Context Dependent Transformation of Expressivity in Speech Using a Bayesian Network

Grégory Beller

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

Abstract

In this paper we describe a transformation system of 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 French expressive database. The system functions for French utterances and several acted emotions. It is employed at artistic ends for the multi-media, the theater and the cinema.

1. Introduction

The capacity to express and to identify emotions, intentions and attitudes through the modulation of the parameters of the voice is prevailing for human communication. It seems that all these controlled or uncontrolled aspects [1] 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 [3] as well as by speech synthesizers [4, 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 [6]:

- intonation (fundamental frequency)
- intensity
- local speech rate (syllable duration)
- degree of reduction (formant parameters)
- voice quality: estimation of the glottal excitation signal (not yet involved in this study)

A first study on speech rate [7] 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 paralinguistic 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*.

Our first expressive transformation system was rule-based like many others [8]. Transposition ratios, time-stretch factors and gain have been chosen according to mean values estimated on the database. For example, "to transform a given utterance from neutral to happy, transpose voiced segments one octave up" is one of the rules that has been hand-written and applied. In order to keep this knowledge while giving more complexity to the model and making it closer to the data, we have decided to enrich these rules by the use of machine learning algorithms. In our new system, transformation parameters are learned in a Bayesian network. An initial rule-based model is partly moved into a data-based model according to the amount of observed data.

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*. The *neutral* sentence which we wish to transform can be, either recorded or produced by a Text-To-Speech synthesizer which supplies then, the phonetic segmentation.

2. Context-dependent model

All the involved processes 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*, give symbolic discrete data noted $S_{variable}$ (repre-

sented by circles in figure 1). On the other hand, the acoustic realization of this text, e.g the recorded speech, gives continuous acoustic data noted $A_{variable}$ (represented by rectangles in figure 1). The segmentation step is predominant as it connects acoustic data to symbolic units. Features involved in the statistical model are thus noted S or A.

2.1. Goals of a generative model

Our system is aimed at transforming a given *neutral* utterance into the same sentence but with a given *expressivity E* with a given expressive *degree D*. First, symbolic descriptors are computed on each phone of the neutral sentence (see section 4). This provides a temporal sequence of *context C* (see section 2.2). Then two corresponding acoustic descriptors sets are predicted, one using *expressivity neutral* and the other using *expressivity E*. Inferred acoustic parameter distributions are then compared so as to provide transformation factors. Hence the problem becomes to infer acoustic data A corresponding to a given context $S = C_i$, i.e. to evaluate $P(A|S = C_i)$.

2.2. Context definition

The context is defined as the set of the symbolic variables that can take different states in closed vocabularies. An example of such a context C_1:

$$C_1 = \begin{cases} S_{gender} & = \text{“}male\text{”} \\ S_{exp} & = \text{“}neutral\text{”} \\ S_{degree} & = \text{“3”} \\ S_{stress} & = \text{“}unstressed\text{”} \\ S_{phonem} & = \text{“/œ/”} \end{cases}$$

These variables can be derived from higher level variables such as $S_{speaker}$ and S_{text}. Table 1 shows the number of states they can take (*cardinality*):

These symbolic variables are supposed independent since phonetic, stress, *expressivity* and speaker related information levels can occur in any combination (assumption discussed in section 4.3). Thus the *Universe* $\mathcal{U}$ of the context-dependent model is composed of 9180 possible contexts per gender.

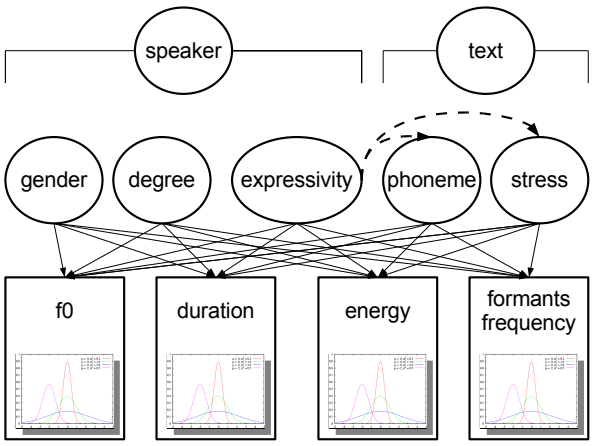

Figure 1: Bayesian network: discrete (circles) and continuous (rectangles) involved variables and their dependencies.

