FULLY AUTOMATIC PROSODY GENERATOR FOR TEXT-TO-SPEECH

F. Malfrère[1], T. Dutoit[1] and P. Mertens[2]

[1] Faculté Polytechnique de Mons, Boulevard Dolez 31, 7000 - Mons (Belgium)
Email : {malfrere,dutoit}@tcts.fpms.ac.be

[2] K.U.Leuven - Département de Linguistique
Blijde Imkomstraat 21 3000 Leuven
Email : Piet.Mertens@arts.kuleuven.ac.be

ABSTRACT

Text-to-Prosody systems based on the use of prosodic databases extracted from natural speech will be a key point for further development of new Text-to-Speech systems.

This paper describes a system using such speech databases to generate the rhythm and the intonation of a French written text. The system is based on a very crude chinks 'n chunks prosodic phrasing algorithm and on a prosodic analysis of a natural speech database. The rhythm of the synthetic speech is generated with a CART tree trained on a large mono-speaker speech corpus. The acoustic aspect of the intonation is derived from a set of prosodic patterns automatically derived from the same speech corpus.

The system has been tested on single sentences and news paragraphs. Informal listening tests have shown that the resulting prosody is convincing most of the time.

1. INTRODUCTION

One of the major problems in text-to-speech synthesis systems consists in the automatic generation of a natural and intelligible prosody. Prosody is typically described in terms of fundamental frequency contours (for voiced portions of speech) and of duration of speech segments (mainly phonemes or syllables). Speech synthesizers driven with correct prosodic information are currently able to produce very high quality synthetic speech [1][2]. Different methods has been proposed for prosody generation, mainly rule-based approaches [3][4] or corpus based approaches based on automatic learning techniques like neural networks [5][6], linear regression [7] or other statistical approaches [8].

This communication presents a prosody generation system based on the use of a fully automatically analyzed speech corpus. Phonetic/syllabic segmentation and the f0 curve are extracted with the MBROLIGN system [9] and grammatical tags are set by LIPSS [10]. The prosody generation system first generates a symbolic description of the prosody from the input text, and then transcribe this abstract representation into acoustic features (F0 and phoneme durations).

In the following section, the prosody generator will be described. This section begins with the description of the symbolic representation and presents the algorithm used to derive prosodic phrasing. Next, the conversion of the symbolic description of prosody to its acoustic counterpart will be outlined.

Section 3 describes the data structure used in our text-to-speech (TTS) system allowing an easy way of communication between the different modules of the TTS system.

Finally, the last section reports on the results obtained for French speech synthesis and on the conversion of the French system into a multi-lingual prosodic generator.

2. PROSODY GENERATION

The prosody generation system is composed of two main steps:

- an abstract symbolic description of the prosody is first derived from its syntax;
- this symbolic string is then converted to an acoustic description of prosody: phonemic duration and f_0 curve.

The phonetic transcription and the syntactic structure of the text are computed with the morpho-syntactic phonetizer LIPSS (LInguistic Processing for Speech Synthesis) [10]. The synthetic prosody is given, with the sequence of phonemes, to the MBROLA [2] speech synthesizer to allow a qualitative evaluation of the prosody generated.

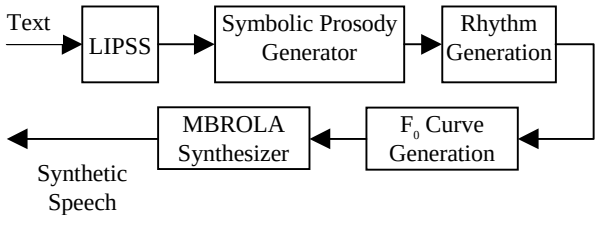

Figure 1 : Text-to-speech synthesizer.

2.1. Symbolic Description of Prosody

The first task of the prosody generation system is to compute the prosodic phrasing of the input text. Its based on a very crude chinks 'n chunks algorithm [11] combining grammatical/content words (See Table 1) detection and rhythmic constrains.

