An Overview of Talkapillar

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

In this paper we present a system devoted to expressive speech for artistic purposes such as cinema, theater and contemporary music. It involves a relational database containing expressive and neutral french utterances. We describe the analysis system partly based on a concatenative Text-To-Speech system. A large set of descriptors of the segmented data permits a statistical approach of the speech rate.

1. Introduction

In previous work [1] a musical concatenative system named CATERPILLAR has been elaborated. This framework has been extended towards 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 should pronounce texts as if they were said by the target specific speaker. The system allows expressive speech analysis for artistic requests. Some contemporary composers are interested in vocal correlates of emotions and want to easily explore and use expressive databases. A film dubbing studio would 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. After a quick overview of related work, this article presents the analysis system. This framework naturally allows for a statistical examination of speech phenomena.

Concatenative unit selection speech synthesis from large databases, also called corpus based synthesis [3], is now used in many TTS systems for waveform generation [4]. Recently, the development of corpus based methods and the increasing size of databases widens TTS systems towards ESS. Some [5] recorded several speech corpora pronounced under different emotions and used classic TTS methods on these separated databases to provide ESS. An attempt to group the different corpora without formal separations into a same speech database have been made [6].

2. General overview of the system

All the processes involved in the expressive speech analysis are presented and summarized in figure 1. Analyzing expressive speech needs to manage with two information levels conveyed by the speech: prosody and voice quality. For instance, it is often related that happiness and anger are demonstrated by an increase of the mean of the fundamental frequency [7, 8]. But differences still remain in the voice quality since anger has sometimes a stronger jitter for instance [9]. The framework presented here involves processes with these two information levels.

The text is first analyzed to provide symbolic informations like phonetic transcription and predicted accentuated syllables [10]. The corresponding audio is segmented by alignment and tagged with expressivity. Different acoustic analyses such as fundamental frequency estimation are calculated on the signal. Profiting from the previous temporal segmentation, *characteristic values* modeling temporal data evolutions among units are computed. All these informations are synchronized and stored in a database. An effective interface allows graphical exploration, concatenative synthesis and content-based transformation.

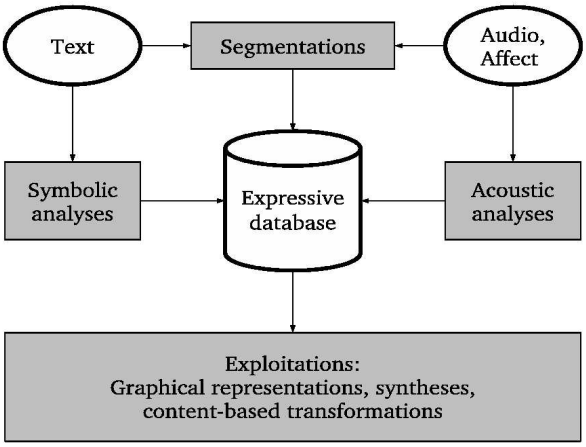

Figure 1: Overview of the system.

3. Database

3.1. Database Content

For this experiment, we built a database of approximately two hours of french speech. It is composed of neutral and expressive utterances pronounced by a french actor. He has been recorded in an anechoic chamber. The corpus is composed of a set of 26 sentences of variable length. To avoid ambiguities,some prosodic boundaries have been specified by punctuations and underlined parts of words. Each sentence was pronounced with the following expressivities: *Neutral, neutral question, angriness, happiness, sadness, boreness, disgust, indignation, positive and negative surprises*. For some expressivities, three occurrences per sentence were pronounced to have a variation of the acti-

vation level (*low, middle, high*). After post-processing of the recordings, the actor has freely eliminated utterances considered as mis-matching the goal. Finally 539 utterances 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 interrelationship and descriptor data. The database is clearly separated from the rest of the system by a database interface, written in Matlab and procedural SQL. Therefore, the current DBMS can be replaced by another one, or other existing speech databases can be accessed. 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', UnitTypes.syllable), FeatureTypes.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 be also browsed with a graphical database explorer that allows users to visualize all data and play units. For instance, figure 2, figure 3, and figure 4 are displayed as the result of a simple mouse click. For the latter, Sound Description Interchange Format (SDIF) [11] is 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 utterances, 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. From the phonetic transcription of the sentence to align (see section 4.1), a rudimentary synthesized sentence is built with diphones coming from a small hand-labelled database. Then MFCC sequences of the two sentences are computed and aligned with a DTW algorithm. This provides a segmentation into *semi-phones, phones, diphones, syllables, prosodic groups and sentences* (see figure 2).