variable	cardinality	state description
S_{gender}	2	female or male
S_{exp}	15	*expressivities* (see section 4.2)
S_{degree}	6	degree or power of *expressivity*
S_{stress}	3	unstressed, secondary stressed, primary stressed
S_{phonem}	34	phonemes (XSAMPA code)

Table 1: Symbolic variables, cardinalities and state descriptions

3. From a rule-based model toward a data-driven model

For artistic purposes, a model needs to be flexible in the sense that most users would like to change understandable and modifiable parameters. A rule-based model is a good start and can be hand-designed. But the cardinality of $\mathcal{U}$ is too large to keep a simple set of rules. Contrary to a rule-based model, a data-driven model can manage many different contexts but suffers from a lack of controllability and from a lack of generalization (the model stay too close to the data). A statistical parametric model is employed because it allows to reconcile these two approaches [5]. In order to make our rule-based model more precise and fitting real data better, we use the Bayes paradigm. After the learning phase, the rule-based model is partially turned into a statistical data-driven model according to the amount of clustered data per context. We briefly present the transition between models which corresponds to our chronological use (see reference [9] for more details).

3.1. Rule-based model

Our first attempt in *expressivity* transformation was based on cumulative hand-designed rules created on a case by case examination of the database. For instance, all *neutral* unstressed “/œ/” with context C_1, are lengthened by a factor 1.5 to be transformed in *extrovert sadness* (context C_2) and again by a factor 1.8 for stressed ones (context C_3). Such a rule can be written:

$$\begin{aligned} A_{duration}(C_3) &= 1.8 \times A_{duration}(C_2) \\ &= 1.8 \times 1.5 \times A_{duration}(C_1) \end{aligned}$$

But the cardinality of the Universe $\mathcal{U}$ makes the task of constructing a rule-based model very complex. This is the reason why we chose a machine learning paradigm involving a database of examples.

3.2. Frequentist approach

The database employed is described in section 4. Let:

- $X = \{X_{(l)}\}_{l=1..N}$ be the set of N observed data
- θ be the parameters of the model
- S_{acous} be a discrete acoustic variable, instead of a continuous one, used for a better explanation.

If all variables are fully observed, i.e. that we have a measure of the acoustic variables A for all possible contexts $S \in \mathcal{U}$, the simplest method is to evaluate the probability of an event $(S_{acous} = S_j)$ by the frequency of appearance of that event within the same context $(S = C_i)$. This approach, called the maximum likelihood (ML), gives:

$$\hat{P}(S_{acous} = S_j|S = C_i) = \hat{\theta}_{i,j}^{ML} = \frac{N_{i,j}}{\sum_j N_{i,j}} \qquad (1)$$

where $N_{i,j}$ is the number of times that $S_{acous} = S_j$ in the context $S = C_i$.

3.3. Bayesian approach

Bayesian probability is a formalism that allows us to reason about beliefs under conditions of uncertainty. Similar to frequentist approach, it possesses an augmented optimization objective which incorporates a prior distribution over the quantity one wants to estimate. It consists of finding the most probable parameters θ knowing that the data have been observed and using a prior on these parameters. This approach, called expectation a posteriori (EAP), gives:

$$\hat{P}(S_{acous} = S_j|S = C_i) = \hat{\theta}_{i,j}^{EAP} = \frac{N_{i,j} + \alpha_{i,j}}{\sum_j (N_{i,j} + \alpha_{i,j})} \quad (2)$$

where α_k are the parameters of a Dirichlet distribution associated to the prior $P(S_{acous} = S_j|S = C_i)$. α_k are the parameters controlling the weights of the rule-based model and of the data-driven model into the final model. If $\alpha_{i,j} \to \infty$, the final model is completely influenced by the prior and if $\alpha_{i,j} \to 0$, the final model is completely influenced by the data. The prior is thus defined by a ratio between a controllable number of simulated cases and the fixed number of really observed cases. Note there exists also the maximum a posteriori (MAP) approach.

