Speech Rates in French Expressive Speech

Grégory Beller, Thomas Hueber, Diemo Schwarz & Xavier Rodet

Ircam, Institut de Recherche et de Coordination Acoustique/Musique
1, place Igor Stravinsky
75004 Paris, France
{beller; hueber; schwarz; rodet}@ircam.fr

Abstract

Expressive speech is a useful tool in cinema, theater and contemporary music. In this paper we present a study on the influence of expressivity on the speech rates of a French actor. It involves a relational database containing expressive and neutral spoken French. We first describe the analysis partly based on a unit-selection Text-to-Speech system. The range of data permits a statistical approach to the speech rate. A dynamic description of the French speech rate is offered which demonstrates its evolution in speech. Finally, several results are given concerning pauses and breathing that help to distinguish between anger and happiness.

1. Introduction

In previous work, we have developped [12] a musical concatenative sound synthesis system called CATERPILLAR. This framework has been extended into a Text-to-Speech (TTS) system, called TALKAPILLAR [2]. One of the aims of this system is to reconstruct the voice of a speaker, for instance a deceased eminent personality. TALKAPILLAR is designed to reproduce fixed texts as if they were spoken by the original specific speaker. The system also allows analysis of expressive speech for artistic purposes. Some contemporary music composers are interested in vocal correlates of emotions and want to easily explore and use expressive databases. A film dubbing studio would like to use an expressive speech synthesizer. Some theater directors would like to transform and to synthesize voices on stage, for instance, to switch between different voice types and expressivities.

In this study, we have recorded a French actor to build an expressive speech database. By acoustical analysis of the speech signal, we have constructed a prosodic model of the ways he has conveyed expressivity. We voluntarily do not treat voice quality in order to concentrate us on prosodic cues. After a quick overview of related work, this article presents a part of the system dedicated to analysis. This framework naturally allows for a statistical examination of speech phenomena.

Contrary to the well defined and consensual definition of fundamental frequency, speech rates are harder to define. They could be defined as an overall movement of what is said [7], taking into account silent pauses, breaks, accelerations and decelerations.

A syllabic time language (in Pike or Abercrombie's typology) like French constructs its rhythm on accentuated syllables which tend to have greater duration than non accentuated syllables independantly of the average speech rate [6].

If a syllable has a greater duration than others, it is for a prosodic reason rather than a phonetic reason, as it seems that a consensus on the existence of the psycho-rhythmic role of the syllable in French has been established [15], and that the speech rate is more related to syllables than to phonetic units [10]. This hypothesis of a constant syllabic rate has been reinforced in a study of voice production [14]. Thus, such a strong relationship between speech rate and syllable duration in French is used in this paper. This article offers results concerning the speech rate under different expressivities. Results show that speech rate is an important prosodic cue to distinguish happy from angry speech.

2. General overview of the system

All the processes involved in our expressive speech analysis are presented and summarized in figure 1. The text is first analyzed to provide symbolic information like phonetic transcription and expected accentuated syllables.The corresponding audio is segmented by alignment and tagged with expressivity markers. Different acoustic analyses such as fundamental frequency estimation are calculated on the signal. Using time based segmentation, *characteristic values* modeling time based evolutions of the features within units are computed. All these information are synchronized and stored in a database. An effective interface allows graphic exploration, concatenative synthesis and content-based transformation.

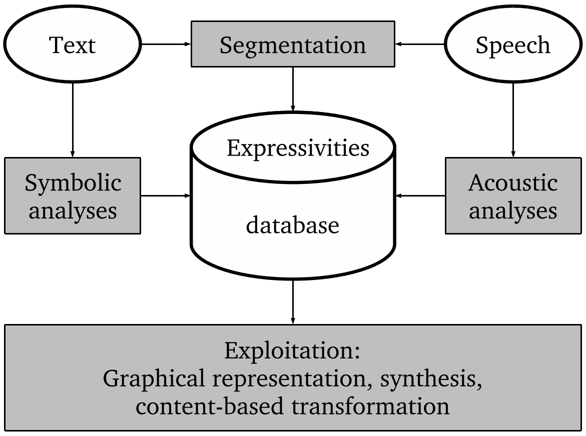

Figure 1: Overview of the system.

