Expresso: Transformation of Expressivity in Speech

Grégory Beller

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

Abstract

This article presents a complete system, Expresso, which can apply to a synthesized or recorded sentence a chosen expression with a chosen degree of intensity, through high quality transformation of the speech signal. The transformation parameters depend on the context and are generated by a Bayesian network, after a learning phase using a corpus of expressive speech examples. This article presents the general system, the recorded expressive corpus, a new hierarchical prosodic model including the degree of articulation and the voice quality, the bayesian network used to generate parameters of transformation, the speech processing algorithms and an evaluation. This system is operational for sentences in French. It has been created to answer the artistic needs of music composers, of dubbing studios and of video production studios.

1. Introduction

The analysis-transformation-synthesis of emotions in speech allows its use in numerous artistic, scientific, therapeutic, commercial and legal applications, according to the associations HUMAINE[1] and ISRE[2]. Film directors, composers, dubbing studios and video game producers are interested in the multiple possibilities that such a system can supply, in terms of recognizing, transforming and feigning emotional reactions in the voice, as demonstrated in the VIVOS[3] project. Because, indeed, if the emotional state is idiosyncratic, that is unique to every individual, the associated reaction is observable by everyone. What offers the possibility of transforming a reaction associated with an emotion to change the perception of this emotion [1].

If there is a neutral internal emotional or psychological state, in which neither emotion, nor humor, nor feeling, nor attitude exist, then it is at a level of zero expressivity. The speaker gives no information about his internal state. This absence of information is referred to as *neutral* [2]. A sentence pronounced with neutral expression thus gives no information on the internal state of the speaker. A sentence of neutral expression has the advantage for expressivity transformation as it contains other levels of available information in the speech, without the influence of an existing expression. So, the comparison of a neutral utterance and of an expressive utterance, expressed by the same speaker with the same speaking style, the same modality and the same semantic message, allows isolating the influence of the expressivity on the speech prosody [3]. The five dimensions of the prosody are comprised of the following characteristics: intonation (fundamental frequency, pitch, melody), intensity (energy, volume), speech rate (speed of delivery), degree of articulation (pronunciation, configurations of the vocal tract, formant dynamics), phonation (glottis source, voice quality (pressed, normal, breathy voice), vibratory mode (fry, normal, falsetto), voicing frequency...)

The transformation of the prosody enables the modification of paralinguistic levels of the speech. The Expresso system of transformation, an overview of which was given in the first part of this article, thus aims at modifying the prosody of a spoken or synthesized neutral utterance, to confer on it a desired expression, while leaving the other information levels intact. For that purpose, a corpus of expressive speech containing pairs of neutral, expressive sentences was recorded and is presented after a general overview of the system. At the end is proposed an evaluation of the Expresso system on the basis of an expressivity recognition test involving transformed and spoken utterances.

2. General presentation

The Expresso system aims at conferring a desired expression with a chosen expressive intensity on one given neutral sentence, called the *source*. Figure 1 illustrates the various treatments that Expresso operates. When a neutral source sentence is presented to the system, it is first analyzed for its context. Then, this context is duplicated to supply a similar target expressive context, in which only the desired expression and the expressive intensity are modified. These symbolic contexts are used by a generative model to deduce acoustic models corresponding to the neutral case and to the expressive case. This generative model includes a Bayesian network of discreet (contextual S) or continuous (acoustic A) variables. These variables were observed during a preliminary learning phase on one expressive speech corpus. Two successive inference phases thus allow defining two acoustic targets, which are then compared to produce transformation functions dependent only on the change of the expression (known as *neutralization*). The resultant functions are then applied to the prosodic model of the source neutral sentence, with the aim of defining one expressive target which corresponds to it. The difference between the expressive target and the neutral source enables the transformation parameters to be defined. Finally, these dynamic parameters control algorithms for speech signal transformation which, when applied to the neutral source, confer on the signal the desired expression.

