions, the longest vowel mean duration is observed in the anger emotion.

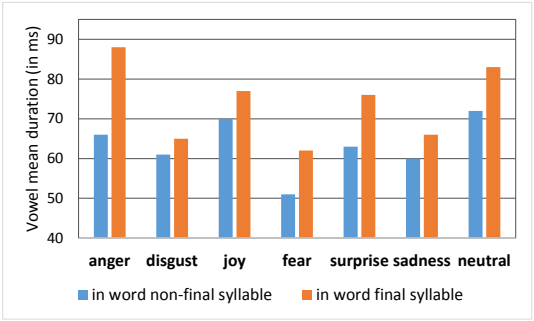

Figure 1: *Vowel mean durations in word non-final and word final syllables.*

For each emotion style, and each word position (i.e., non-final vs. final syllables), the duration of the vowels are compared to a reference vowel duration which corresponds to the vowel mean duration calculated on neutral style in same positions (i.e., respectively non-final vs. final syllables). Figure 2 reports the percentage of vowel occurrences that have longer duration than the corresponding reference vowel duration. Anger is the emotion which has the higher percentage of lengthened vowel durations (in both word non-final and final positions). The smallest percentage of lengthened vowels is observed in the fear and disgust emotion styles.

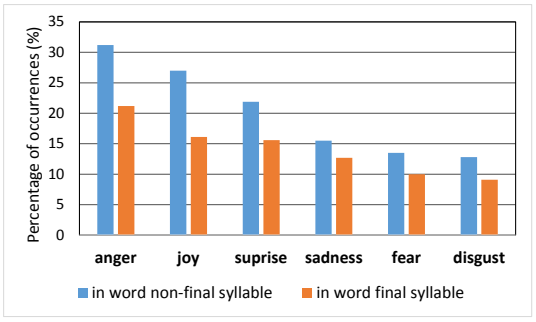

Figure 2: *Percentage of vowel occurrences, for the six emotion styles, that are longer than the corresponding mean duration (estimated on the neutral style).*

The lengthening of the vowels in word non-final syllables has been quantified on a four degree scale, as defined in the PROSOTRAN prosodic annotator [28]. The first degree (1) corresponds to vowels that are longer than the mean duration plus one time the standard deviation. The second degree (2) corresponds to vowels that are longer than the mean duration plus two times the standard deviation. Degree 3 is defined in a similar way; and degree 4 corresponds to vowels that are longer than the mean duration plus four times the standard deviation. Figure 3 illustrates, for each emotion style, the amount of vowels in word non-final position that are lengthened (according to the above scale). It thus displays only the percentages of vowels that are longer than the mean duration plus at least one time the standard deviation. The sadness emotion style contains mostly moderate vowel lengthening while the anger emotion style contains the higher percentage of heavily lengthened vowels (degree 4).

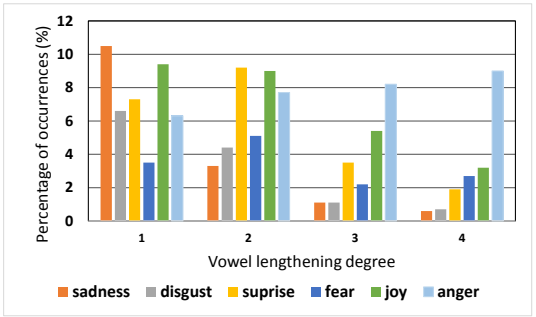

Figure 3: *Percentage of vowel occurrences with respect to the four degree lengthening scale (from 1 – moderate lengthening to 4 – heavy lengthening).*

3.2. Vowel energy

An analysis similar to the one done for vowel lengthening is carried out for vowel energy, considering vowels in word non-final position. A four degree scale is also defined for quantifying the level of increase of vowel energy. The first degree (1) corresponds to vowels whose energy is higher than the mean vowel energy plus one time the standard deviation and similarly for degrees 2, 3, and 4. Figure 4 illustrates, for each emotion style, the percentage of vowel occurrences in word non-final position that have an increase in energy (according to the defined scale). It thus considers only the vowels for which the energy is higher than the vowel mean value plus N times the standard deviation. For sadness, all vowels, with higher energy than the mean reference value, are only slightly higher than the

reference value. Also, but to a lesser extent, a similar behavior is observed for the disgust emotion. On the other hand anger has the highest percentage of vowels with a large increase in energy (degree 4: energy strongly higher than the reference value).