3.4. Bayesian network

Bayesian networks have been used for several speech processing purposes. Emotion recognition [10] use both naïve bayes classifier or dynamic Bayesian networks. 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 1), and of a quantitative description represented by a generalized (joint) probability density function:

$$GPDF = P(A, S) \quad (3)$$

3.4.1. Qualitative part: Graphical model

The structure of the graphical model is arbitrarily given and presented on figure 1. This qualitative view of the statistical model shows the variables involved during learning and inference steps. Circles represent discrete variables related to the symbolic context and rectangles represent continuous variables that are vectors of *characteristic values* (see section 5.1) computed on dynamic acoustic descriptors. Arrows represent dependencies between variables.

3.4.2. Quantitative part: Generalized probability density function: $P(A, S)$

The Generalized Probability Density Function $GPDF$ quantifies all dependencies between variables of a Bayesian network. Each continuous variable is assumed to follow a Linear Conditional Gaussian (LCG) distribution, conditional to the configuration of its discrete parent variables. $GPDF$ is estimated by the use of the Bayes rule:

$$P(A, S) = P(A|S)P(S) \quad (4)$$

Once the $GPDF$ is estimated (learning step), the LCG distributions $P(A|S = C_i)$ of acoustic data are inferred using equation 4 (inference step).

4. Database

In order to estimate a $GPDF$, we recorded a database of French expressive speech. It presents 3996 contexts per gender. $\mathcal{U}_{observed}$ is thus none exhaustive and covers less than the half of the Universe $\mathcal{U}$. This lack of data raises difficulties when a new sentence (context) is presented to the system. This problem is considered and partially resolved in the section 5.3.

4.1. Recordings

The database is composed of recordings of four actors ($S_{speaker}$), two males and two females (S_{gender}), during approximately one 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 [11] have been marked with prosodic boundaries using punctuation and underlined parts of words.

4.2. Expressivities

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*.

4.3. Phonetic segmentation

The first step of the analysis process is the segmentation of recorded utterances, in phones (S_{phonem}).The automatic segmentation method used [12] is classical and achieved by a Hidden Markov Model trained on a *neutral* multi-speaker database [13]. 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. In fact, the phonetic realization of an expressive utterance is influenced by *expressivity* as it is shown in section 6.1.

4.4. Prosodic segmentation

Phonetic segmentation provides XSAMPA labels that are used by post-processings to define boundaries and durations ($A_{duration}$) of other unit types: *syllable, prosodic group, phrase* and *sentence*. A rule-syllabifier uses phonetic labels to define syllable boundaries. Prosodic boundaries are defined using an automatic text-based prediction of the stress level of these 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 which allows the storage of hierarchical relationships.

4.5. Acoustic descriptors

All the labeled units are segment of analyzed speech signal. Dynamic descriptors are acoustic analysis data varying during the time span of a unit.

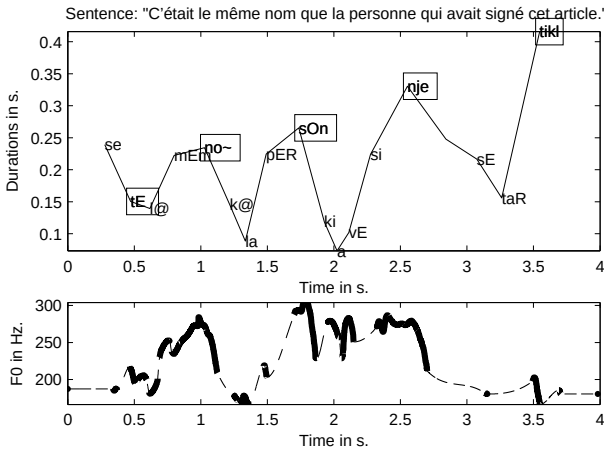

Figure 2: Duration and fundamental frequency of syllables of a French sentence pronounced with extrovert happiness: "It was the same name as the person who had signed up this paper". Squared syllables are stressed.