3. Expressive corpus

The expressive corpus, *IrcamCorpusExpressivity* [4], contains recordings of the voices of four dubbing actors, two men and two women approximately forty years of age. Expressivities

[1] HUMAINE:http://emotion-research.net/

[2] ISRE:http://isre.org/index.php

[3] VIVOS:http://www.vivos.fr

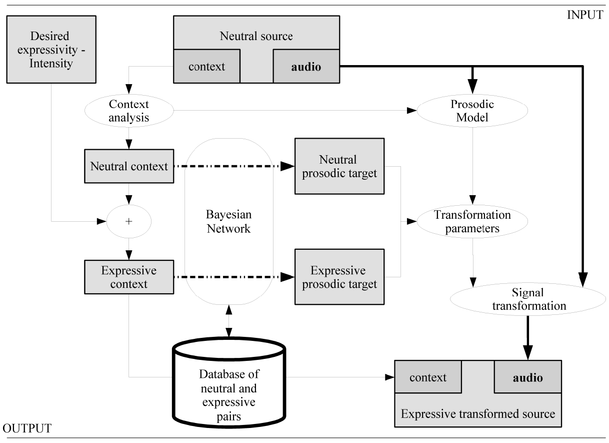


Figure 1: Overview of the system.

have been uttered with introversion and extraversion and with several intensity levels. The expressivities recorded were: *neutral* (reference state), *introverted anger* (contained or cold anger), *extroverted anger* (explosive or warm anger), *introverted enjoyment* (sweet or maternal enjoyment), *extroverted enjoyment* (explosive enjoyment), *introverted fear* (contained fear), *extroverted fear* (explosive or alarming fear), *introverted sadness* (contained sadness), *extroverted sadness* (explosive or tearful sadness), discretion, disgust, confusion, positive surprise (for the speaker), negative surprise (for the speaker) and excitement. The data collected consisted of an audio file for every sentence and a corresponding XML file, containing the metadata relative to the expressivity (category, activation and intensity), to the declaimed text, and to the identity of the speaker (age, sex, name). Some stages of manual labeling completed the data (phonetic segmentation, annotated prominence level and paralinguistic labeling). At the end, more than 500 utterances were recorded for each actor, forming a corpus of total duration about 12 hours of expressive speech in French.

4. Prosodic model

The utterances of the expressive corpus, as well as the source utterance to be transformed, are prosodically modeled in the same way. Prosodic units defined by the analysis of the context allow the stylization of the acoustic data relative to the five prosodic dimensions.

4.1. Context-dependent model

A first operator, called *context analysis* (see figure 1), analyzes the text, the metadata and the phonetic segmentation, and supplies a set of informed *units*. A unit is defined by its context which is a combination of symbolic discrete variable states, and by two temporal boundaries connecting it to a part of the audio signal. The definition of a prosodic model strongly depends on the chosen units (utterance, breath group - phrase, word, melism, syllable). These units possess different durations and can overlap. The units used are listed here in an hierarchical order of inclusion, from smallest to largest: phones, syllables, breath group and sentence. This stage also connects these units internally, due to tree construction representing the prosodic hierarchy. Units can thus inherit contextual data generated from nearby units or parents. Table 1 lists the variables that form together the context of a unit, and their cardinalities (number of states which these variables can take).

Variables	units	card.	description
$Sspeaker$	sentence	4	Name of the actor
$Ssex$	sentence	2	Gender of the speaker
$Smodality$	sentence	4	Modality of a sentence
$Sprominence$	syllable	5	Prominence of a syllable
$Sphoneme$	phone	38	Phoneme of the phone
$Stext$	sentence		orthographic text
$Sexpressivity$	sentence	15	Expression
$Sdegree$	sentence	6	Degree of expressive intensity

Table 1: Names, units, cardinalities and descriptions of the symbolic variables S

4.2. Acoustic analysis

The YIN algorithm [5] is used to estimate $f0$ only during phonation (voiced part of the speech signal). The measure of the intensity, as a prosodic dimension, is rather different from the measure of instantaneous energy.we use a perceptive measure of the intensity, called *loudness*. Speech rate is defined by the interpolated inversion of the durations of the syllables [6] which allows the definition of local speech rate which shows accelerations and decelerations within an utterance (see Figure 2). The degree of articulation is influenced by the phonetic content, the speech rate and the spectral dynamics that corresponds to the rate of change of the vocal tract configuration. The proposed measure of the degree of articulation results from the joint statistical observation of changes between the area of the vocalic triangle and the speech rate according to the intensity of the expressivity [7]. Voice quality is measurable due to the estimation of the relaxation coefficient, Rd [8].

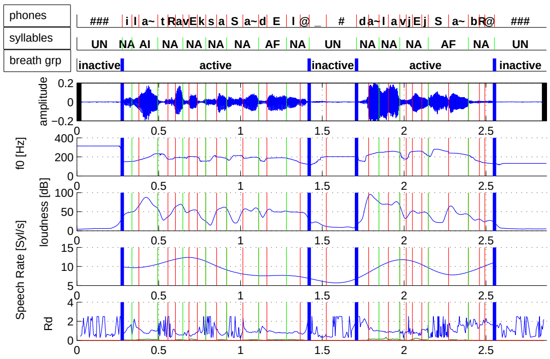


Figure 2: Example of raw prosodic data.