4. Descriptors

All these units are labeled and informed with three types of descriptors:

4.1. Symbolic descriptors

Category descriptors express the membership of a unit to a category or class and all its base 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 [12] issued from the TTS project MBROLA. This module 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 are accentuated or not. Predicted accentuated syllables are employed to define prosodic boundaries since one prosodic group corresponds to a sequence of none accentuated syllable ended by an accentuated one [10]. A test on a database

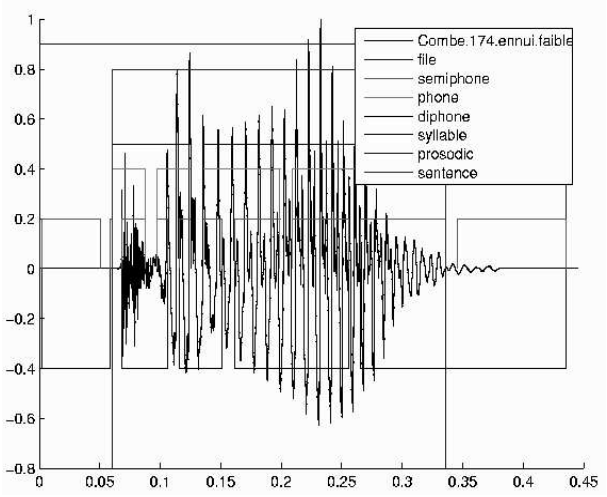

Figure 2: Example of speech segmentation for the french sentence "Comment?".

of 1153 neutral french utterances shows that the mean of the accentuated syllable's durations is approximately the double of the mean of the non accentuated syllable's durations. These descriptions and other added symbolic descriptors corresponding to the relative places of the units from one to each others are stored into SDIF files.

4.2. Dynamic descriptors

Dynamic descriptors are analysis data evolving over a unit (e.g. fundamental frequency).

4.2.1. Speech rate

Speech rate is often defined as a means over all an utterance, in syllables per second (see [7, 8]). Because the most prominent syllables have often a longer duration, we prefer 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 the segmentation and the alignment with the symbolic syllable's boundaries given by EULER, we obtain a dynamic evolution of the speech rate over the utterance. A deceleration corresponds to a rising of the curve and an acceleration is represented by a falling of the curve.

figure 3 shows several informations: The speech rate curve presents local maxima corresponding to decelerations. The 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 speech and not for expressive utterances. 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 as in the F0 curve.

The main advantage of this description of the speech rate is its relative dynamicity. In fact, we see clearly on figure 3 that the actor emphasizes the sentence by an increase of the accentuated syllable duration which gives a certain rhythm to his performance.

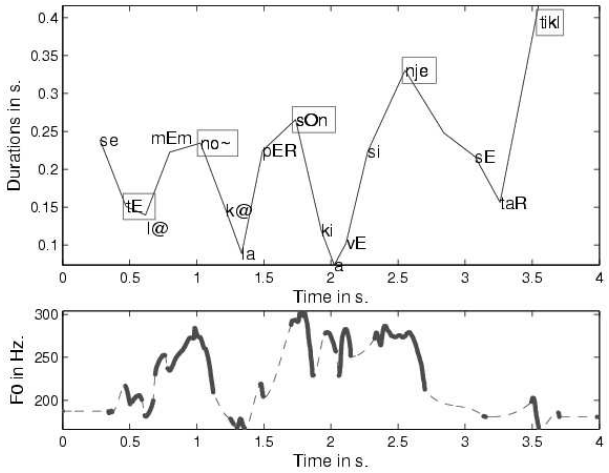

Figure 3: Durations of syllables and F0 of a french sentence pronounced with happiness: *"C'était le même nom que la personne qui avait signé cet article."*

4.2.2. Fundamental frequency and energy

Fundamental frequency (F0) is computed by the YIN algorithm [13]. 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 just keep F0 and energy estimations on voiced part of the signal.

4.3. Static descriptors

Static descriptors take a constant value for a unit. They mainly model the temporal evolution of the preceding dynamic descriptors (see 4.2). A vector of *characteristic values* is represented on figure 4 and is composed of:

- arithmetic and geometric mean, standard deviation
- minimum, maximum, and range slope, giving the rough direction of the descriptor movement, and curvature (from 2nd order polynomial approximation)
- value and curve slope at start and end of the unit (used for the calculation of the concatenation cost)
- the temporal center of gravity/anti-gravity, giving the location of the most important elevation or depression in the descriptor curve and the first 4 order temporal moments
- the normalized Fourier spectrum of the descriptor in 5 bands, and the first 4 order moments of the spectrum. This reveals if the descriptor has rapid or slow movement, or if it oscillates (used to measure Jitter and Shimmer).