3. Database

3.1. Database Content

For this experiment, we built a database of approximately two hours of French speech. It is composed of the neutral and expressive speech of a French actor recorded in an anechoic studio. The corpus is composed of a set of 26 sentences of variable length. To provide the same prosodic boundaries for all expressivities, stressed syllables have been signalled to the speaker in the text by punctuation and underlined parts of words. This prevents from possible prosodic reorganizations due to speech rate changes [6] and leads to a nearly constant distribution of ”accentuated” syllables. Each sentence was pronounced with the following expressivities: *Neutral, neutral question, anger, happiness, sadness, boredom, disgust, indignation, positive and negative surprise*. For some expressivities, three occurrences per sentence were spoken in order to have a variation of the activation level (*low, middle, high*). After post-processing of the recordings, the actor was invited to delete any speech which did not match the goal.Finally 539 sentences have been retained for the analysis.

3.2. Database Interface

Since a large amount of data has been used for this study (see section 4), an efficient database architecture is needed. A relational database management system (DBMS) is used in this project to reliably store data files, tens of thousands of units, their interrelationships and feature data. The database is clearly separated from the rest of the system by a database interface, written in *Matlab* and *PostgreSQL*. As an example of the power of the interface language, let us show a typical command of this language as given to *Matlab*:

>>dbi('getunidata', 'unit', dbs('getuidsfromsymbol', 'sOn', UnitType.syllable), FeatureType.f0, 'slope').

It returns the slopes of the evolutions of the fundamental frequency F0 of all the syllables ”s0n” in the database. All sorts of similar queries can be done, and the result easily further filtered, if necessary, in Matlab. The database can also be browsed with a graphic database explorer that allows users to visualize all data and play units. For instance, figures 2, 3 and 4 are displayed as the result of a simple mouse click. Up to now, Sound Description Interchange Format (SDIF) [13] has been used for well-defined exchange of data with external programs (analysis, segmentation).

3.3. Database segmentation

The first step of the analysis is the segmentation of recorded speech, in variable length units. A simple speech alignment is employed for this segmentation. Speech alignment connects units in a text to corresponding points on the speech signal time axis. In order to align a sentence, from its phonetic transcription (see section 4.1), a rudimentarily synthesized sentence is built with diphones coming from a small manually labelled database. Then mel frequency cepstrum coefficients (MFCC) sequences of the two sentences are calculated and aligned with a dynamic time warping (DTW) algorithm. This provides a segmentation into *semi-phones, phones, diphones, syllables, prosodic groups and sentences* (see figure 2).

4. Features

All these units are labeled and endowed with three types of features:

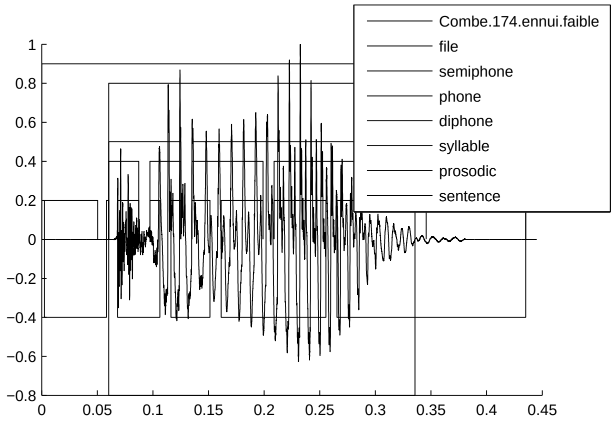

Figure 2: Example of speech segmentation for the French sentence: ”*Comment?*”

4.1. Symbolic features

Category features position a unit in a category or class and its relationship to all classes in the hierarchy (e.g. speaker → actor → male, for the sound source hierarchy). The phonetic and syntactic description of the text is provided by the EULER program [1] issued from the TTS project MBROLA. EULER analyzes a text and gives several symbolic representations such as a phonetic transcription (XSampa) and a grammatical analysis. It also gives boundaries of syllables and it predicts if they will be accentuated or not. Predicted accentuated syllables are employed to define prosodic boundaries since one prosodic group corresponds to a sequence of unaccentuated syllables followed by an accentuated one for the specific case of French [8]. A test on a database of 1153 neutral French sentences shows that the mean of the accentuated syllable's duration is approximately twice the mean of the non accentuated syllable. These descriptions and other added symbolic features, corresponding to the relative places of the units, are stored in SDIF files.

