GRÉGORY BELLER

Transformation of Expressivity in Speech

1. Introduction

The need to express and understand emotions is a fundamental part of human communication. This level of information does not always inform us of the emotional state the person is in. Even if the actors are driven by their emotions, they can also be simulated. From reviewing different theories on the control of emotional reactions and of several emotional data collecting methods, it appears necessary to separate the inner emotional state from the corresponding outer perceived expression, by the introduction of the term 'expressivity'.

With regard to speech, expressivity is presented here as a level of information in the spoken message which gathers the expression of emotions, simulated or not, attitudes and moods. It is different from speaker identity, speaking style, modality and prominence, which justifies the interest in defining a state reference called 'neutral'. Expressivity is accessible beyond words because it is also conveyed by paralinguistic indices such as restructurings, non-verbal sounds, and prosody.

Since it is an essential level of information, human-machine interfaces try to manage it by recognising and synthesising expressive speech; indeed, current speech synthesis methods provide natural and intelligible speech. Art directors, film studios and video game producers are now interested in the many possibilities of a system which offers analysis, synthesis, and transformation of the expressivity of the voice, as demonstrated by the VIVOS project.

Statistical models of emotional prosody have been used by voice conversion systems (Hsia *et al.* 2007) as well as by concatenative (Bulut *et al.* 2007) and HMM-based (Yamagishi *et al.* 2005) speech synthesizers. Our approach consists of splitting the problem into two parts: firstly, we establish a neutral synthesis from a

neutral corpus of excellent quality; then we transform the expressivity of this synthesis. This does not prevent us from using expressive targets in the stage of synthesis, and it also allows us to transform recorded and non-synthesised sentences. The second part of this chapter focuses on transformation of the expressivity of an utterance.

2. Emotion and expressivity

A growing community of researchers study emotional data to infer or confront theories on emotions. A commonly recognised difficulty expressed in these studies is evaluating the degree of control/spontaneity over the emotional data. The acted emotional data from long studies in laboratories is now widely criticised by those who wish to observe the demonstrations of the felt emotional states.

However, it must be acknowledged that several techniques employed by actors, for example, self-induction of an emotional state, can lead to real emotional expressions. Furthermore, this difficulty shows that a clear boundary has to be drawn between inner emotional state studies and outer expression observations. That is why we provide an attempt at defining expressivity.

2.1. Theoretical background on emotions

Different topics related to emotions have motivated theories. Many theories have been designed on the function, the role and the mechanism of emotions, though few of them deal with the control of emotions and related emotional reactions. Here we try to provide a chronological view of such theories.

From Aristotle to the Middle Ages, emotions were considered as reflecting the animal part of the human kind. To get to perfection, one has to control them (Saint Thomas d'Aquin). At rebirth, Descartes describes emotions as bodily responses to external stimuli (Descartes 1664). They are thus uncontrollable since they are part of rational determinism. LeBrun draws a dictionary of composed and basic facial emotional responses. They both related the existence of an emotion organic centre, called the pineal gland.

In 1872, Charles Darwin characterises emotions as vestiges of patterns of actions for the survey, designed by evolutionary development (Darwin 1872). This description of systematic emotional responses, which are considered as universals (for mammals), does not allow for any control, as they are biologically defined. Shortly after, James (1884) agrees with Darwin, but changes the causal relations. Emotions are appraisals of emotional reactions which are direct responses to external stimuli.

During the past century, cognitivists aim at describing this appraisal (Frijda 1986). Scherer especially describes two parallel processes that occur in vocal expression of emotions (Scherer 2006). The 'push' effect underlies psycho-physiological activations displayed by uncontrollable paralinguistic signs. The 'pull' effect is shown by the use of socio-cultural codes. Each of these two effects affects the prosody of emotional speech. The 'push-pull' distinction aims to distinguish between the controlled part (pull) and the uncontrolled one (push). This distinction can be seen in other theories as making a difference between the spontaneous part and the symbolic/acted part of the expression of an emotional state (Buck 1985). The argument for totally, culturally designed emotional response is central in the social-constructivist perspective that claims that emotional responses are not biologically but culturally motivated (Averill 1980).

Nowadays neuroscientists aiming at localizing the pineal gland refresh the debate. On the one hand, emotional stimuli seem to be directly managed by the limbic system and then interpreted by the cortex (LeDoux 2005). On the other hand, somatic effects related to emotions can be activated consciously (Changeux 1983), that is to say, one can self-induce emotional reactions without the need of external stimuli. This debate about our ability to control emotional response is at the heart of methods for collecting emotional data.

2.2. Collecting emotional data

Emotional data is, by nature, rare and heterogeneous. Nevertheless, some experimental protocols allow their acquisition (Douglas-Cowie *et al.* 2003). These methods are generally divided into three adopted classes:

- Naturalistic data:
  - Free context: All emotions can occur. The difficulty remains in the lack of knowledge about the context and consequently, in the variability of recording conditions (Chung 2000).
  - Constrained context: A service is dedicated to specific scenarios where a limited number of emotional states can occur. As the emotional state is not the goal of one who is under it, this method is part of the category of indirect measurements, using the perception test terminology (indirect method). It provides naturalistic emotional data with a stable recording process (Vidrascu/Devillers 2005).
- Induced data: An emotional situation is artificially created in a predefined context with good recording conditions. Again, it involves indirect methods, the subject of which is not aware of what data is really observed (Aubergé *et al.* 2004).
- Acted data: Actors simulate desired emotions while speaking a chosen text and are recorded in ideal conditions (direct method).