Grammatical tag	Grammatical word (G)	Content word (C)
Verb		X
Infinitive		X
Noun		X
Adjective		X
Determiner	X	
Adverb		X
Pronoun	(X)	(X)
Preposition	X	
Coordinating Conj.	X	
Subordinating Conj.	X	
Other		X

Table 1 : Grammatical/content words classification.

For the French language, the intonation groups (as defined by Mertens [12]) are determined with the rule:

'If a grammatical word follows a content word then the content word is the end of a intonation group and receives a final accent (FA).'

A special final accent is assigned to the last word of the sentence depending of the modality: L-L- for affirmation, H/H for interrogation and H+H+ for exclamation (Tags coming from the tonal model of French intonation described in [12]).

Following this first tagging, rhythmic constraints mainly related to the number of consecutive unaccented syllables in a group are applied to add some final accents (as for the final accent on the word *canard* on Figure 2).

G	C	C	C	G	C	
NA	NA	FA	FA	NA	L-L-	
Le	petit	canard	apprend	à	nager	.
l@	p@--ti	ka--naR	a--pRa~	a	na--Ze	_
NA	NA NA	NA AF	NA AF	NA	NA L-L-	

Figure 2 : Prosodic phrasing stage

Following word accentuation, a syllable accentuation process is performed to assign a final accent on the last syllable of each accented word (this is only valid for French). The other syllables receive a non accented tag (NA).

Figure 2 gives an example of prosodic phrasing for the French utterance *'Le petit canard apprend à nager'* (*'The little duck learns to swim'*).

2.2. From Symbolic to Acoustic Prosody

The speech synthesizer must receive an acoustic description of the prosody, in the form of a sequence of phoneme durations and pitch targets. These values are computed from the abstract description outlined in section 2.1.. The conversion is made using machine learning techniques trained on speech corpora.

The training corpus (1 speaker) is composed of 12 short news bulletins (around 50 seconds per bulletin), 13 long news bulletins (from 3 to 10 minutes per bulletin) and from 40 short sentences (from 1 to 15 words). The total duration is about 70 minutes. The prosody transplantation tool MBROLIGN [9] has been used to obtain automatically the phonetic and syllabic segmentation and the f_0 curve of the speech signals. Prosodic tagging has been obtained from the text by use of the same chinks 'n chunks algorithm as the one described in section 2.1..

Rhythm Generation

Phonetic durations are derived from the symbolic representation with a classification and regression tree [13]. The following criteria, coming from different linguistic levels, have been used to build the tree :

- at the phoneme level: the current phoneme, its phonetic class and its position in the syllable (onset, nucleus or coda);
- at the syllabic level: the syllable type (CV, CVC, …), its accent type (NA, FA, L-L-, …) and its size (in phonemes);
- at the phoneme contextual level : the phonetic class of the following phoneme;
- at the rhythmic level : the position (in syllables) of the last accented syllable.

The CART tree has been trained with WAGON, a tool available with the FESTIVAL Speech Synthesis system of CSTR [14]. The training of the CART on 90 % of the corpus and its testing on the other 10 % give a mean duration prediction error of less than 20 ms.

F0 Curve Generation

To convert the symbolic representation of intonation into an f_0 curve represented by a set of pitch targets, an intonation pattern dictionary is automatically derived from the training corpus described in section 2.2. .

Each intonation pattern in the dictionary is composed of a key and a set of pitch targets (targets are coded in semitones (ST) in regards to a mean pitch value) to apply to the different syllables supporting the pattern. The key is composed of 4 parts:

- the position of the pattern in the sentence;
- the last final accent found (AF) and the mean value of the pitch targets on the syllable supporting the accent (in ST);
- a number of unaccented syllables (NA);
- the final accent of the pattern (AF) and the mean value of the pitch targets on the syllable supporting the accent (in ST).

For example, for an intonation pattern in second position in a sentence with 3 unaccented syllable the key will look like : 2 FANANANAFA 2.5ST 4.0ST. During the training stage, a dictionary of 1838 patterns has been derived from the labeled speech corpus and 5 pitch targets have been assigned to each syllables of each pattern.