4.5.1. 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 [15]. This algorithm also gives the energy (A_{energy}) and the harmonic to noise ratio (also called aperiodicity) of the signal for each computed frame. Fundamental frequency is interpolated within unvoiced segments the boundaries of which are defined by a threshold process on the aperiodicity (see figure 2).

4.5.2. Speech rate

Local speech rate is defined from the duration of syllables [7] (see figure 2). Contrary to the well defined mean speech rate computed over an entire utterance 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.

4.5.3. Formant frequencies

The formant parameters are computed by an estimation algorithm of formant trajectories [2]. At first, the method finds the poles of an auto regressive model, estimated on the LPC of the time-framed windowed signal. Then, it defines the most important poles according to several criteria of which the group delay [16]. Finally, it makes correspond some of these poles to formants, while assuring that the trajectories of formants are smoothed in the time-frequency space. Trajectories are decoded recursively thanks to dynamic programming.

5. Transformation

Once the learning step is over ($GPDF$ estimated), a new sentence can be presented. Phonetic, stressing and other contextual information such as the wished *expressivity*, build up a sequence of symbolic contexts. Acoustic variables parameters are then inferred for each phoneme taking into account its context. The processes are summarized in the algorithm 1.

5.1. Temporal model

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

- start, middle, end, minimum, maximum and range values
- arithmetic mean, geometric mean, standard deviation
- temporal center of gravity/anti-gravity giving the location of the most important elevation or depression in the descriptor curve
- 2^{nd} order Legendre polynomial approximation giving the slope and the curvature
- inflexion point corresponding to the target value occurring at the time the derivative of the 2^{nd} order approximation reaches zero or at the middle if evolution is linear.
- Fourier spectrum and spectral centroid of the descriptor (not represented in figure 3), related to descriptor rapid or slow movements, and oscillation. *Jitter* is related here to a deviation of the gravity center of the Fourier spectrum of f0. *Shimmer* is related to a similar deviation for energy.

Some of these *characteristic values* are only computed for analysis. Up to here, all acoustic variable is correspond to the value of its inflexion point.

5.2. Inference

Two inference steps give plausible acoustic realizations of the *neutral* utterance and of the *expressive* one. Comparisons of these two sets of acoustic data lead to transposition, time-stretch, gain and spectral frequency warping factors that evolve along the sentence since context changes at each phone. After a

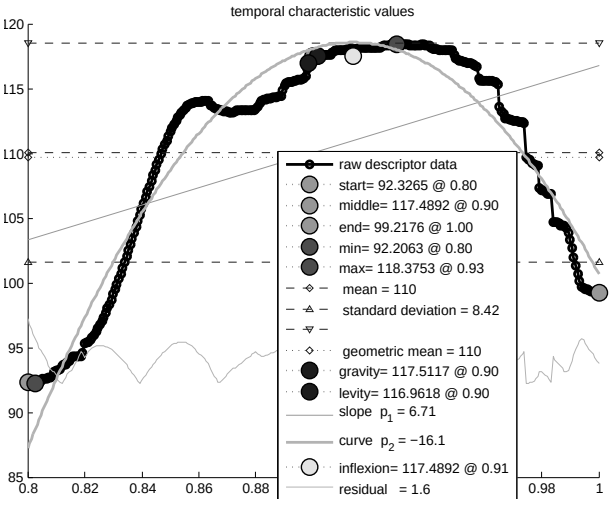

Figure 3: Example of *characteristic values* of the fundamental frequency (in [Hz]) computed over a vowel time span (in [s]) pronounced with introvert anger.

smoothing step of transformation parameters, a phase vocoder technology [17] transforms the *neutral* speech signal according to these parameters.

5.3. Unobserved contexts

A new sentence can present a context which was not observed during the learning phase because our data base does not cover all the universe $\mathcal{U}$. In that case, our generative model has all the same to propose a solution and supply parameters of transformation. This is achieved by using two inference steps. The first one used acoustic data of the *neutral* utterance to infer the most probable known context. The *expressivity* and the degree of *expressivity* wished are added/modified to this context. A second inference allows then to predict acoustic data. The learning does not thus allow the generalization for all the contexts, what is a crucial question which exceeds the subject of this article. Nevertheless, the solution presented here allows to deduce parameters of transformation for all possible contexts by proceeding by analogy.