A review on techniques employed by actors reveals that some can generate naturalistic data, e.g self-induction methods (Beller *et al.* 2008c), though these techniques have not been involved in a scientific study yet. All recorded emotional data by use of direct or indirect methods handle a compromise between their spontaneity and their controllability (or the knowledge about their context). The case of acted versus spontaneous emotional expression is even harder to define taking into account that the person in an emotional state cannot measure the degree of control he has on his own expression. From a perceptual point of view, it is also difficult when regarding an emotional expression, to infer whether the speaker is in the corresponding emotional state. Nevertheless, an emotional expression simulated or not, often contains enough information for people to agree on the inner emotional state, i.e. on a supposition about the inner emotional state. Even if we do not know that the emotion is true or acted (felt or not), we can observe the same emotional expression and then reproduce it in order to give the information of the corresponding inner motional state. Thus, it seems that a clear boundary needs to be defined between an emotional state and its expression. To distinguish

the emotional (inner) state, i.e., what one can feel, from the (outer) expression of this state, i.e., what people can perceive, we specify this last phenomenon by the term 'expressivity'.

2.3. A definition of expressivity

Expressivity is a level of information in communication. This level groups together the external demonstrations, controlled or not, which can be attributed to uncontrolled internal states. These internal states include the emotions, feelings, attitudes, moods, and psychological states which make up the actor's performance style. Expressivity is thus a part of the communication system, defined as a demonstrator of an internal state which is inaccessible, by definition, in others. This internal state can be present or not, and its demonstration can be thus feigned or not. In every case, the expressivity refers only to the internal states which we can suppose are uncontrollable.

As for any common perceptual category, a significant number of people should agree on the term used to define expressivity. Since the perception of expressivity is dependent on the context in which it occurs, and since the way one perceives it is influenced by its proper experienced inner emotional states (neural perception-action theory (LeDoux 2005)), a large number of items compose the panel of expressivity. Actually, the literature reveals a lot of related terms that can be said to be components of a heterogeneous field with ambiguous and badly determined words. Even if standardisations have been proposed by the observation of the universal emotional responses (Ekman 1999), perceptive tests (Devillers *et al.* 2003) still show that people prefer to use more than one of these words for labeling emotional data. Furthermore, one can consider anger as a category, yet others will differentiate cold anger from hot anger since involved taxonomy is strongly application-driven. Thus, it is hard to define explicitly a closed set of linguistic categories that compose expressivity. That is why we attempt to define this information level by using 'the opposite definition paradigm'. In other words, we report here information levels that cannot be considered as part of expressivity (see Figure 1).

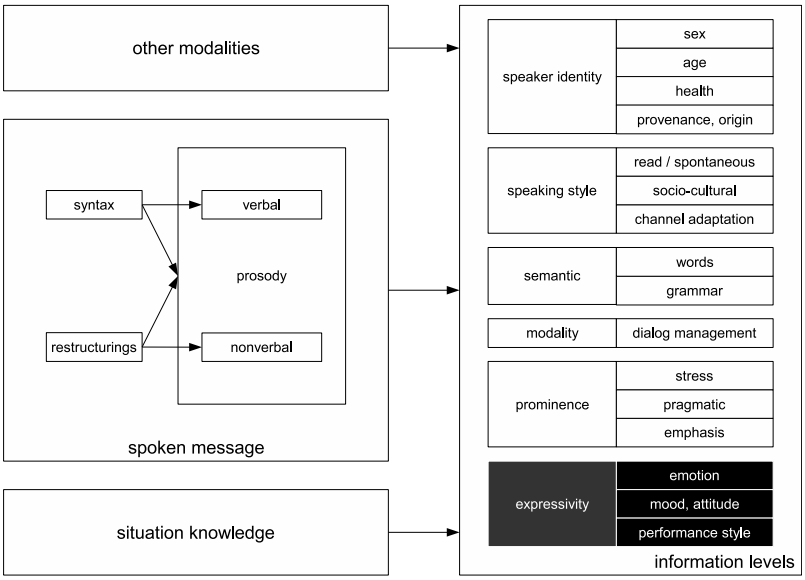

Figure 1. Presentation of various levels of information involved in verbal communication (right part). Expressivity refers to a hypothetical internal state. The spoken message, which conveys this information, can be decomposed into several supports (left part).

2.3.1. Speaker identity

First of all, in any act of vocal communication there is a speaker involved. The various information relative to the identity of this speaker is then implicit in communication. In the case of spoken communication, the identity of a speaker is conveyed by its voice. The sex, age, health, origin (foreign status, regional accent, social standing, level of language) all contain a wealth of information relative to the personality of the speaker made accessible to the others by the voice.

2.3.2. Speaking style

Numerous contextual factors can affect the manner in which an individual speaks. First of all, physical constraints can bring them to

speak louder or to repeat words they have said, for example, in the case of a noisy environment (Garnier 2007). Social constraints also come into play. The professional constraints especially bring complex speaking styles as is the case in political speech or in a teaching environment. Finally, important differences appear between spontaneous speech, speech read from a text and speech recited 'off by heart'.