4.2. Dynamic features

Dynamic features are analysis data evolving trough a unit (e.g. fundamental frequency).

4.2.1. Speech rate

Speech rate is often defined as the average number of syllables per second in a whole sentence (see [5, 9]). 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 a linear interpolation of the durations of syllables (see figure 3). By use of segmentation and alignment with the symbolic syllable boundaries given by EULER, we obtain a dynamic evolution of the speech rate over the sentence. A deceleration corresponds to a rising of the curve and an acceleration is represented by a falling of the curve.

In figure 3, the speech rate curve presents local maxima in syllable duration corresponding to decelerations. Framed syllables are the ones considered as accentuated by the text-to-prosody generator of EULER. For this example, EULER gives a good prediction of the prosody pronounced by the actor although it has been designed for neutral and not for expressive speech. It can be seen that there is a strong correlation between the speech rate curve and F0. Moreover the final accent is more distinguishable in the speech rate curve than in the F0 curve.

The main advantage of this description of the speech rate is its relative dynamic. In fact, we see clearly in figure 3 that

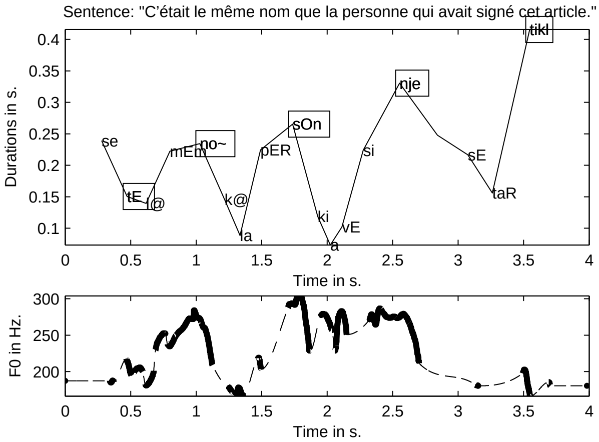

Figure 3: Durations of syllables and F0 of a French sentence pronounced with happiness: *"C'était le même nom que la personne qui avait signé cet article."*

the actor seems to emphasize the sentence with an increase of the accentuated syllable duration which gives a certain rhythm to his performance.

4.2.2. Fundamental frequency and energy

Fundamental frequency (F0) is computed by the YIN algorithm [4]. This algorithm also gives the energy and the harmonic-to-noise ratio (also called aperiodicity) of the signal for each computed frame. By thresholding the aperiodicity curve, we keep F0 and energy estimation only on the voiced parts of the signal.

4.3. Static features

Static features give to a unit constant values. They mainly model time-based evolution of dynamic features (see 4.2). A set of *characteristic values* is represented in figure 4 and is composed of:

- arithmetic and geometric mean average and standard deviation
- minimum, maximum, and range slope, giving the rough direction of the feature's movement, and curvature (from 2nd order polynomial approximation)
- value and curve slope at start and end of the unit
- temporal center of gravity/anti-gravity, giving the location of the most important elevation or depression in the feature curve and the first four order temporal moments
- normalized Fourier spectrum of the feature in five bands, and the first four order moments of the spectrum. This reveals if the feature has rapid or slow movement, or if it oscillates (used to measure jitter and shimmer).

5. Results

The framework presented previously is a tool for statistical exploration of speech data where a large amount of data is accessible by a simple and powerful interface. We show here results in figure 5 and figure 6 where statistical properties of each expressivity are summarized by an ellipse. Center coordinates of each ellipse correspond to the averages of the two involved features. The X and Y widths represent standard deviations.

