CONTENT-BASED TRANSFORMATION OF THE EXPRESSIVITY IN SPEECH

Grégory Beller, Xavier Rodet

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

ABSTRACT

In this paper we describe a transformation system for speech expressivity. It aims at modifying the expressivity of a spoken or synthesized neutral utterance. The phonetic transcription, the stress level and the other information about the corresponding text supply a sequence of contexts. Every context corresponds to a set of parameters of acoustic transformation. These parameters change along the sentence and are used by a phase vocoder technology to transform the speech signal. The relation between the transformation parameters and the contexts is initialized by a set of rules. A Bayesian network transforms gradually this rule-based model into a data-driven model according to a learning phase involving an expressive French speech database. The system functions for French utterances and several acted emotions. It is employed at artistic ends for multi-media applications, the theater and the cinema.

Keywords: emotions, expressivity, speech, transformation, Bayes

1. INTRODUCTION

The capacity to express and to identify emotions, intentions and attitudes through the modulation of the parameters of the voice is essential in the human communication. It seems that all these controlled or uncontrolled aspects [13] belong to more than one category. We group them in the term *expressivity* whether they are simulated or not.

Current speech synthesis methods provide speech with good naturalness and intelligibility. Art directors, contemporary composers and film dubbing studios are now interested by the multiple possibilities of a system which offers to analyze, to synthesize and transform the *expressivity* of the voice [2]. Statistical models of emotional prosody have been used by voice conversion systems [8] as well as by speech synthesizers [5]. Our system should change *expressivity* of a sentence as an actor does. Therefore we have recorded French actors to build an expressive speech database. Then recordings are studied according to the five dimensions of prosody [12]: intonation, intensity, speech rate, degree of reduction and voice quality.

A first study on speech rate [3] has shown the importance of the stress level of syllables. For instance, stressed syllables last much longer than unstressed in the case of happiness whereas all syllables last approximately the same duration in the case of fear. This categorization of syllables helps to analyze and modify *expressivity*. The degree of reduction is also influenced by *expressivity*. In order to analyze it (to draw a vocalic triangle, for instance), we need at least the phonetic labels of the vowels. This level of annotation offers a categorization in phonetic classes in which the spectra of the corresponding vowels can be compared. Thus the degree of reduction can be estimated for all utterances independently of the phonetic context and then used to compare *expressivity*.

Indeed, one major difficulty in the analysis of para-linguistic features is the influence of the verbal content of the sentence. Context-dependent categorization is a useful tool to analyze para-linguistic aspects of speech. The recordings divided in linguistic units can be classified according to phonetic label, stress level or other symbolic information. Statistics of acoustic values can be estimated in each class. Then statistic values between the different classes can be compared and related to various expressivities.

After a quick overview of the system, of the database and of the involved features, this article explains how context dependent acoustic transformations are inferred with a Bayesian network and applied to the speech signal to modify *expressivity*.

2. GENERAL OVERVIEW

All the involved processes are shown in figure 1. They use two information levels conveyed by speech. On one hand, the linguistic part of the spoken message, i.e. the text and supplementary information, such as expressivity, provide symbolic discrete data noted as S_x. On the other hand, the acoustic realization of this text, i.e the recorded speech, gives continuous acoustic data noted as A_x.

The stage of segmentation is essential because it connects the acoustic data to the symbolic units.

3. DATABASE

3.1. Database content

Our database of expressive French speech is composed of recordings of four actors ($S_{speaker}$), two males and two females (S_{gender}), during approximately an hour and a half each. They were all recorded in the same professional conditions following an identical procedure. Ten sentences (S_{text}) extracted from a phonetically balanced corpus [7] have been marked with prosodic boundaries using punctuation and underlined parts of words. Chosen *expressivities* (S_{exp}) were acted emotions: *neutral, introvert and extrovert anger, introvert and extrovert happiness, introvert and extrovert fear, introvert and extrovert sadness*, as well as *positive and negative surprises, disgust, discretion, excitation and confusion*. Each sentence was pronounced in all the *expressivities*. Furthermore, in the case of acted emotions, every sentence was repeated six times with an increasing degree (power) of *expressivity* (S_{degree}). Finally, the corpus is composed of approximately 550 utterances per actor. Some *fillers* have been also uttered for each expressivity.