2.3.3. Semantic message

A speaker taking into account external constraints communicates a semantic message. Interestingly, expressivity can only be perceived by the way words are pronounced and flourished by non-verbal sounds. This means that semantically non-expressive texts (said neutrally) can be spoken in a way that people can perceive an intended expressivity. By non-expressive texts, we rely on sentences that can be pronounced in any expressivity without losing any sense. To a larger extent, musical performance can carry expressivity without using any word. Furthermore, universal emotions can be perceived by foreign listeners that do not understand the language of the expressive speaker (Burkhardt *et al.* 2006). Expressivity can thus be explicitly recognized by means of the semantic message, but is not dependent on it.

2.3.4. Modality

Modality is a special part of the vocal communication due to its important role in dialog management. Modality is generally defined by three common categories: question, assertion and exclamation. In normal dialogical situations, modality governs the tour of word. At some point of a discussion, it reflects the position of the speaker on the issue being discussed. He can then adopt the opposite position, generally called case irony, scepticism and doubt. Irony and doubt are commonly classified as an instance of expressivity. But, unusually, we leave it to the reader to classify them as a part of the modality group. This is because of the strong relationship between the linguistic message and the way it is uttered. This does not prevent the message from being accompanied by any expressivity such as confusion or embarrassment.

2.3.5. Prominence

Prominence is the perceptive result of controlled acoustic contrasts (culmination, distinction, demarcation) performing several functions. First of all, it is evident in the accentuation of certain syllables which can be defined by linguistic rules (Lacheret-Dujour/Beaugendre 1999) and which is thus language-dependent. Prominence sometimes plays a role in the disambiguation of the sense, as in pragmatic accent. Finally, prominence also serves to emphasise certain elements (accent of focus, accent of emphasis, and accent of instigation). It seems that prominence plays a particular role in verbal communication, because its realisation requires a more important part of control during its production. That is why prominence distinction is very important for expressivity analysis (see Section 4.1.2). In a certain way, the 'push-pull' theory can be scaled down to a local and dynamic version: the non prominent, obvious syllables, which are brought about by the 'push' effect vs. the prominent, more elaborated syllables, the realisations of which are marked by the 'pull' effect.

Supprimé : nce

2.4. Neutral as a reference

In this chapter, expressivity is explicitly presented as a level of information among others, within a vocal message. An individual (speaker identity) pronounces (modality and prominence) a linguistic message (semantic message), in a certain physical environment and under certain social constraints (speaking style) and according to his internal state, which he decides to express or suppress (expressivity). To observe the variations due to this last level of information, that is, due to expressivity, it is necessary to marginalise the variations due to the other levels of information. If it has not been demonstrated yet that there is a neutral internal emotional or psychological state, in which neither emotion, nor mood, nor feeling, nor attitude exists, the existence of a level of zero expressivity, in which the speaker gives no information about his internal state, seems widely accepted by those who address this absence of information by the case: neutral. This is why researchers and engineers usually compare two versions of the same sentence, pronounced by the same person and in the same

conditions, and containing two different expressivities, in order to analyse, measure and compare them (Tao *et al.*, 2006). This empirical and widely used process seems evident here because it neutralizes the variations due to the other information levels in order to enlighten the variations relative to the change of expressivity. The measured variations can be reused for the other speakers pronouncing other texts and so lead towards general, generative models of expressivity. The realised corpus is based on this process (see Section 4).

3. Expressivity in speech

The 'double coding of speech' framework (Fónagy 1983) differentiates the linguistic channel from the paralinguistic channel. The linguistic channel contains the semantic information level of words structured by syntax. The paralinguistic channel handles nonverbal sounds, restructuring and prosody that describe not only the way words are pronounced but also the way nonverbal sounds are uttered. From a perceptual point of view, the listener perceives an acoustic signal where information is expressed, through syntax and restructuring, in a structured sequence of verbal and non-verbal sounds, all modulated by prosody. Previous information levels and their relationships are retrieved by the listener decoding the spoken message together with other modalities such as visual, and with internal and situational knowledge.

3.1. Non-verbal sounds

Non-verbal sounds are a common phenomenon in speech. This category groups all sounds produced during speech that do not handle any linguistic function. For instance, fillers (laughter, scream, etc.), and respiration (breaths, pauses, etc.) are important information for expressivity but not for semantic sense. In fact, nonverbal sounds act as a semantic mark-up of expressivity. One can infer sadness only from a cry, and fear only from a scream. The importance of nonverbal

sounds can be emphasized by the fact that they are often written in a text, in a phonetic-spelling way. For instance, *ha ha ha* is often used to write laughter in comic strips. Despite a formal textual representation, non-verbal sounds appear to be of first importance for expressivity. That is why expressive text-to-speech systems start to synthesise nonverbal sounds such as laughter (Beller Forth.).

3.2. Restructuring

Another part of the paralinguistic channel relies on possible restructuring cases that are commonly considered as disfluencies in neutral speech (Piu/Bove 2007). Repetitions, resettings, and other a-syntactical restructuring of speech are informative and appear a lot in expressive speech (Beller *et al.* 2008c). An efficient and common language needs to be defined in order to annotate these expressive non-verbal phenomena (both restructuring and non-verbal sounds) and to compare them among corpora. Several transcriptions have been proposed (Aubergé *et al.* 2006). They almost all consist in extensions of the SSML[1] (Eide *et al.* 2004) (Blankinship/Beckwith 2001).