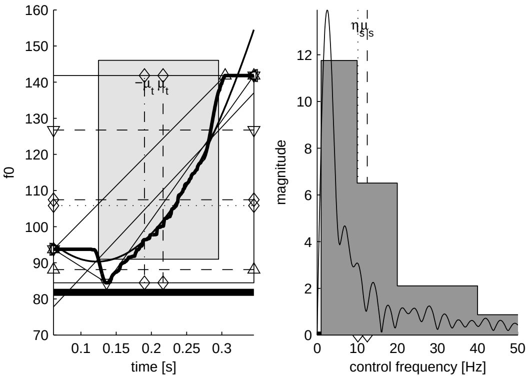

Figure 4: Example of *characteristic values* of the fundamental frequency computed over the French utterance: *"Comment?"*

5.1. Distinction between happy and angry speech

It has been mentioned that happiness and anger share the same prosodic aspects [5]. In fact, we can see on the X-coordinate of figure 5, that they both have an average F0 around 250Hz for this actor. It is well known that F0 is a prime prosodic cue. [11]. Thus many algorithms fail to distinguish between happiness and anger by over accentuating the effect of the fundamental frequency. The Y-coordinate of figure 5 shows the distribution of the speech rate for each expressivity. We see that happy speech is pronounced slightly slower than angry speech. But the main difference is that happy speech exhibits a very large standard deviation of speech rate. This means that the prosodic strategy taken by the actor for happiness involves frequent fast changes in the speech rate. These changes occur precisely in accentuated syllables, some of which (final ones) are almost sung. We can see in figure 5 that sadness is also expressed with a high F0, contrary to some papers relating a low F0 for this expressivity. Such a difference shows that multiple strategies can be employed by an actor to express an affect (whining for sadness in this case). Thus, the results presented in this paper should be taken carefully as examples rather than generalities. However, this information is vital to simulate the style of an actor and could be used to distinguish more clearly between happiness and anger in an expressivity recognition system.

5.2. Pauses and breathing

Another interesting result comes from the examination of the duration of the pause and breathing. When breathing is audible, it gives a prosodic cue which is part of the speech rate (see section 4.2.1). The examination whether pauses are silent or filled by an audible breathing could improve the distinction between negative and positive surprise. As is shown in figure 6, it seems that the actor used silent pauses for positive surprise and shorter and breathy pauses for negative surprise. An examination of plosive explosion duration reinforces the distinction. We can also see that the strategy to express fear employs long pauses that cover almost half the duration of the sentence. It has to be correlated with a high average speech rate for this expressivity.

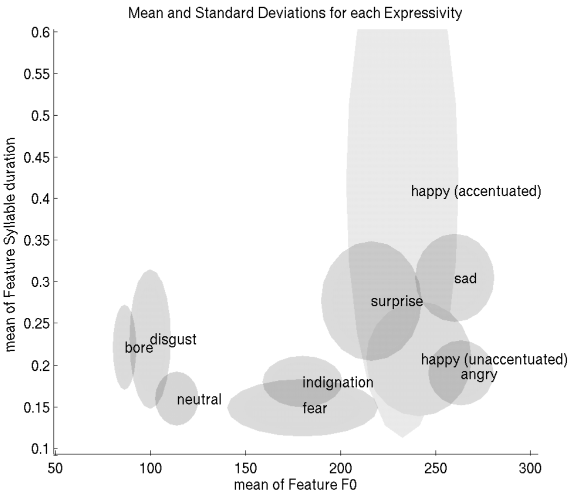

Figure 5: Average speech rate vs average F0.

5.3. Content-based transformations

A first attempt at expressive speech synthesis has been done with content-based transformations. The observations presented in this section are used to transpose and time-stretch segmented audio with a phase vocoder [3]. The coefficients of these elementary transformations change along the sentence, depending on the context of the units. Some examples exhibit the need of a voice quality treatment and can be heard at the following address: <http://www.ircam.fr/anasyn/concat>.

6. Conclusion and future work

In this paper, we have presented different results about the influence of expressivity on the speech rate of a French actor. They could help to distinguish expressivities, to improve expressive speech synthesis and to inspire rule-based transformations. Different steps of the analysis process have been presented so as to provide a global comprehension of a framework, able to statistically analyze large speech corpora. We use it for artistic purposes dealing with cinema, theater and contemporary music.