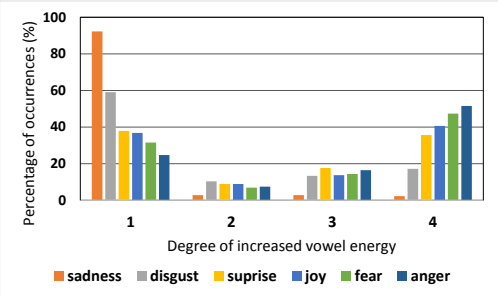

Figure 4: *Percentage of vowel with respect to the four degree scale of energy increase.*

3.3. F0 values

F0 values vary according to the general emotional state of the speakers. Values are higher when the speaker is angry and lower when the speaker is sad. Figure 5 displays the F0 values for a given sentence uttered in angry and in neutral styles: the angry melody (red curve) is much higher all over the sentence than the neutral melody (black curve).

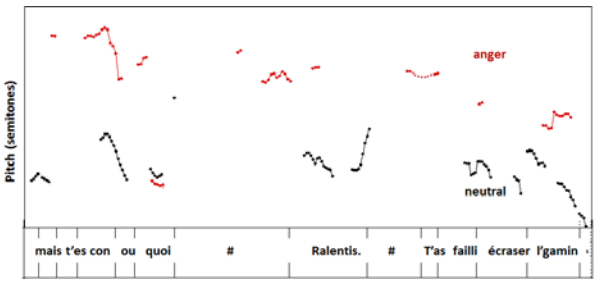

Figure 5: *F0 values measured on neutral (black) and on angry emotion (red) styles for a French sentence (text at bottom of figure, meaning "Are you a prick or what? Slow down. You've almost run over the kid.").*

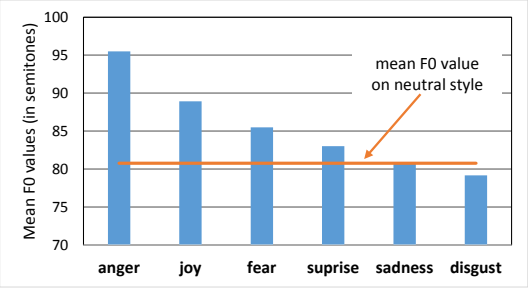

Figure 6: *Mean F0 value for the six emotion styles (bars), and for the neutral style (red line).*

Mean F0 values are calculated for each emotion style. Results are displayed in Figure 6 along with mean F0 value calculated on the neutral style (red line). Mean F0 value on the sadness emotion is the closest to mean F0 value of the neutral style; although surprise emotion (slightly higher mean F0) and disgust emotion (slightly lower mean F0) are also very close to the mean F0 value of the neutral style. The three remaining emotions (fear, joy and anger) have higher mean F0 values than the neutral style.

To get a more detailed analysis of F0 values, their distribution is reported in Figure 7 for each emotion style (blues curves). On each graph, the orange curve reports the distribution of F0 values on the same sentences uttered in a neutral style. These statistics rely on the quantification of the F0 values on a 14 degrees scale defined according to the F0 range of the speaker.

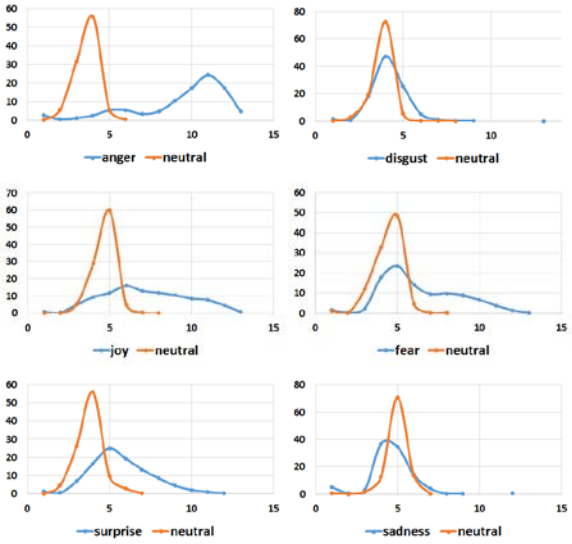

Figure 7: *Distribution of F0 values for each emotion style, and comparison to F0 values for the corresponding sentences in neutral style. The points on the curves indicate the frequency (vertical axis) of quantified F0 values that fall in each degree of the 14 degree scale (horizontal axis).*