3.3. Prosody

The verbal and non-verbal sounds of speech are ordered in a sequential way. At the same time, the acoustic realization of these sound elements is modulated by prosody. Prosody deals with phonological features, known as suprasegmental features, which apply to groups larger than a single segment (e.g. the phoneme), such as the syllable, the word, or the breath group. Five features are cited in literature as the five dimensions of prosody (Pfitzinger 2006):

- intonation: fundamental frequency, pitch, F0;
- loudness: intensity, energy;
- speech rate: speed of delivery;

1 SSML: Speech Synthesis Markup Language: <http://www.w3.org/TR/speech-synthesis>.

- articulation degree: pronunciation, vocal tract configuration, formant dynamics;
- phonation: glottal excitation, relaxation (pressed / normal / breathy), vibration mode, voicing.

4. Prosody analyses

The significant variation of these features, allotted to the change of expressivity, makes it possible to build models of actors' performance styles. The recording of the latter provides sound examples which are both symbolically and acoustically analysed. The machine then learns the existing relationships between the expressivities and the configurations from acoustic parameters such as intonation (for example), whilst taking into account contextual information. Lastly, these generative statistical models are able to propose parameters of transformation for new sentences from outside of the corpus, and more importantly from a different speaker. The use of these transformation parameters by speech processing algorithms makes it possible to confer a desired degree of expressivity to recorded or synthesised neutral speech. Therefore, we build an expressive corpus involving actors that utter expressivity. After presenting the corpus, we shall offer some prosodic analyses to prepare the design of a context-dependent generative model of the prosody of expressivity in speech.

4.1. IrcamCorpusExpressivity

IrcamCorpusExpressivity contains recordings of four French actors (two males and two females for a varying speaker identity) each with a duration of approximately one hour and a half (Beller *et al.* 2008c). They were all recorded in the same professional conditions and followed the same procedure. Ten neutral sentences extracted from a phonetically balanced French corpus (Combescure 1981) were uttered with expressivities: introvert and extrovert anger, introvert and

extrovert happiness, introvert and extrovert fear, introvert and extrovert sadness, and, positive and negative surprises, disgust, discretion, excitation, confusion, and, of course, neutral. Moreover, six repetitions per sentence occurred for basic emotions (written in italics), starting from the neutral version and afterwards with an increasing degree of the expressive level (power, intensity, activation). The corpus was composed of approximately 550 utterances per actor. Each utterance was phonetically hand-segmented (to discard semantic level) and prominence was hand-labelled (to discard prominence level). Some non-verbal sounds for each emotion were also recorded. All these data and their interrelationships were managed by a relational database management system (Beller *et al.* 2008b).

4.1.1. Intonation and loudness

The first analysis of the corpus focuses on intonation and loudness. Pitch has been estimated with the Yin algorithm (De Cheveigné/Kawahara 2002) and loudness with a perceptual measurement (Peeters 2004). Jitter and shimmer are related to the spectral centre of gravity of windowed pitch and loudness respectively. Pitch is modifiable, using dynamic transposition, just as loudness is modifiable, using dynamic gain, thanks to SuperVP, a phase vocoder technology (Bogaards *et al.* 2004). The known results in agreement with the literature set apart (a strong correlation between pitch and loudness), show that extrovert emotions often distinguish themselves by a pitch the average of which is situated an octave or two over that of the neutral case (see Figure 3). So, a simple estimation of the average does not allow distinguishing between extrovert happiness and extrovert anger. This leads us to investigate other prosodic dimensions such as speech rate.

4.1.2. Speech rate

Speech rate is often defined as the average number of syllables per second in a whole sentence (Chung 2000, Pereira/Watson1998). Because the most prominent syllables often have a longer duration, we prefer to define speech rate by the sequence of individual syllable durations. The speech rate curve is thus represented by an interpolation of the durations of syllables (see Figure 2). A

deceleration corresponds to a rising of the curve and an acceleration is represented by a falling of the curve. These movements can be simulated using SuperVP by locally time-stretching (compressing or dilating) the speech signal. Figure 2 shows that the final accent is more distinguishable in the speech rate curve than in the F0 curve. Indeed, it has been shown that speech rate analysis with prominence care performs a better discrimination between extrovert happiness and extrovert anger than pitch analysis only (Beller *et al.* 2006). The author emphasises here the need for prominence information level to analyse expressivity (see Section 2.3.5) since prominent parts of the speech are more controlled than others.

Another advantage of the dynamic speech rate estimation is the facilitation of the study of rhythm. In fact, Figure 2 shows clearly that the actor emphasises the sentence with an increase of the prominent syllable duration which gives a certain rhythm to his performance. Research is now focused on rhythm extraction so as to measure the metricity of the speech, that is to determine whether it possesses an inherent perceptual rhythmical structure or not.