3.2. Database interface

Since a large amount of data has been used for this study (see section 5.), an efficient database architecture is required. The free, open source, relational database postgreSQL is accessed by a database interface written in Matlab®. Thus, the database can be browsed with a graphical explorer that allows users to visualize analyses and play units of the database. For the latter, the SDIF format is used to store acoustic analysis results and the XML format for all symbolic data. Finally the database is used to reliably store thousands of sounds and data files, tens of thousands of segmented units and their interrelationship.

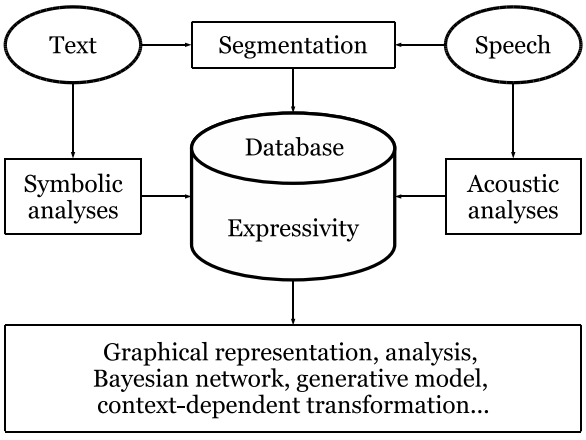

Figure 1: Overview of the system.

3.3. Database segmentation

The first step of the analysis process is the segmentation of recorded utterances, in variable length units. The automatic segmentation method used is classical and achieved by a Hidden Markov Model trained on a *neutral* multi-speaker database [9]. Then this initial segmentation has been hand-corrected by phoneticians using wavesurfer [14] to better match the phonetic realization which is often different from the automatically predicted phonemic transcription in the case of expressive speech. This provides XSAMPA phonetic labels (S_{phonem}) that are used by post-processings (see 4.1.) to define boundaries and durations ($A_{duration}$) of other unit types: *syllable, prosodic group, phrase* and *sentence*.

4. DESCRIPTORS

All these units are annotated with three types of descriptors.

4.1. Symbolic descriptors

Category descriptors express the membership of a unit to a category or class and to all its base classes in the hierarchy (e.g. speaker → actor → male for the sound source hierarchy). The syntactic description of the text is provided by the LIAPHON program [1]. Then a syllabifier is used and prosodic boundaries are defined using an automatic text-based prediction of the stress level of syllables, corrected by hand (S_{stress}). These descriptions and symbolic descriptors defining the relative place of a unit with respect to others are stored in XML files.

4.2. Dynamic descriptors

Dynamic descriptors are acoustic analysis data varying during the time span of a unit.

4.2.1. Fundamental frequency and energy

Pitch curve, also called intonation contour, is a prominent perceptual cue for *expressivity*. Fundamental frequency (A_{f0}) is calculated by the YIN algorithm [6]. This algorithm also gives the energy (A_{energy}) and the harmonic to noise ratio (also called aperiodicity) of the signal for each computed frame.

4.2.2. Speech rate

Local speech rate is defined from the duration of syllables [3] (see figure 2). Contrary to the well defined mean speech rate computed over an entire utterance

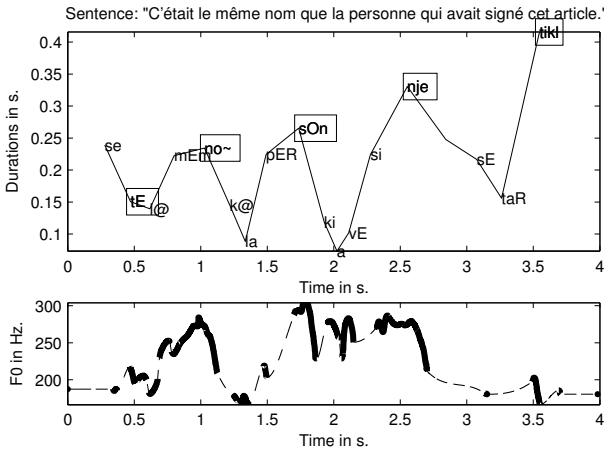

Figure 2: Durations and fundamental frequency of syllables of a French sentence pronounced with happiness: "It was the same name as the person who had signed up this paper."