4.3. Stylization

The stylization process produces a simplified shape relative to the trajectory of a continuous value, which is supposed to protect its functional and audible phenomena. The proposed stylization algorithm, ProsoArchy [1], takes advantage of various pre-existing stylization processes such as Prosogram [9], MoMel [10], Fujisaki's model [11], B-splines models [12], and hierarchical models [13]. ProsoArchy is an additive hierarchical model of log-quadratic curves. The algorithm uses phonetic segmentation to define prosodic units: sentence, breath group and syllable. After applying a logarithmical function to the raw curve of the sentence, it estimates a quadratic model. It subtracts then from the initial raw curve, the estimated model. The byproduct of this substraction is also quadratically modeled but takes into account the breath group segmentation. In a recursive way, the same stylization process occurs for child syllable units. Every unit of the sentence is thus modeled by one quadratic model, regardless of its duration. Figure 3 shows an example.

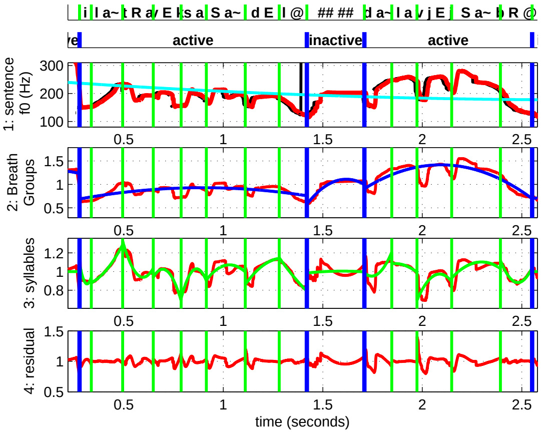


Figure 3: Example of a stylization using ProsoArchy. Sentence model (blue sky), breath group model (dark blue), syllable model (green) and residual (red).

5. Model learning

When all the sentences of the corpus were symbolically and acoustically analyzed, the database, managed by the platform *IrcamCorpusTools* [14], collects numerous units informed by their contexts and their modelled prosody. These units thus appear as sets of both discrete variables (symbolic S) and continuous variables (acoustics A). The purpose of the generative model is to supply a possible acoustic realization for a given context. So, the problem means deducing sets of acoustic values A corresponding to a given context: S = C_i, i.e. to estimate $P(A|S = C_i)$. To take advantage of the expert approach which allows the flexibility and the introduction of arbitrary rules, and the data-driven approach which allows a greater complexity [15], Bayes paradigm is then applied [1]. After a learning phase, an initial rule based model is partially transformed into a data-driven model, according to the number of available observed examples. The structure of our graphic model is illustrated by Figure 4. Rectangles represent the discrete variables indicating the context. Circles represent the continuous variables that are vectors of stylization parameters. The rounded and tinted rectangles represent the prosodic dimensions that involve the acoustic variables. Arrows represent the dependencies between variables. The heterogeneous of the nature of these variables enables the model to be context-dependent.

6. Transformation of the speech signal

Once the model has been learnt, an inference phase predicts a prosodic behavior, according to a set of contexts. Following the example of the utterances of the corpus, the neutral source sentence, presented to the system, is represented by a temporal and hierarchical sequence of contexts C and by a set of acoustic values describing its prosody. From the sequence of contexts extracted from the source, a first prosodic behavior is inferred, representing the way an actor in the corpus would say the same sentence in a neutral way. Then, this context is slightly modified, so as to include the expression and the intensity desired. A second prosodic behavior is then inferred taking into account the modified context's sequence. It aims to represent the way the same actor in the corpus would have uttered the same sentence with the new expression and intensity. Two inferences thus supply two possible acoustic realizations of the same sentence pronounced in a neutral and in an expressive way. The comparison of these two sets of acoustic values supplies transformation functions. The application of these functions to the prosodic behavior of the source sentence supplies target values. Comparison of the source values and of the target values then allows the definition of transformation parameters which drive the speech processing algorithms: Transposition (intonation), time-stretching (speech rate), gain (loudness), non-linear stretching of the spectral envelope (articulation degree and voice quality) [16].