At run time, the intonation curve is computed by searching the dictionary for the pattern that is the closest to the one that is required for each intonation group. The look-up strategy is based on a comparison forcing the first and the last accents of the pattern to be equal to those in the target and trying to minimize first the difference between the number of syllables of the researched pattern and those of the dictionary and next the difference between the positions of the patterns in their respective sentences. This look-up procedure ensures that at least one pattern will be found in the dictionary for each pattern to be synthesized. In practice, more than one pattern is often found.

Following the definition of the intonation patterns, two successive patterns overlap respectively on their first and last syllable.

Some pattern concatenation is therefore needed, like in diphone speech synthesis, to obtain the overall f_0 curve. This is where multiple candidate patterns found in the dictionary for each target pattern are put to profit. A dynamic programming algorithm is used to select, among the candidate patterns for each target pattern, those which are supposed to lead to the most natural f_0 curve when concatenated.

The criterion used to evaluate the naturalness of the concatenation of two patterns is only based on the mean value of the pitch targets of the last and the first final accents of two successive patterns and tries to minimize their difference. This criterion must be refined in the future in order to consider not only mean pitch value but also pitch variation and pitch curve concavity.

After the optimal sequence of patterns has been found, the global f_0 curve is built by applying the different pitch targets to the syllables supporting each pattern. At concatenation points, pitch targets are smoothed to allow a good matching between patterns.

This approach of f_0 contours generation recalls in some way the unit selection process [1][15] only applied here to the selection of prosodic units.

3. MULTI LAYER CONTAINER (MLC)

In Text-to-Speech systems, data is most of the time organized into levels (the orthographic level and the phonetic level are two good examples). In order to allow different modules to work together, they must use the same data structure.

The prosodic module described in this paper has been implemented using a data structure that recalls the MLDS (Multi Level Data Structure) described in [16]. This data structure, called Multi Layer Container (MLC) has been specially designed for TTS applications.

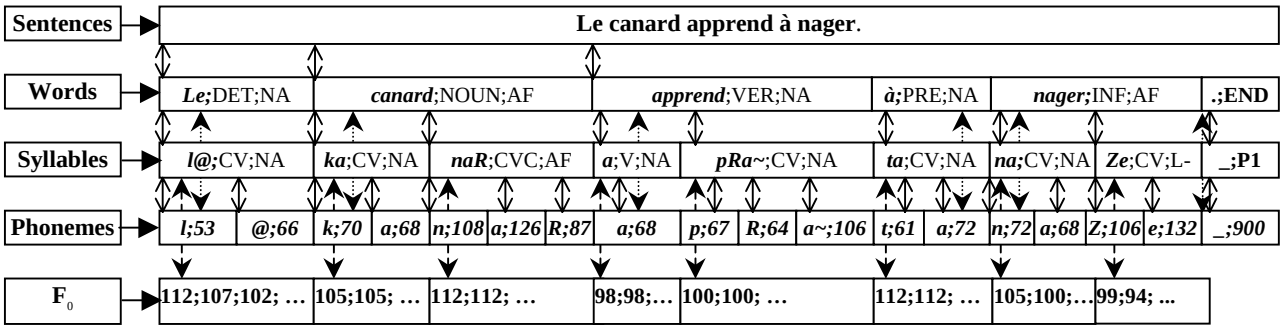

Figure 3 : Example of Multi Layer Container for a sentence.

It can be seen as an extension of the C++ Standard Template Library for multi-layered data objects. Such a structure is represented on Figure 3 for the French sentence *'Le canard apprend à nager'* (*'The duck learns to swim'*).

4. CONCLUSIONS AND PERSPECTIVES

Preliminary tests have been made for isolated sentences and for paragraphs extracted from newspapers. The first results are very promising. An example of text-to-speech synthesis using the prosody generation system described here is given on the CD-ROM [SOUND355_01.wav] and other are available at **http://tcts.fpms.ac.be/synthesis/prosody**.