For the disgust emotion, although its mean F0 value is slightly lower than the mean F0 value of the neutral style (see Figure 6), the F0 distribution curve is slightly enlarged to the right (i.e., towards higher F0 values) compared to the neutral curve. The distribution curves of fear, joy and surprise emotions are also moved to the right side having at least half of their F0 values higher than those measured in neutral style. As far as anger is concerned, there is almost no intersection between the distribution curves representing the neutral style (orange curve) and the anger emotion (blue curve). As for the sadness emotion, it consists of one exception for which the F0 distribution is slightly enlarged to the left (i.e., towards lower F0 values) compared to the neutral style.

4. Pauses and prosodic structures

4.1. Pauses

The number of pauses is compared between emotion data and the data uttered in a neutral style. For analysis purpose, the pauses are classified into three categories: short pauses (shorter than 250 ms), mid pauses (shorter than 400 ms) and long pauses (longer than 400 ms). Table 1 reports for each emotion the percentage of short, mid and long pauses in the columns "E". The columns "N" reports the percentages calculated on the same sentences uttered in a neutral style. It shows that the greatest difference is observed between anger and neutral styles,

and between joy and neutral styles. Indeed, the anger emotion uses a higher amount of short pauses and fewer long pauses. As for the joy emotion, a higher amount of short and mid pauses is used at the expense of long pauses.

Table 1. *Percentage of pause occurrences in emotion (columns "E") and neutral styles (columns "N").*

	Short pauses		Mid pauses		Long pauses	
	E	N	E	N	E	N
anger	63	17	9	7	28	76
disgust	13	9	2	1	85	90
joy	40	10	16	6	44	84
fear	35	33	9	4	56	63
surprise	12	12	5	3	83	85
sadness	36	30	2	4	62	66

As for the total number of pauses in the various emotion styles compared to the neutral style, their number is much higher in anger emotion while sadness and surprise are very close to the neutral style, and joy and fear emotions contain significantly less pauses than the neutral style.

4.2. Prosodic groups

A segmentation of utterances into prosodic groups is carried out using an automatic approach [29] based on [30], in which the decision on prosodic group frontiers relies on F0 slope, F0 level and vowel duration. Figure 8 displays, for each emotion style, the percentage of prosodic groups with respect to their length expressed in syllables. The results show that the anger emotion gives a preference to shorter prosodic groups compared to the other emotion styles. The disgust and the sadness emotion styles are the ones that use the longest prosodic groups.

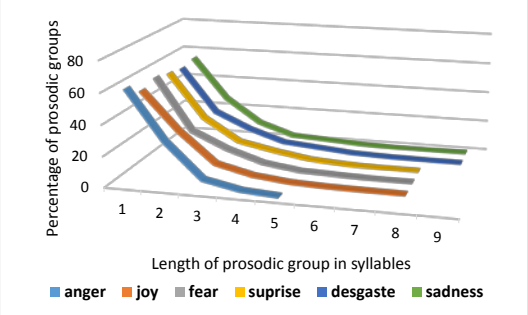

Figure 8: *Percentage of prosodic groups (vertical axis) with respect to the length in syllables (horizontal axis) for each emotion style.*

Table 2. *For each emotion style, comparison of the length of the prosodic groups with respect to the same sentences uttered in a neutral style.*

anger	shorter (++)
joy	longer (+)
fear	longer (+)
surprise	shorter
disgust	same
sadness	longer (+)

Table 2 summarizes the results of the comparison of the length of the prosodic groups between each emotion style and the same sentences uttered in a neutral style. Prosodic groups in the anger style are much shorter than prosodic groups in the corresponding sentences uttered in a neutral style. Sentences in surprise style contain slightly shorter prosodic groups. The sentences in the joy, fear and sadness emotions have longer prosodic groups than the corresponding sentences uttered in a neutral style. As for the disgust style, prosodic groups have the same length as in neutral speech.