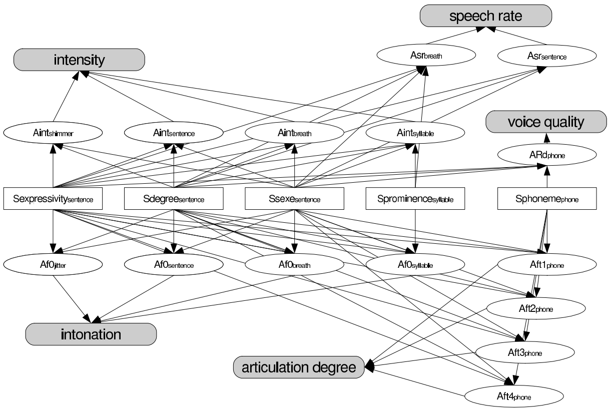


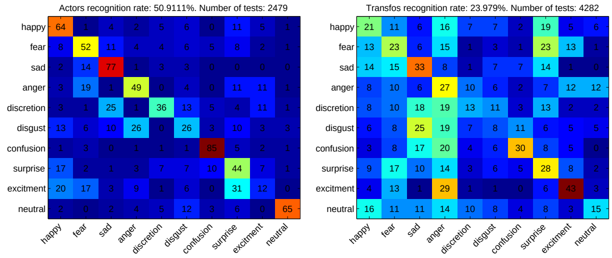
Figure 4: Bayesian network used as generative model.

7. Evaluation

A perceptive test was developed on to evaluate the Expresso system. It is a recognition test for the expressivity, available on the internet[4]. The purpose of this experiment is to compare the recognition rates of original and transformed stimuli. A semantically neutral sentence was extracted from the corpus: “He cannot see me if I switch off the lamp.” (“Il ne pourra pas me voir si j’éteins la lampe.”). Fourteen expressive versions of this sentence, performed by each of the two actors are a part of the test, as controls. To this set, the products of expressive transformations were added. Two models, trained on two male actors, were evaluated: *ModelCombe* and *ModelRoullier*. The neutral sentences of these two actors were transformed in all the expressivities, using both models. The test presented in random order 90 stimuli, of which one third were control sentences which had not been transformed. The participant task was to classify these stimuli using the same expressive categories as the actors. Every stimulus had been listened to at least 75 times, to produce statistically significant results. Overall, the recognition rates are rather weak, despite being above chance. Results are presented

[4] `http://perspection.ircam.fr/beller/expresso_quiz/`

after grouping extroverted and introverted versions (randomness rate 10%). Figure 5 presents the confusion matrices for the control stimuli and the transformed stimuli with grouping of expressive categories. Recognition rates are readable along the diagonal of each confusion matrices.



Actors recognition rate: 50.9111%. Number of tests: 2479

	happy	fear	sad	anger	discretion	disgust	confusion	surprise	excitment	neutral
happy	64	1	4	2	5	6	0	11	5	1
fear	8	52	11	4	4	6	5	8	2	1
sad	2	14	77	1	3	3	0	0	0	0
anger	3	19	1	49	0	4	0	11	11	1
discretion	3	1	25	1	36	13	5	4	11	1
disgust	13	6	10	26	0	26	3	10	3	3
confusion	1	3	0	1	1	1	85	5	2	1
surprise	17	2	1	3	7	7	10	44	7	1
excitment	20	17	3	9	1	6	0	31	12	0
neutral	2	0	2	4	5	12	3	6	0	65

Transfos recognition rate: 23.979%. Number of tests: 4282

	happy	fear	sad	anger	discretion	disgust	confusion	surprise	excitment	neutral
happy	21	11	6	16	7	7	2	19	5	6
fear	13	23	6	15	1	3	1	23	13	1
sad	14	15	33	8	1	7	7	14	1	0
anger	8	10	6	27	10	6	2	7	12	12
discretion	8	10	18	19	13	11	3	13	2	2
disgust	6	8	25	19	7	8	11	6	5	5
confusion	3	8	17	20	4	6	30	8	5	0
surprise	9	17	10	14	3	6	5	28	8	2
excitment	4	13	1	29	1	1	0	6	43	3
neutral	16	11	11	14	10	8	4	8	3	15

Figure 5: Confusion matrix for the actor's stimuli (left) and for the transformed stimuli (right). X-axis: Target expression. Y-axis: Perceived expression