→**Initialization**
- $GPDF$ estimated (learning step over) ;
- observed contexts: $\mathcal{U}_{observed}$;
- new *neutral* [N] sentence (audio and text) ;
- desired *expressivity* [E] with *degree* [D] ;

→**Analyses**
- segmentation in P phones ;
- context sequence definition: $\{C_N(t)\}_{t\in[1:P]}$;
- acoustic analyses ;
- computation of *characteristic values* ;

→**Inference of acoustic data**

for $t \in [1:P]$ **do**
- check if context has already been observed:

if $C_N(t) \in \mathcal{U}_{observed}$ **then**
inference of *neutral* acoustic data $A_N(t)$ with context $C_N(t)$;

else
inference of *neutral* context $C_N(t)$ with acoustic data $A_N(t)$;

end
- compute *expressive* context $C_E(t)$:

$$C_E(t) = C_N(t)$$

$$C_E(t) = \begin{cases} S_{exp} & = \text{E} \\ S_{degree} & = \text{D} \end{cases}$$

- inference of *expressive* acoustic data $A_E(t)$ with context $C_E(t)$;

end

→**Transformation parameters**
- compute transformation parameters $T_{N\to E}$ from A_E and A_N
- Smooth/filter transformation parameters $T_{N\to E}$;

→**Transformation of the speech signal**
- dynamic transposition ;
- dynamic time-stretching ;
- dynamic gain ;
- dynamic frequency warping ;

Algorithm 1: Algorithm of the transformation of a new sentence

This procedure leads to a prediction of acoustic transformation parameters for any symbolic context even if it was not observed before, by use of analogy [18].

6. Discussions

Learning with a Bayesian network has a number of advantages over a rule-based system. Our previous rule-based model is kept and used to initialized the data-driven model. Application of the Bayesian rule allows to compute acoustic parameter distributions that fit the data according to the amount of observations. The Bayesian network gradually and partially turns the rule-based model into a data-driven model. The Matlab® Bayesian Network Toolbox [19] efficiently computes conditional probability distributions of either discrete or continuous variables. This heterogeneity of the nature of the descriptors gives the means to the model of being context-dependent. This Bayesian network approach raises several questions, especially on the independency of variables as mentioned in 4.3.

6.1. Interdependencies of variables within a context

Phoneticians that have corrected the segmentation of the database, have observed that for several *expressivities*, some expected phonemes were undershot, absent or added and required re-labeling (An open "/E/" could sound like a "/œ/", for instance). This means that, even if the same text has been pronounced, the S_{phonem} tabular distribution is different according to *expressivity*. $P(S_{phonem}|S_{exp})$ has been estimated by adding an arrow in the graph: $S_{exp} \to S_{phonem}$ (dotted arrow in figure 1). The mean frequencies of the appearance of some phonological classes per sentence, computed with two actors data sets (one male and one female), are represented by height of bars in the figure 4. It shows how *expressivity* influences the prononciation of the text. A study on the correspondences between the wished phonemic transcription and the phonetic annotation of the realized pronunciation supplies more information, but exceeds the scope of this article.

A similar context interdependency occurs at stress level: $S_{exp} \to S_{stress}$ (dotted arrow in figure 1). For instance, in the case of *extrovert anger*, almost all syllabes could be perceived stressed, since they are separated by caesura.

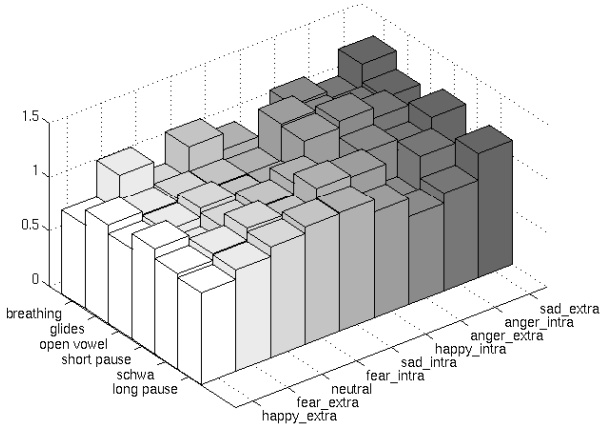

Figure 4: Tabular distributions of the mean frequency of appearance of phonological classes per sentence. The same text has been pronounced with different *expressivities*.