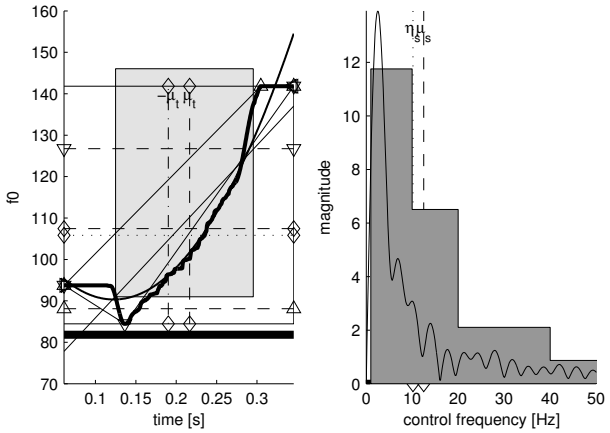

Figure 3: Example of *characteristic values* of the fundamental frequency computed over a syllable pronounced with introvert anger.

and measured in syllables per second, we keep on the syllable duration as the unity. Because the most prominent syllables have often a longer duration, the speech rate curve is thus shown by the durations of syllables which draw an "instantaneous" view of the evolution of the speech rate. A deceleration corresponds to a rising of the curve and an acceleration is represented by a falling of the curve.

4.2.3. Formant parameters

The formant trajectory estimation algorithm [2] uses the group delay function [11] and is based on three hypotheses:

1. Formants deal with the most important poles of an AR estimation.
2. Formants can be classified according to their frequencies.
3. Formant's trajectories are continuous in the time-frequency space.

An observation probability matrix (Hyp_1) is made up of weighted characteristics: *frequency, group delay, intensity of the spectrum around the frequency and corresponding bandwidth*. Formant trajectories ($A_{formant}$) are decoded recursively (Hyp_2) by a Viterbi algorithm insuring continuity (Hyp_3).

4.3. Static descriptors

The evolution of dynamic descriptor is modeled over the time span of a unit by a vector of *characteristic values* (see figure 3):

- minimum, maximum, range, slope, arithmetic mean, geometric mean, standard deviation
- 2^{nd} order Legendre polynomial approximation giving the slope and the curvature
- temporal center of gravity giving the location of the most important elevation in the descriptor curve
- Fourier spectrum of the descriptor named *control frequency*, and its first 4 order moments. *Jitter* and *Shimmer* are respectively related here to deviations of the gravity centers of the Fourier spectra of f0 and energy.

5. STATISTICAL MODEL

Bayesian networks have been used for several speech processing purposes. A Bayesian network models dependencies between some discrete and continuous variables. It is composed of a qualitative description represented by a graph the structure of which can be learned or given arbitrarily (see figure 4), and of a quantitative description represented by a generalized (joint) probability density function: $GPDF = P(A, S)$, estimated by an iterative Expectation-maximization algorithm [10].

5.1. Estimation/Learning

A context is defined as the set of the symbolic variables (see figure 4) that can take different states in closed vocabularies. An example of such a context: C_i = {S_{gender}="male"; S_{exp}="extrovert fear"; S_{degree}="3"; S_{stress}="unstressed"; S_{phonem}="/e/"}. A normal distribution of each acoustic *characteristic values* is computed on clustered data for every contexts.

Our relatively small database does not present all the plausible observable contexts that would be required to transform any new sentence. To compute reliable distributions of acoustic parameters for a

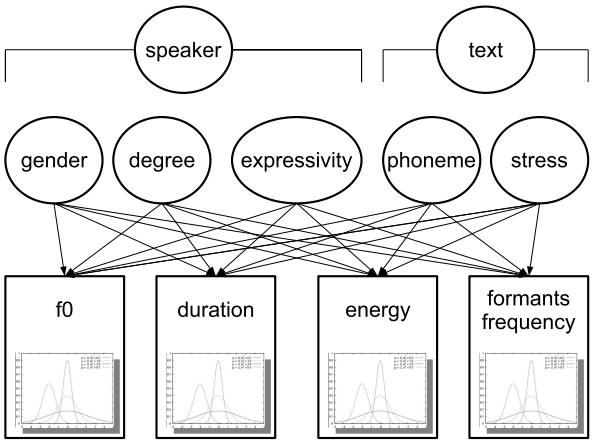

Figure 4: Bayesian network used for learning and inference steps.

context that has not been presented during the learning step, we complete the statistical model by analogy. The acoustic data of an unobserved context are used to infer the most probable observed context of the database that can explain them. This analog context is modified (*expressivity*) and used for another inference step.