5. Discussion

Table 3 summarizes the behavior of the parameters analyzed in this study. It displays for each emotion style the degree of lengthening (L) or shortening (S) for vowel and pause durations as well as the lengths of the prosodic groups, compared to the neutral style. It also displays the degree of higher (H) or lower (L) energy and F0 values measured on the vowels in comparison with the neutral style. For the anger emotion, all the parameters studied have a very different value in comparison to neutral data: heavy lengthening of vowel duration, short pauses, much higher energy and F0 values, and much shorter prosodic groups. Joy and fear emotion styles are also prosodically marked and are quite different from the neutral style. The disgust and sadness emotions have shorter vowel durations than neutral style, and the vowel energy is either lower than in the neutral style (as for sadness), or very similar to neutral style (as for disgust).

Table 3. *Summary of the parameter behavior in the different emotion styles, compared to the neutral style (see text for notations).*

	duration	pauses	energy	F0	Prosodic groups
anger	L+++	S++	H++	H+++	S+++
joy	L++	S+	H+	H++	L+
fear	S++	S++	H+	H++	L+
surprise	L+	-	H+	H+	S+
disgust	S+	-	-	-	-
sadness	S++	-	L+	-	L+

6. Conclusions

This paper provides an analysis of the behavior of the prosodic parameters on French speech data for various emotion styles: anger, fear, sadness, disgust, surprise, and joy. Beside the analysis of vowel duration, vowel energy and fundamental frequency in comparison to neutral reading style, the behavior obtained for pausing and prosodic phrasing is also investigated on the six emotional styles. This latter prosodic dimension is a phenomenon that deserves further investigation in order to correctly predict certain emotional prosodic features in speech synthesis systems. Indeed, statistical speech synthesis systems take benefit from adaptation techniques for modeling the impact and taking into account prosodic parameters; however such modeling relies on the sequence of units (sequence of phones and pauses, with their associated linguistic labels) as input both for training and for synthesis. Hence, pause distribution and lengths of the prosodic groups need to be predicted beforehand.

7. Acknowledgements

This study consists of a pilot study carried out within the ANR-JCJC project *SynPaFlex* (PI: Damien LOLIVE, 2015 edition).

8. References

[1] E. Couper-Kühlen. “An Introduction to English Prosody”. *Forschung & Studium Anglistik*, 1, Tübingen: Max Niemeyer & London: Edward Arnold, 1986.

[2] I. Fonagy. “*La vive voix*", Paris, Payot, 1983.

[3] I. Fonagy and K. Magdics. “Emotional patterns in intonation and music”, *Zeitschrift für Phonetik*, 16, pp. 293-326, 1963.

[4] K. R. Scherer. “Vocal correlates of emotion”, in *Handbook of psychophysiology: Emotion and social behavior*, Wagner H. & Manstead A. (éds), London, Wiley, pp. 165-197, 1989.

[5] K. R. Scherer, D. R. Ladd and K. E. A. Silverman. “Vocal cues to speaker affect: Testing two models”, *Journal of the Acoustical Society of America*, 76, pp. 1346-1356, 1984.

[6] Å Abelin and J. Allwood. "Cross linguistic interpretation of emotional prosody", in *ITRW workshop on Speech & Emotion*, Newcatle, UK, sept. 2000, pp. 110-113, 2000.

[7] D. Ververidis and C. Kotropoulos. “A review of emotional speech databases”, in *Panhellenic Conference on Informatics (PCI)*, pp. 560-574, 2003.

[8] V. Aubergé, N. Audibert and A. Rilliard. “Why and how to control the authentic emotional speech corpora”, in *INTERSPEECH’2003*, Geneva, Switzerland, 2003.

[9] D. Ververidis and C. Kotropoulos. "Emotional speech recognition: Resources, features, and methods”, *Speech communication*, 48(9), pp. 1162-1181, 2006.

[10] H. Gunes, B. Schuller, M. Pantic and R. Cowie. “Emotion representation, analysis and synthesis in continuous space: A survey”, in *FG 2011, IEEE Int. Conf. on Automatic Face & Gesture Recognition and Workshops*, pp. 827-834, 2011.

