A REAL-TIME FRENCH TEXT-TO-SPEECH SYSTEM GENERATING HIGH-QUALITY SYNTHETIC SPEECH

*E. Moulines, F. Emerard, D. Larreur, J.L. Le Saint Milon, L. Le Faucheur, F. Marty *, F. Charpentier, C. Sorin*

CNET LAA/TSS/RCP 22301 Lannion (France)
* French and Computer-based Education Research Laboratory, University of Illinois, Urbana, Illinois 61801 (USA)

ABSTRACT

This paper describes the main features of the new CNET diphone-based text-to-speech system for French language. The linguistic analysis works in three steps. First, a morphosyntactic analysis module assigns a grammatical value to each word in the text and transcribes it phonetically. A second module parses the text into hierarchical "syntactico-prosodic" groups. Finally, prosodic patterns are automatically assigned to each word by queries to a data base of prosodic events. The phonetic and prosodic informations serves as commands to the synthesis component. The synthesis component is based on diphone concatenation. A time-domain formulation of the pitch-synchronous overlap-add scheme (TD-PSOLA) is used to modify the speech prosody and to concatenate diphone waveforms. It is combined with a low bit-rate speech decoder to reduce the memory requirement for storing the diphone inventory.

The system runs in real-time on a PC equipped with a TMS320C25 DSP board and provides notably improved sound quality and naturalness in comparison to commercially available systems.

INTRODUCTION

To make synthetic speech a close imitation of natural speech, i.e. making it understandable without undue effort, requires to work at both the signal processing level (to obtain a natural voice timbre) and the linguistico-prosodic level (to attain a good pronunciation of any words and a natural, non monotonous intonation).

The first part of this paper presents the main characteristics of the new linguistico-prosodic module CNETVOX90 developed for text-to-speech (TTS) synthesis of French. This module includes several sets of rules and a large lexicon of words and idiomatic expressions.

The synthesis component is described in a second part. It combines a low bit rate speech coding algorithm (to reduce the memory requirement for storing the diphone data base) and a time-domain pitch-synchronous overlap-add (TD-PSOLA) to modify speech prosody and to concatenate diphone units [1-3]. Some implementation constraints are reviewed.

1 LINGUISTICO-PROSODIC MODULE

The linguistico-prosodic module of the previous CNET TTS system for French [4-5] had several limitations, principally due to the weakness of the extracted linguistic information: heterophonous homographs ambiguities (e.g. "fils") were not solved; liaisons and schwas were not well processed; sentence prosodic parsing, based mainly on rhythmic criteria, leads to some mislocation of prosodic group boundaries, interfering with the comprehension of the message, etc...

In order to solve these problems in an efficient way, a complete linguistic analysis was included in the high-level processing module, which was then redesigned to maximally exploit, at the phonetic and prosodic levels, any reliable grammatical and syntactic information [6].

1.1 Linguistic processing

The present linguistic processing combines the linguistic analysis with the phonetic transcription and parses the input text into hierarchical "syntactico-prosodic" groups.

Linguistic analysis and orthographic-phonetic transcription:

At this stage, the objective is to transcribe each word and to assign it to a grammar category. To achieve this goal, the program uses [7]:

- 1 table with 229 roots,
- 1 table with 65 prefixes,
- 1 table with 174 suffixes,
- 553 transcription rules,
- 11 transformation rules,
- a lexicon containing 1,655 idiomatic expressions and 13,118 individual words (including common abbreviations and acronyms) and their phonetic transcription.

The grapheme-to-phoneme transcription is provided either by a direct look-up into the lexicon or by the application of the transcription rules. At the end of this process, over 99 % of the words are definitively transcribed; the remaining words are heterophonous homographs which are assigned temporary transcriptions by the lexicon.

The grammatical labelling module uses a set of 255 categories. The words which are found in the lexicon and those whose suffix occurs in the suffix table are assigned to one of the 255 categories; this assignment is temporary if the word belongs to several grammar categories (e.g. "les, court, bien, son, entre, fait") . At this stage, about 58 % of words are assigned to a definite category, about 39 % are assigned to a temporary category and about 3 % remain unassigned. Therefore, the linguistic analysis needs to be continued for only about 42 % of the words. This analysis uses 1,112 rules which examine the syntactic environment (up to five slots to the left and four slots to the right). Upon completion of this linguistic analysis, each word has been assigned an unambiguous grammar category and a single phonetic transcription.