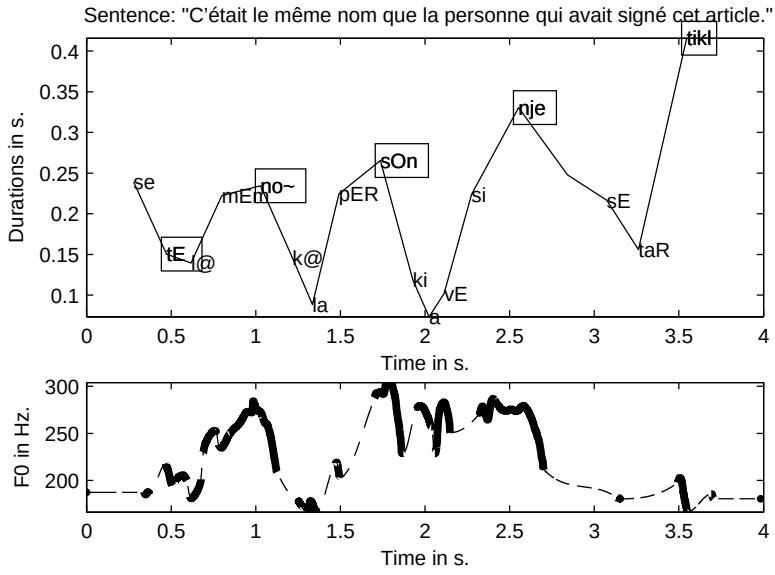

Figure 2. Duration of syllables and F0 curve of a French sentence uttered by a male actor with extrovert happiness: “C’était le même nom que la personne qui avait signé cet article”. Squared syllables are perceived as prominent.

4.1.3. Articulation degree

Articulation degree (Lindblöm 1983) originates from interactions between phonetic context, speech rate, and spectral dynamics (which correspond to speed changes in the configuration of the vocal tract). A study concerning the influence of segment duration on the vocalic triangle in neutral speech (Gendrot/Adda-Decker 2004) shows that formants aim towards a central vowel for segments of short duration, resulting in a diminishing of the vocal triangle area (neutral natural tendency). A major difference exists between neutral and other expressivities. Given a phonetic context, the articulation degree is not only dependent on the speech rate variable but also on expressivity.

Supprimé : s

Indeed, we conducted a statistical analysis on the influence of expressivity on the articulation degree (Beller *et al.* 2008a). Using phonetic segmentation, previous dynamic speech rate estimation and a proposed robust formant analyser, articulation degrees for different expressivities were computed. The neutral natural tendency is clearly not designed for expressivities like extrovert sadness, introvert fear, and surprises which show a reduction of the vocalic triangle despite slower speech (see Figure 3). Conversely, extrovert anger shows an expansion of the vocalic triangle area which exceeds that of the neutral natural tendency. Dynamic frequency warping manages articulation degree modifications by moving dynamically formant frequencies. Again, expressivity analysis and transformation require phonetic information derived from the semantic message information level (see Section 2.3.3).

4.1.4. Voice quality

Voice quality has been recently introduced as a prosodic dimension (Campbell/Mokhtari 2003), especially for expressive speech (Gobl/Chasaide 2003). This recent and exciting framework groups several analyses aimed at describing phonation. Several problems arise as this approach intrinsically involves an inverse problem – the separation of the effects of the glottis excitation signal and of the

vocal tract filter. Nevertheless, some features already exist and allow for an analysis, such as the four parameters Liljencrants-Fant (LF) model together with the Rd relaxation coefficient (Fant *et al.* 1985). Our first step in voice quality analysis thus involves an ARX-LF model estimation of the speech signal (Vincent *et al.* 2005). From the parameters of the LF models, the Rd coefficient is computed (Henrich *et al.* 2002) to estimate the relaxation degree of the speech (pressed/normal/breathy). To change this coefficient and to transform voice quality, a filter is designed by the ratio of the corresponding spectral envelope of the glottis models and applied by SuperVP. Voice quality analysis also investigates other phonation phenomena such as voicing (dynamic spectro-temporal voicing frequency), vibratory modes (falsetto, normal, fry), abnormalities (asymmetry of the vocal folds) and micro-temporal variations (jitter, shimmer, growl).

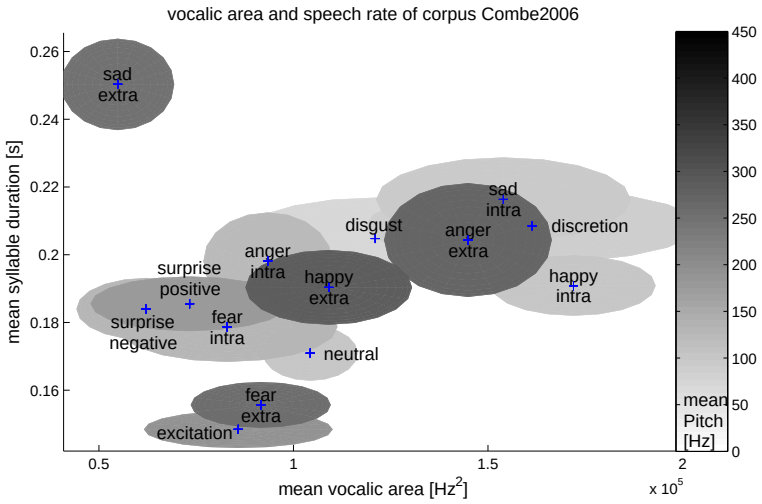

Figure 3. Expressivities performed by a male actor, represented according to their vocalic area (X axis), to their syllabic duration (Y axis) and to their mean pitch (Z axis -colours). Ellipsoids show mean values (centre coordinates) and variances (widths).

4.2. Paralinguistic paradox

Interestingly, voice quality parameters (open quotient) seem to be dependent on the pronounced vowel (probably due to production efficiency). Similarly, in order to analyse the articulation degree (to draw a vocalic triangle; for instance), we need at least the phonetic labels of the vowels. This level of annotation offers a categorisation in phonetic classes in which the spectra of the corresponding vowels can be compared. Thus the articulation degree can be estimated for all utterances independently of the phonetic context and then used to compare expressivity. Again, the study of speech rate has shown the importance of the prominence level of syllables.