Future work will be concentrated on the analysis of rhythmic patterns in expressive speech such as the one shown in figure 3. Another future direction is the analysis and transformation of voice quality as connected to expressivity.

7. Acknowledgments

The authors would like to thanks the French actor Jacques Combe for this performance.

8. References

[1] Bagein, M.; Dutoit, T.; Tounsi, N.; Malfrère, F.; Ruelle, A.; Wynsberghe, D.,2001. Le projet EULER, Vers une synthèse de parole générique et multilingue. *Traitement automatique des langues*, 42[1].

[2] Beller, G.; Schwarz, D.; Hueber, T.; Rodet X., 2005. Hybrid concatenative synthesis in the intersection of speech and music. *JIM2005* Paris:CICM, 41-45.

[3] Bogaards, N.; Roebel, A.; Rodet,X., 2004. Sound analysis and processing with audiosculpt 2. *ICMC* Miami.

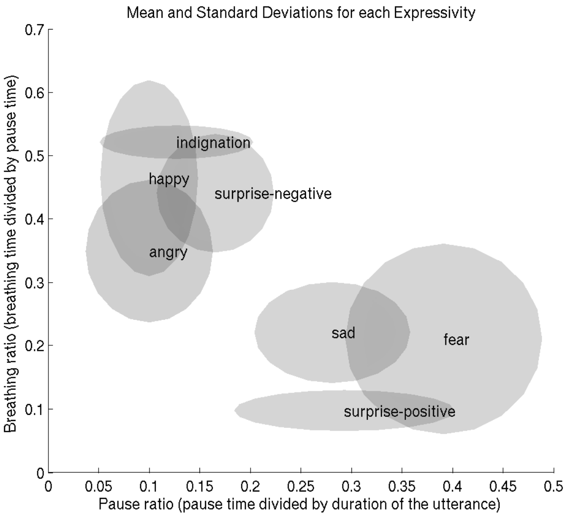

Figure 6: pause durations divided by sentence durations in X-coordinate and breathing durations divided by pause durations in Y-coordinate.

[4] de Cheveigné, A.;Kawahara,H., 2002. YIN, a Fundamental Frequency Estimator for Speech and Music. *JASA*, 111, 1917-1930.

[5] Chung, S.-J., 2000 L'expression et la perception de l'émotion extraite de la parole spontannée: évidences du coréen et de l'anglais. *Thèse de phonétique, Université PARIS III - Sorbonne Nouvelle* Paris.

[6] Fougeron, C.; Jun, S.-A., 1998 Rate effects on French intonation: prosodic organization and phonetic realization. *Journal of Phonetics*, 26, 45-69.

[7] Galarneau, A.; Tremblay, P.; Martin, P., Dictionnaire de la parole. *laboratoire de Phonétique et Phonologie de l'Université Laval à Québec*.

[8] Malfrère, F.; Dutoit, T.; Mertens, P., 1998. Automatic prosody generation using suprasegmental unit selection. *SSW3*, 323-328.

[9] Pereira, C.; Watson, C., 1998. Some acoustic characteristics of emotion *Fifth International Conference on Spoken Language Processing* Sydney.

[10] Rouas, J.-L.;Farinas, J.; Pellegrino, F., 1998. Evaluation automatique du débit de parole sur des données multilingues spontannées.

[11] Scherer, K.R., 1989. Vocal correlates of emotion. *Handbook of psychophysiology: Emotion and social behavior* London:Wiley, 165-197.

[12] Schwarz, D., 2004. Data-Driven Concatenative Sound Synthesis. *thèse d'informatique, Université Paris VI - Pierre et Marie Curie* Paris.

[13] Schwarz, D.;Wright,M.,2000. Extensions and Applications of the SDIF Sound Description Interchange Format. *ICMC*, Berlin, 481-484.

[14] Wheeldon, L., 1994. Do speakers have access to a mental syllabary. *Cognition*, 50, 239-259.

[15] Zellner, B., 1998. Caractérisation du débit de parole en français. *Journées d'Etude sur la Parole*.