6.2. Interdependencies between acoustic variables

A second strong interrelationship is the interdependence between acoustic variables. For instance, the variance of f0 seems highly correlated to the speech rate. The degree of reduction is also often raised when speech is accelerated, although the contrary has been observed for several *expressivities* like *anger* (see reference [7]). Thus relations like $A_{f0} \leftrightarrows A_{duration}$ should be added.

6.3. Dependencies between successive contexts

Hidden Markov Models are widely used in speech recognition and share the same formalism as Bayesian networks (graphical models). It has been shown that the knowledge of the probability of transitions between phonemes increases recognition rate. Therefore, we expect an improvement in prediction results by adding connections from the previous context: $S(i-1) \rightarrow S(i)$. Hence coarticulation could also be modeled by a dynamic Bayesian network.

6.4. Expressivity variable: discrete or continuous?

Finally, the nature of S_{exp} can be discussed because there are various representations of the *expressivity*, of which some are category-specific (discrete) and the others continuous [8]. No strong motivation has yet guided our choice and it is still an open question. However, we envision to replace the discrete expressive degree variable S_{degree} by a continuous variable. This would lead to LCG distribution the parameters of which (means) are linearly dependent (W_i) of the expressive degree:

$$P(A_{f0}|S_{degree} = d, S = C_i) = \mathcal{N}(\mu_i + W_i \times d, \sigma_i) \quad (5)$$

7. Conclusion

In this paper, we have described a transformation system of speech *expressivity*. A statistical model is learned on a multi-speaker expressive database in a Bayesian Network. Acoustic transformation 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 possibilities and improvements evoked in this paper have not been tested yet, the reliability of this learning algorithm to the problem and its relative transparency/interpretability make the system full of promises. It is now working for several acted emotions in French. Some examples can yet be listened to at the following address: `http://recherche.ircam.fr/equipes/analyse-synthese/beller`. Future work will now be focused on quantifying the results by perception tests and introducing voice quality as an acoustic modifiable parameter.

8. 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.

9. References

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

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

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

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

[5] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in hmm-based speech synthesis," in *IEICE Trans. on Inf. & Syst.*, vol. E88-D, no. 3, March 2005, pp. 503–509.

[6] H. Pfitzinger, "Five dimensions of prosody: Intensity, intonation, timing, voice quality, and degree of reduction," in *Speech Prosody*, ser. Abstract Book, H. Hoffmann, R.; Mixdorff, Ed., no. 40, Dresden, 2006, pp. 6–9.

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

[8] M. Schröder, "Emotional speech synthesis–a review," in *Eurospeech, Aalborg*, DFKI, Saarbrücken, Germany: Institute of Phonetics, University of the Sarland, 2001, pp. 561–564.

[9] P. Naïm, P.-H. Wuillemin, P. Leray, O. Pourret, and A. Becker, *Réseaux bayésiens*. Paris: Eyrolles, 2004.

[10] E. Ball, "A bayesian heart: computer recognition and simulation of emotion," in *Emotions in humans and artifacts*, P. S. Trappl R, Petta P, Ed. Cambridge, The MIT Press, 2003.

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

[12] A. Morris, "Automatic segmentation," IRCAM, Tech. Rep., 2006.

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

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

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

[16] H. Murthy, K. M. Murthy, and B. Yegnanarayana, "Formant extraction from phase using weighted group delay function," in *Electronics Letters*, vol. 25, no. 23. IEE, 1989, pp. 1609–1611.

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

[18] J. van Santen, L. Black, G. Cohen, A. Kain, E. Klabbers, T. Mishra, J. de Villiers, and X. Niu, "Applications of computer generated expressive speech for communication disorders," in *EUROSPEECH*, 2003.

[19] K. Murphy, "The bayes net toolbox for matlab," in *Computing Science and Statistics*, vol. 33, 2001.