Indeed, one major difficulty in the analysis of paralinguistic features is the acoustical influence of the verbal/segmental content. The realization of phonetic segment sequences driven by a text implies a co-articulation phenomenon and other physiological constraints that may affect pitch course and other prosodic features. The produced infra-segmental variations can falsely be attributed to prosody if the linguistic context is unknown. In order to observe supra-segmental features, segmental effects need to be removed. That is the paradox of paralinguistic analysis that requires linguistic information to be achieved. This is done by adding contextual information at each step: analysis, transformation and synthesis.

5. Prosody model and transformation

Our system is aimed at transforming a given neutral source utterance into a target utterance with the same sentence, from the same speaker, in the same speaking style, and with respect to original modality and prominence, but with a given expressivity *E* and a given expressive power *P*. First, linguistic information is derived from the text for each syllable of the neutral sentence (see Section 4). This provides a temporal sequence of contexts C^N_{src}. Then two corresponding acoustic descriptor sets are predicted, one using expressivity neutral A^N_{model} and the other using expressivity *E* A^E_{model}. Inferred acoustic parameter distributions are then compared so as to provide transformation

factors. Hence the problem becomes to infer acoustic data A^X_{model} corresponding to a given context $C^X_{model} = C_i$, i.e. to evaluate:
$P(A^X_{model} | C^X_{model} = C_i)$

5.1. Context-dependent model

The recordings divided in syllables can be classified according to symbolic information such as expressivity, prominence level, syllable type (V, CV, CVC...), or other linguistic annotation (relative to other information levels). A context is defined as a set of symbolic variables that can take different states in closed vocabularies. An example of such a context is given in Table 1. The syllable produced in context C_1 is a non-prominent syllable made of a consonant (CV) and of a vowel /a/, uttered by a male expressing extrovert anger with an expressive power of three (third repetition of the sentence in the database).

Variable	*Information level*	*State description*	*Example C_1*
Gender	speaker identity	female or male	"male"
expressivity	expressivity	see section 4	"extrovert anger"
Power	expressivity	degree or power of expressivity	"3"
prominence	prominence	non-secondary, primary prominent	"non-prominent"
type	semantic message	V, CV, CVC, VC	"CV"
vowel	semantic message		"/a/"

Table 1. Context definition: symbolic variables of different information levels and their state descriptions. C1 is an example

5.2. Model parameters: stylization of acoustic features

Once the context $C^X_{model} = C_i$ is defined, temporal evolution of the acoustic signals related to prosodic dimensions can be modelled: e.g. F0, loudness, local speech rate, articulation degree, relaxation coefficient. At this point there were many attempts at stylizing intonation contours. Different stylization models were proposed and

tested on the data. They could briefly be summed up using Legendre polynomial family, starting from mean value on the segment (order 0), by way of linear model (first order), to higher order stylization models like quadratic (second order) or cubic (third order) contours. This family is generally used since it confers the model a specific scalability on its fitting precision (Schwarz 2004), providing controllable complexity to the model. Every syllable of expressivity *X* is modelled by the stylization parameter values of its acoustic features A^{X}_{model}(*syllable*) together with its corresponding context C^{X}_{model}(*syllable*). These are the parameters of the proposed prosodic model of expressivity.

5.3. Model robustness: statistics

The power of a model is not only in its capacity to fit well and infer real data, but also in its simplicity (reduction of complexity) and its generality. Therefore, statistics are used to evaluate part of the model parameters that remain stable and robust in a significant amount of data. Again, contextual information is used to group syllables into classes. Statistics of model parameter values can be estimated in each context, leading to smoother, more robust, and more significant model parameter values.

5.4. Model generalness: marginalized decision tree

A new sentence to transform can present a context which was not observed during the recording phase. In that case, our generative model has the ability to propose a solution and supply parameters for the transformation. Several solutions exist depending on the knowledge about the context relationships. Firstly, the ability to compare contexts (needing a definition of a symbolic distance) can lead to a choice of the closest context to infer model parameter values or to interpolate values of the neighbouring contexts. Secondly, unobserved cases can be filled by model parameter distributions that are given a priori, using a Bayesian paradigm (Beller 2007). Finally, and as presented in more detail here, knowledge about context can be

defined by hierarchical relationships between contextual variables. We view the order of the presentation of contextual variables (see Section 1) as a hierarchical order (expressivity is more important than prominence which is more important than syllable type, and so on). Figure 4 represents this hierarchy by means of a decision tree. Each variable can take as many states as labels in its dictionary. Each node groups leaf syllables the context of which is the same as described by the path of the node.

Let us suppose that the new sentence to transform presents a context C_1 that was unobserved during the learning phase. For instance, it may happen because none of the syllables of the database contains the desired syllable type ("CV"). That part of this too specific request can be eliminated by marginalising context with regard to the variable type. This modified request is then applied to find syllables from the database which correspond less to our initial contextual request C_1, but which correspond to the most important part of that request (to the sense of the predefined hierarchy). Successive marginalisation of the initial request can be repeated until a syllable is found, or until a significant number of syllables have been accumulated. This procedure leads to a prediction of acoustic transformation parameters for any symbolic context even if it was not observed before; it allows the model to become general.