[11] D. Erickson. “Expressive speech: Production, perception and application to speech synthesis”, *Acoustical Science and Technology*, 26(4), pp. 317-325, 2005.

[12] M. El Ayadi, M. S. Kamel and F. Karray. “Survey on speech emotion recognition: Features, classification schemes, and databases”, *Pattern Recognition*, 44(3), pp. 572-587, 2011.

[13] R. B. Lanjewar and D. S. Chaudhari. “Speech Emotion Recognition: A Review”, *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 2, 2013.

[14] L. Chen, X. Mao, Y. Xue and L. L. Cheng. “Speech emotion recognition: Features and classification models”, *Digital Signal Processing*, 22(6), pp. 1154-1160, 2012.

[15] S. G. Koolagudi, N. Kumar and K. S. Rao. “Speech emotion recognition using segmental level prosodic analysis”, in *ICDeCom 2011, IEEE Int. Conf. on Devices and Communications*, pp. 1-5, 2011.

[16] M. Schröder. “Emotional speech synthesis: a review”. In *INTERSPEECH 2001*, Aalborg, Denmark, pp. 561-564, 2001.

[17] M. Schröder. “Expressive speech synthesis: Past, present, and possible futures”, in *Affective information processing*, Springer London, pp. 111-126, 2009.

[18] M. Bulut, S. S. Narayanan and A. K. Syrdal. “Expressive speech synthesis using a concatenative synthesizer”, in *INTERSPEECH 2002*, Denver, Colorado, USA, 2002.

[19] N. Audibert, V. Aubergé and A. Rilliard. “The prosodic dimensions of emotion in speech: the relative weights of parameters”, in *INTERSPEECH’2005*, Lisbon, Portugal, pp. 525-528, 2005.

[20] A. Iida, N. Campbell, F. Higuchi and M. Yasumura. “A corpus-based speech synthesis system with emotion”, *Speech Communication*, 40(1), pp. 161-187, 2003.

[21] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza and M. Picheny. “The IBM expressive text-to-speech synthesis system for American English”, *IEEE Trans. on Audio, Speech, and Language Processing*, 14(4), pp. 1099-1108, 2006.

[22] J. Yamagishi, T. Masuko and T. Kobayashi. “HMM-based expressive speech synthesis-Towards TTS with arbitrary speaking styles and emotions”, in *Special Workshop in Maui*, Maui, Hawaı, 2004.

[23] R. Tsuzuki, H. Zen, K. Tokuda, T. Kitamura, M. Bulut and S. Narayanan. “Constructing emotional speech synthesizers with limited speech database”, in *ICSLP’2004*, Jeju Island, Korea, vol. 2, pp. 1185-1188, 2004.

[24] D. N. Jiang, W. Zhang, L. Q. Shen and L. H. Cai. “Prosody analysis and modeling for emotional speech synthesis”, in *ICASSP 2005 - IEEE International Conference on Acoustics, Speech, and Signal Processing*, March, Philadelphia, Pa., USA, Proceedings, vol. 1, pp. 281-284, 2005.

[25] J. Tao, Y. Kang and A. Li. “Prosody conversion from neutral speech to emotional speech”, *IEEE Trans. on Audio, Speech, and Language Processing*, 14(4), pp. 1145-1154, 2006.

[26] Z. Inanoglu and S. Young. “Data-driven emotion conversion in spoken English”, *Speech Communication*, 51(3), pp. 268-283, 2009.

[27] *“Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; extended advanced front-end feature extraction algorithm; compression Algorithms”*, ETSI ES 202 212, 2005.

[28] K. Bartkova, E. Delais-Roussarie and F. Santiago. « Prosotran: a tool to annotate prosodically non-standard data”, in *Speech Prosody 2012*, Shangaï, pp. 55-58, 2012.

[29] K. Bartkova and D. Jouvet. “Automatic detection of the prosodic structures of speech utterances”, in *SPECOM 2013, 15th, International Conference on Speech and Computer*, pp. 1-8, 2013.

[30] P. Martin.; “Prosodic and rhythmic structures in French”, *Linguistics* 25, pp. 925–949, 1987.