5. Exploitations of the databases

5.1. Synthesis

Interestingly, ESS is widely expected for its artistic purposes. Many composers and directors want to work with vocal expressivity. Furthermore, several computer-game's creators wish to add expressive voices to characters in non-predicted scenario. The TALKAPILLAR synthesis system is not aimed at a classical TTS use as is the case of the majority of similar TTS systems.

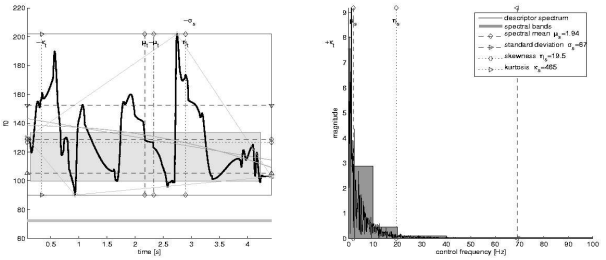

Figure 4: Example of *characteristic values* of the fundamental frequency computed over all an angry utterance.

It is mainly designed to (re)produce a specific expressivity. It offers an excellent framework for an artistic purpose since the user has access to all the steps of the process, which permits a full control of the result. A friendly user interface permits to avoid bad diphones (black list) and to transform the results with its own speech rate or f0 contour thanks to SVP. That's why interests for multimedia and cinema are shown and illustrated by the use of the system by a french dubbing company.

5.1.1. Features Involved

The database is filled with plenty of heterogeneous units: semiphones, phones, diphones, syllables, prosodic groups... Each of them is described by a set of features involving symbolic and acoustic analysis. These features are the basis for comparison between units and unit selection. Adequacy of units (target cost) and concatenation cost are computed as a weighted combination of the features. Minimization of the global cost of a unit selection is done by a Viterbi algorithm.

5.1.2. Unit Selection Algorithm

The classical unit selection algorithm finds the sequence of database units u_i that best match the given synthesis target units t_τ using two cost functions: The expresses the similarity of u_i to t_τ including a context of r units around the target. The predicts the quality of the concatenation of u_i with a preceding unit u_{i-1}. The optimal sequence of units is found by a Viterbi algorithm finding the best path through the network of database units.

5.1.3. Prosodic Unit Selection

The first step of TALKAPILLAR is to select supra-segmental units in a database. By a Viterbi algorithm applied on the symbolic representation of prosodic units, we access to the best prosodic sequence able to traduce the a special expressivity of a speaker. By choosing those prosodic groups in an expressive database, we can have a representation of a real evolution of acoustic parameters for a chosen affect. After having been rescaled by a speaker-dependent identity factor, we add these acoustic parameters of the chosen and transformed prosodics units to the segmental units of the target in the aim of select the best sequence of segmental units that fit to a real prosody excerpt from the database.

5.1.4. Segmental Units Selection

The second step consists in selecting the segmental units according to their symbolic representation and to the acoustic representation derived from the selection of prosodic units coming

from the previous step. A similar Viterbi selection algorithm is then applied to find the best sequence of segmental units that match the target string and the prosody selected.

5.1.5. Concatenation

When a correct sequence of units has been selected, it just needs to be concatenated in order to build the desired phrase. This is the last step of the synthesis process and could be aimed at two different goals. Concatenative synthesis has been designed to preserve all sound details so as to improve the quality and the naturalness of the result. In the TALKAPILLAR TTS system, a first strategy is to not transform chosen units. They are concatenated with a slight cross fade at the junction and a simple period alignment to not produce clicks and other artifacts.

5.2. Content-based Transformations

Another strategy consists in transforming some of the selected units before concatenating them. For instance, some of the units (the voiced ones) are slightly pitch-transposed to best match the prosody selected beforehand. They could also be time-stretched if the selected diphones length are to short compared to the desired speech rate. One can record the sentence with a different prosody and import it immediately in the database so as to provide new prosodic groups and then force the synthesis to follow its own expressivity. These transformations are accomplished with a phase vocoder technology based on a speech algorithm [14]. A first attempt to ESS has been done with content-based transformations. Observations are used to transpose and time-stretch segmented audio. The coefficients of these elementary transformations change along the sentence depending on the context of the units.