Considering the weakness of these recognition rates, it is necessary to look at the causes of discrepancies. we look at the further the recognition results of both used models, *ModelCombe* and *ModelRoullier*. On average, the model *ModelCombe* seems to give better results than the model *ModelRoullier* for the extroverted enjoyment, the extroverted sadness and the excitement while the others do not really possess the capacity to be perceived as the desired expression. The *ModelRoullier*, is successful for the extroverted sadness, the extroverted anger and the confusion. The first conclusion of this comparison is that it seems manifest that the models possess performances dependent on the expression. It seems while a hybrid model, gathering parts of both models according to their recognition rates by expression could supply better results. Unfortunately, the difference between neutral recognition rates shows that this evaluation suffers from a lack of stimuli. Indeed, neutral transformation does not change anything in the utterance, resulting in two neutral utterances with different recognition rates. That shows that resulting recognition rates are under the influence of the chosen utterance. However, a comparison of the performances of the two learnt models leads us to the choice of an hybrid model, taking parts of each depending on the expression.

8. Conclusion

In this paper, we presented Expresso, a system that transforms speaker expressivity. Expressivity is one level of accessible information through speech such as semantic sense, speaker identity, speaking style and modality. This level of information manifests through non verbal sounds, restructurings and prosody. Through detailed study of an expressive French speech corpus, we established a new context-dependent hierarchical prosodic model. This model uses a Bayesian network and generates from the text the prosodic patterns corresponding to expressivities. By the comparison of neutral utterances and expressive utterances of the corpus, it is then possible to estimate the impact that a change in expression can have on the prosody of an utterance. The influence of change on expression can then be applied to a new sentence, upon entry into the system, to confer a desired expression, through transformation of the speech signal. Two inference phases provide transformation functions that are applied to the sentence to be transformed by a phase vocoder. The pitch, intensity, speech rate, degree of articulation and voice quality are five prosodic dimensions transformed automatically by the Expresso system. The last one was estimated on the basis of a perception study that produced encouraging results. Future research tracks comprise the modification of the other phonation attributes like the voicing frequency, the breathiness, and the vibratory mode as well as synthesis of non verbal sounds.

9. References

[1] G. Beller, "Analyse et Modèle génératif de l'expressivié: Application à la parole et à l'interprétation musicale," Ph.D. dissertation, Université Paris XI, IRCAM, June 2009.

[2] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1145 – 1154, July 2006.

[3] G. Beller, "Transformation of expressivity in speech," *Linguistic Insights*, vol. 97, pp. 259–284, 2009.

[4] G. Beller, C. Veaux, and X. Rodet, "Ircamcorpus-expressivity: Nonverbal words and restructurings," in *LREC2008 - workshop on emotions*, 2008.

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

[6] G. Beller, D. Schwarz, T. Hueber, and X. Rodet, "Speech rates in french expressive speech," in *Speech Prosody 2006*, SproSig. Dresden: ISCA, 2006, pp. 672–675.

[7] G. Beller, N. Obin, and X. Rodet, "Articulation degree as a prosodic dimension of expressive speech," in *Speech Prosody 2008*, Campinas, 2008, pp. 681–684.

[8] G. Fant, "The voice source in connected speech,," *Speech Communication*, vol. 22, pp. 125–139, 1997.

[9] P. Mertens, "The prosogram : Semi-automatic transcription of prosody based on a tonal perception model," in *Speech Prosody*, 2004.

[10] D. Hirst, A. D. Cristo, and R. Espesser, *Prosody : Theory and Experiment*. Kluwer Academic, M. Horne (ed), 2000, ch. Levels of representation and levels of analysis for intonation, pp. 51–87.

[11] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing. acoustical analysis and physiological interpretations," Dept. for Speech, Music and Hearing, Tech. Rep., 1981.

[12] D. Lolive, N. Barbot, and O. Boeffard, "Modélisation b-spline de contours mélodiques avec estimation du nombre de paramètres libres par un critère mdl," in *JEP*, 2006.

[13] Y. Morlec, "Génération multiparamétrique de la prosodie du français par apprentissage automatique." Ph.D. dissertation, INPG, 1997.

[14] G. Beller, C. Veaux, G. Degottex, N. Obin, P. Lanchantin, and X. Rodet, "Ircam corpus tools: Système de gestion de corpus de parole," *TAL*, 2009.

[15] 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.

[16] N. Bogaards, A. Roebel, and X. Rodet, "Sound analysis and processing with audiosculpt 2," in *ICMC*, Miami, USA, Novembre 2004.