5.2. Inference/Transformation

Our statistical model is aimed at transforming a given *neutral* utterance into the same sentence but with a given *expressivity E*. First, symbolic descriptors, forming a sequence of contexts, are computed on the neutral sentence. Then two acoustic descriptors sets are inferred with the Bayesian network, one using expressivity *neutral* and the other using expressivity *E*. Inferred acoustic parameter distributions are then compared so as to provide transformation factors. After a smoothing step of transformation parameters, a phase vocoder technology [4] transforms the *neutral* speech signal according to these parameters.

6. CONCLUSION

In this paper, we have described a transformation system of speech *expressivity*. A statistical model is learned on an expressive multi-speaker database in a Bayesian Network. Acoustic transformations parameters are time-varying and dependent of symbolic contexts extracted of the text and of a speaker state definition. It has been shown how a Bayesian network achieves the mutation of a rule-based model into a data-driven model. Even if all the possibilities and the improvements evoked in this paper were not estimated yet, the system seems to be adequate to the problem. It is now working for several acted emotions in French. Some examples can be listened to at the following address: http://recherche.ircam.fr/equipes/analyse-synthese/beller. Future work will be focused on quantifying the results by perceptual tests and introducing voice quality as an acoustic modifiable parameter.

7. ACKNOWLEDGMENTS

This work was partially funded by the French RIAM network project VIVOS. The authors would like to thanks the actors involved in this study for their performances.

8. REFERENCES

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

[2] Beller, G. 2007. Influence de l'expressivité sur le degré d'articulation. *RJCP, Rencontres Jeunes Chercheurs de la Parole*.

[3] Beller, G., Schwarz, D., Hueber, T., Rodet, X. may 2006. Speech rates in french expressive speech. *Speech Prosody* Dresden. SproSig ISCA.

[4] Bogaards, N., Roebel, A., Rodet, X. Novembre 2004. Sound analysis and processing with audiosculpt 2. *International Computer Music Conference (ICMC)* Miami, USA.

[5] Bulut, M., Lee, S., Narayanan, S. 2007. a statistical approach for modeling prosody features using pos tags for emotional speech synthesis. *ICASSP*.

[6] de Cheveigné, A., Kawahara, H. 2002. Yin, a fundamental frequency estimator for speech and music. *JASA* 111, 1917–1930.

[7] Combescure, P. 1981. 20 listes de dix phrases phonétiquement équilibrées. *Revue d'Acoustique* 56, 34–38.

[8] Hsia, C.-C., Wu, C.-H., , Wu, J.-Q. 2007. conversion function clustering and selection for expressive voice conversion. *ICASSP*.

[9] Morris, A. 2006. Automatic segmentation. Technical report IRCAM.

[10] Murphy, K. 2001. The bayes net toolbox for matlab. *Computing Science and Statistics* volume 33.

[11] Murthy, H., Murthy, K. M., Yegnanarayana, B. 1989. Formant extraction from phase using weighted group delay function. *Electronics Letters* volume 25. IEE 1609–1611.

[12] Pfitzinger, H. 2006. Five dimensions of prosody: Intensity, intonation, timing, voice quality, and degree of reduction. Hoffmann, H., R.; Mixdorff, (ed), *Speech Prosody* number 40 in Abstract Book Dresden. 6–9.

[13] Scherer, K. 1989. *Handbook of Psychophysiology: Emotion and Social Behavior* chapter Vocal correlates of emotion, 165–197. London, Wiley.

[14] Sjölander, K., Beskow, J. 2000. Wavesurfer - an open source speech tool. *International Conference on Spoken Language Processing*.