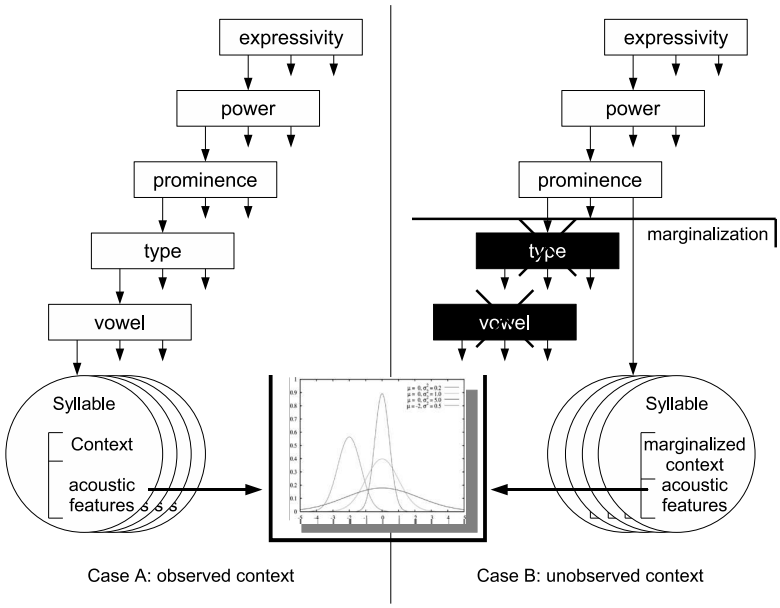

Figure 4. Decision tree involving a hierarchy of contextual variables. Left: Observed case of the database. Right: Unobserved case that requires recursive context marginalisation until it reaches an observed case.

5.5. Prosody transformation

After the two inference phases, the proposed prosodic model, fed by the input sequence of contexts (C^{N}_{src}), provides two sequences of acoustic parameters: A^{N}_{model} (neutral) and A^{E}_{model} (desired expressivity). Significant differences can be seen as effects of changing from one expressivity to another (see Section 2.4). These two sequences are compared in order to provide dynamic speaker-independent transformation controls that drive signal processing algorithms (SuperVP and ARX-LF model). Some examples can be listened to at the following address: <http://www.ircam.fr/anasyn/-beller>

6. Conclusion

In this chapter, we have presented a review of theories that deal with control of emotional reactions. They raise the difficulty of evaluating the degree of spontaneity in an emotion by observing only its external expression. Although empirical methods involving direct and indirect measurements have been used to collect emotional data, it is difficult for them to identify the inherent degree of control the speaker has on his emotional reaction, whether it is simulated or not. Therefore, emotions and their corresponding expressions have to be distinguished.

We have provided an attempt at defining expressivity as an information level of communication. This involves all perceivable demonstrations, controlled or not, of an internal state that can be interpreted as uncontrolled. Expressivity can formally be separated from other information levels, e.g. speaker identity, speaking style, modality and prominence. In practice, the task is harder since all these information levels share the same paralinguistic channel composed of restructurings of verbal words and non-verbal sounds, all modulated by prosody. This problem can be solved by discarding prosodic variations due to other information levels, e.g. involving the so-called neutral case.

By contextually clustering an expressive French corpus, significant prosodic variations can be attributed to the change of expressivity. A statistical context-dependent method has been described. The resulting speaker independent model of expressivity has been used to modify the expressivity of a neutral utterance, either recorded or synthesised. The proposed method is aimed at artistic ends, but can also be applied to the analysis of anthropomorphisation of Human-machine interfaces.

Acknowledgments

This work was partially funded by the French RIAM network project VIVOS (<http://www.vivos.fr>). The authors thank the actors involved in this study for their performances.

References

Aubergé, Véronique / Audibert, Nicolas / Rilliard, Albert, 2004. E-Wiz: A trapper protocol for hunting the expressive speech corpora, *LREC04*, Lisbon, 179-182.

Aubergé Véronique / Audibert Nicolas /Rilliard Albert, 2006. Auto-annotation: an alternative method to label expressive corpora, *LREC06*, Workshop on Emotional Corpora, Genova, 45-46.

Averill James R., 1980. A constructivist view of emotion, In *Theory, Research and Experience*, New York: Academic Press, 1/305-339.

Beller Grégory, 2007. Context Dependent Transformation of Expressivity in Speech Using a Bayesian Network, *ParaLing*, Saarbrücken.

Beller Grégory, Forthcoming. Semi-Parametric Synthesis of Speaker-Like Laughter, *The Phonetics of Laughing*. Berlin : de Gryuter.

Beller Grégory / Obin Nicolas / Rodet Xavier. 2008a. Articulation Degree as a Prosodic Dimension of Expressive Speech, *Speech Prosody 2008*, Campinas.

Beller Grégory / Schwarz Diemo / Hueber Thomas / Rodet Xavier, 2006. Speech Rates in French Expressive Speech, *Speech Prosody 2006*, Dresden, 2006.

Beller Grégory / Veaux Christophe / Degottex Gilles / Obin Nicolas / Lanchantin Pierre / Rodet Xavier, 2008b. IRCAM Corpus Tools: Système de Gestion de Corpus de Parole, In *TAL*.

Beller Grégory / Veaux Christophe / Rodet Xavier, 2008c. IrcamCorpusExpressivity: Nonverbal Words and Restructurings, *LREC workshop on emotions*.