5.3. Hybrid synthesis

An interesting aspect of the system is that the TTS synthesizer TALKAPILLAR and the musical synthesizer CATERPILLAR share the same framework. Created on the same software architecture, these two synthesis systems joined together give a powerful tool for composers interested in interaction between music and speech [2]. For instance, voiced parts of a sentence could be replaced by cello's sustained units respecting the prosody of the replaced speech segments. The flexibility of the selection process's parametrization (features involved, cost weights, etc.) sets the user free to create very innovative hybrid synthetic phrases. Experiments have been made in creating hybrid synthetic phrases with speech units and any other sound units [15].

5.4. Prosody Extraction

An option of the system permits to extract prosodic units out of the database. Exportation of the acoustic features (f0, energy, Speech rate, etc.) and of the symbolic descriptors (grammatical structure, type of the final accent, . . .) into a SDIF file allows to exploit these informations in other frameworks.

6. Conclusion and Future Works

In this paper, we have presented a system to explore influences of the expressivity on the speech of a french actor. Different steps of the analysis process have been presented so as to provide a global comprehension of a framework able to analyze, transform and synthesize expressive speech. The framework presented previously is a tool for statistical explorations of databases. A large amount of data combined with a simple and powerfull interface makes it really effective. We use it for artistic purposes dealing with cinema, theater and contemporary music.

Some examples can be listened from the following address: <http://recherche.ircam.fr/equipes/analyse-synthese/concat>.

Future works will be now concentrated on the analysis of rhythmic patterns in expressive speech such as the one showed in figure 3. Another future direction is the analysis and transformation of voice quality.

7. acknowledgments

The authors would like to thanks the french actor Jacques Combe for its performance.

8. References

[1] D. Schwarz, "New Developments in Data-Driven Concatenative Sound Synthesis," in *Proceedings of the International Computer Music Conference (ICMC)*, Singapore, Oct. 2003, pp. 443–446.

[2] G. Beller, D. Schwarz, T. Hueber, and X. Rodet, "Hybrid concatenative synthesis in the intersection of speech and music," *JIM*, vol. 12, pp. 41–45, 2005. [Online]. Available: http://mediatheque.ircam.fr/articles/textes/Beller05c/

[3] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Atlanta, GA, May 1996, pp. 373–376.

[4] R. Prudon and C. d'Alessandro, "A selection/concatenation TTS synthesis system: Databases developement, system design, comparative evaluation," in *4th Speech Synthesis Workshop*, Pitlochry, Scotland, 2001.

[5] A. Black, "Unit selection and emotional speech," *Eurospeech*, 2003.

[6] M. Bulut, S. Shrikanth, S. Narayanan, and A. K. Syrdal, "Expressive speech synthesis using a concatenative synthesizer," in *ICSLP*, ATT Labs-Research, Florham Park, NJ, 2002.

[7] S.-J. Chung, "L'expression et la perception de l'émotion extraite de la parole spontannée: évidences du coréen et de l'anglais," phonétique, Université PARIS III - Sorbonne Nouvelle: Institut de Linguistique et Phonétique Générales et Appliquées, Paris, 2000, sous la direction de J. Vaissière.

[8] C. Pereira and C. Watson, "Some acoustic characteristics of emotion," in *Fifth International Conference on Spoken Language Processing, Sydney*, 1998.

[9] K. Scherer, *Vocal correlates of emotion*, 1989, pp. 165–197.

[10] F. Malfrère, T. Dutoit, and P. Mertens, "Automatic prosody generation using suprasegmental unit selection," in *SSW3*, 1998, pp. 323–328.

[11] D. Schwarz and M. Wright, "Extensions and Applications of the SDIF Sound Description Interchange Format," in *Proceedings of the International Computer Music Conference (ICMC)*, Berlin, Germany, Aug. 2000, pp. 481–484.

[12] M. Bagein, T. Dutoit, N. Tounsi, F. Malfrère, A. Ruelle, and D. Wynsberghe, "Le projet EULER, Vers une synthèse de parole générique et multilingue," *Traitement automatique des langues*, vol. 42, no. 1, 2001.

[13] A. de Cheveigné and H. Kawahara, "YIN, a Fundamental Frequency Estimator for Speech and Music," *Journal of the Acoustical Society of America (JASA)*, vol. 111, pp. 1917–1930, 2002.

[14] N. Bogaards, A. Roebel, and X. Rodet, "Sound analysis and processing with audiosculpt 2," in *International Computer Music Conference (ICMC)*, Miami, USA, Novembre 2004. [Online]. Available: http://mediatheque.ircam.fr/articles/textes/Bogaards04a/

[15] G. Beller, "La musicalité de la voix parlée," Maitrise de musique, Université Paris 8, Paris, 2005. [Online]. Available: http://mediatheque.ircam.fr/articles/textes/Beller05a/