The main advantages of the approach described in this paper are :

- it can be easily adapted to other languages. Currently, we are working on a English prosody generator using the same technique;
- the system can also be quickly adapted to a special speaker (new training corpus) or to a special intonation style. We are now busy applying the same technique to produce emotional prosody automatically obtained from emotional speech databases.

The following directions are currently being explored to improve the quality of the synthesized prosody:

- replacing the very crude chinks 'n chunks algorithm with a more efficient model like the one described by P. Mertens [12] for French;
- using more contextual information at the tone level in the similarity criterion used during the dynamic programming process;
- improving the criterion used to check the naturalness of pattern concatenation during dynamic programming.

5. ACKNOWLEDGMENTS

The authors are especially grateful to the FRIA (Fonds pour la Formation à la Recherche dans l'Industrie et l'Agriculture) for its financial support and to Michel Bagein and Alain Ruelle for their implementation of the MLC.

6. REFERENCES

1. Campbell N., " Prosody and the Selection of units for Concatenation Synthesis ", Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, pp. 61-64, 1994.
2. Dutoit T., Pagel V., Pierret N., Bataille F. and van der Vrecken O., " The MBROLA Project : Towards a Set of High Quality Speech Synthesizers Free of Use for non commercial purposes ", Proceedings of ICSLP'96, pp. 1393-1396, 1996.
3. K.J. Kohler K.J., " Improving the prosody in German text-to-speech output ", Proceedings of the ESCA Workshop on Speech Synthesis, pp.189-192, 1990.
4. Collier R., " Multi-lingual intonation synthesis: principles and applications ", Proceedings of the ESCA Workshop on Speech Synthesis, pp. 273-276, 1990.
5. Traber C., " F0 Generation with a Database of Natural F0 Patterns and with a Neural Network ", in 'Talking Machines: Theories, Models and Designs', G. Bailly and C. Benoît, Eds. North Holland, pp. 287-304, 1992.
6. Morlec Y., " Génération multiparamétrique de la prosodie du français par apprentissage automatique ", Ph. D. dissertation, Institut de la Communication Parlée, Grenoble, 1997.
7. Black A.W. and Hunt A.J., " Generating F0 contours from ToBI labels using linear regression ", Proceedings of ICSLP'96, p. 1385-1388, 1996.
8. Möbius B., Pätzold M. and Hess W., " Analysis and synthesis of German F0 contours by means of Fujisaki's model ", Speech Communication, pp. 53-61, 1993.
9. Malfrère F. and Dutoit T., " High Quality Speech Synthesis for Phonetic Speech Segmentation ", Proceedings of EuroSpeech'97, pp. 2631-2634, 1997.
10. Dutoit T., " High Quality Text-to-Speech Synthesis for the French Language ", Ph. D. dissertation, Faculté Polytechnique de Mons, 1993.
11. Liberman M.J. and Church K.W., " Text Analysis and Word Pronunciation in Text-to-Speech Synthesis ", in 'Advances in Speech Signal Signal Processing', S. Furui and M.M. Sondhi, Eds. Dekker, pp. 791-831, 1992.
12. Mertens P., " L'intonation du français. De la description linguistique à la reconnaissance automatique ", Thèse de Doctorat, Katholieke Universiteit Leuven, 1987.
13. Brieman L., Friedman J.H., Olshen R.A. and Stone C.J., " Classification and regression trees ", Wadsworth & Brooks Press, 1984.
14. Black A.W., Taylor P. and Caley R., " The Festival Speech Synthesis System : System Documentation ", University of Edinburgh, 1997.
15. Hunt A.J. and Black A.W., " Unit Selection in a Concatenative Speech Synthesis System using Large Speech Database ", Proceedings of ICASSP'96, pp. 373-376, 1996.
16. van Leeuwen H.C. and te Lindert E., " Speech Maker: a flexible and general framework for text-to-speech synthesis, and its application to Dutch ", Computer, Speech and Language, pp. 149-167, 1993.