Blankinship Eric / Beckwith Richard, 2001, Tools for expressive text-to-speech markup, *UIST'01,* New York, 159-160.

Bogaards Niels / Roebel Axel / Rodet Xavier, 2004. Sound Analysis and Processing with AudioSculpt 2, *ICMC,* Miami.

Buck Ross, 1985. The Communication of Emotion, *Paperback Edition.* New York : Guilford Press.

Bulut Murtaza / Lee Sungbok / Narayanan Shrikanth, 2007. A statistical approach for modelling prosody features using pos tags for emotional speech synthesis, *ICASSP*, Hawaï.

Burkhardt Felix / Audibert Nicolas / Malatesta Lori / Turk Oytun / Arslan Levent / Aubergé Valérie, 2006. Emotional Prosody - Does Culture Make A Difference ?, *Speech Prosody*, Dresden.

Campbell Nick / Mokhtari Parham, 2003. Voice quality: the 4th prosodic dimension, *XVth Int. Congress of Phonetic Sciences*, Barcelona, 3/2417-2420.

Changeux Jean-Pierre, 1983, *L'Homme neuronal*, Paris : Odile Jacob.

Chung Soo-Jung, 2000. L'expression et la perception de l'émotion extraite de la parole spontannée : évidences du coréen et de l'anglais, phonétique, *PhD thesis Université PARIS III*, Paris.

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

Darwin Charles, 1872. *Expression of Emotion in Man and Animal.* Oxford : Oxford University Press.

De Cheveigné Alain / Kawahara Hideki, 2002, YIN, a Fundamental Frequency Estimator for Speech and Music, *JASA*, 111/1917-1930.

Descartes René, 1664. *Les passions de l'Âme*. Paris : Michel Bobin et Nicolas Le Gras

Devillers Laurence / Lamel Lauri / Vasilescu Ioana, 2003. Emotion detection in Task-oriented spoken dialogs. *IEEE ICME*.

Douglas-Cowie Ellen / Campbell Nick / Cowie Roddy / Roach Peter, 2003. Emotional speech: towards a new generation of databases, *Speech Communication*., 40/1-2/33-60.

Eide E. / Aaron A. / Bakis R. / Hamza W. / Picheny M. / Pitrelli J., 2004. A Corpus-Based Approach to <Ahem/> Expressive Speech Synthesis, *5th ISCA Speech Synthesis Workshop*.

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

Fant Gunar / Liljencrants Johan / Lin Qiguang, 1985. A four-parameter model of glottal flow, *STL-QPSR*, 4/1-13.

Fónagy Ivan, 1983. *La vive voix : essais de psychophonétique*. Paris ; Payot.

Frijda Nico H., 1986. *The emotions*. Cambridge : Cambridge University Press.

Garnier Maëva, 2007. Communiquer en environnement bruyant : de l'adaptation jusqu'au forçage vocal, *PhD thesis University of Paris 6*.

Gendrot Cédric / Adda-Decker Martine, 2004. Analyses formantiques automatiques de voyelles orales: évidence de la réduction vocalique en langues française et allemande, *MIDL*..

Gobl Christer / Chasaide Ailbhe Ni, 2003. The role of voice quality in communicating emotion, mood and attitude, *Speech Communication*. 40/1-2/189-212.

Henrich Nathalie / d'Alessandro Christophe. / Doval B., 2002. *Glottal Flow Models: Waveforms, Spectra And Physical Measurements*.

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

Lacheret-Dujour Anne / Beaugendre Frédéric, 1999. *La prosodie du Français*, Paris : CNRS langage.

LeDoux Joseph, 2005. *Le Cerveau des émotions*, Paris : Lavoisier.

Lindblom Bjorn, 1983. Economy of Speech Gestures, In *The Production of Speech*, New-York: Spinger-Verlag.

Peeters Geoffroy, 2004. A large set of audio features for sound description (similarity and classification) in the CUIDADO project, *Technical report IRCAM*.

Pereira Cecile / Watson Catherine, 1998. Some acoustic characteristics of emotion, *Fifth International Conference on Spoken Language Processing*, Sydney.

Pfitzinger Helmut, 2006. Five Dimensions of Prosody: Intensity, Intonation, Timing, Voice Quality, and Degree of Reduction, In H. Hoffmann, R.; Mixdorff (eds.), *Speech Prosody*, Dresden, 40/6-9.

Piu Marie / Bove Rémi, 2007. Annotation des disfluences dans les corpus oraux, *RECITAL*.

Scherer Klaus. R., 2006. The Affective and Pragmatic Coding of Prosody, In *Chinese Spoken Language Processing*, 4274/13-14.

Schwarz Diemo, 2004. Data-Driven Concatenative Sound Synthesis, *PhD thesis Université Paris 6.*

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

Vidrascu Laurence / Devillers Laurence, 2005. Detection of Real-Life Emotions in Call Centers, *Interspeech.*

Vincent Damien / Rosec Olivier / Chonavel Thierry., 2005. Estimation of LF glottal source parameters based on an ARX model, *Interspeech*.

Yamagishi Junichi / Onishi Koji / Masuko Takashi / Kobayashi Takao, 2005, Acoustic modelling of speaking styles and emotional expressions in HMM-based speech synthesis, *IEICE Trans. on Inf. & Syst.*, E88-D/503-509.