Syntactico-prosodic parsing:

The task of the parsing module is to establish syntactico-prosodic boundaries within each sentence and to define their nature and relative strength in order to determine subsequent prosodic processing. The parsing is based on the assumption that, in French, a prosodic boundary can be derived, in most cases, from

CH2847-2/90/0000-0309 $1.00 © 1990 IEEE

the grammar category of the word and its left and right neighbours. To detect these boundaries, 140 parsing rules are used. Each rule is composed of a sequence of grammar category sets, each set consisting in one or several categories (e.g. the first part of the negation or all nouns + all adjectives + all verbs).

Parsing is carried out in the following way:

- The program scans each word and determines whether the grammar categories of this word and of the surrounding words match the grammar categories within the sets of any of the 140 rules. If a match is found, a prosodic boundary is set and the corresponding parsing rule number is inserted after the boundary word.

- Once all the words of the sentence have been examined, a hierarchical analysis is carried out to decide whether the last boundary word shall be followed by a pause or by a simple lengthening of the final syllable. This decision depends either on the syntactic-prosodic value assigned to the boundary (for example, an obligatory pause is associated with punctuation marks and subordinate clause conjunctions) or on the number of syllables of the prosodic group in which the boundary appears.

This parsing module is very reliable: over 95 % of the prosodic boundaries are correctly detected.

1.2 Prosodic processing

At this stage, the input text has been changed into a sequence of phonetically transcribed words, some of which are followed by a syntactico-prosodic parsing rule number. Since a given type of boundary can be yielded by several rules (for example, six rules set up the same boundary between a pre-verbal and a verbal syntagm), it is possible to cluster the 140 parsing rules into a set of 30 prosodic markers. 9 prosodic markers are added for the pauses and 2 markers for the words which are not located at a parsing boundary (one for the function words and the other for the lexical words). Once each word has been assigned a prosodic marker, the prosodic processing module operates as follows:

- each marker is revised after an arborescent analysis of the right context [8]

- a pitch pattern and a segmental duration corresponding to the prosodic marker is assigned to each word; if there is a pause marker, the duration of the pause is also determined. These prosodic informations are contained in tables elaborated after an exhaustive analysis of a prosodic data-base [6] [9].

The arborescence which determines the duration of the segments is close to the one proposed in [10], but it emphasizes the notion of syllable. Thus, the duration of a segment depends on:

. the prosodic marker and any potential pause marker which govern the word,

. the position, in the word, of the syllable to which the segment belongs,

. its position in the syllable,

. its phonetic environment.

A table provides fundamental frequency patterns adapted to all word lengths and all prosodic markers. Since this melodic table handles only the vowel phonemes, a micromelody is added to the voiced consonants. Moreover, for each marker and for each word length, several melodic contours, derived from the analysis of the prosodic data-base, are now available: therefore, some melodic diversity may be introduced in synthesizing longer texts.

At this stage, the input text has been changed into a list of phonemes with their requested durations and pitch contours (specified by 3 pitch values). This representation serves as a command for the synthesizer.

2 SYNTHESIS MODULE

In our system, synthetic speech is obtained by concatenation of acoustic units [4]. Our diphone inventory includes 1500 units (1200 diphones and 300 context-dependent allodiphones). Such an inventory corresponds roughly to 3 minutes of speech and requires more than 5 Mbytes of memory (at 16 kHz sampling frequency). Low bit-rate coding of this inventory is therefore, most of the time, a pre-requisite to meet industrial requirements.

Therefore, in case of a preliminary off-line coding of the diphone inventory, the synthesis algorithm must involve two different processes:

- a decoding process: the waveform of the acoustic units is reconstructed from the low bit-rate coded informations,

- a prosodic modification and concatenation process: the sequence of acoustic units is concatenated after an appropriate modification of their intrinsic prosody to match the intonation contour and the phonemic duration specified by the prosodic generation module.

In our current TMS320C25 implementation, these two processes are intimately interlaced in order to maximally exploit the common structure of the decoding and synthesis scheme. We present here only the main features of the low bit-rate coding algorithm and recall the salient characteristics of the TD-PSOLA synthesis scheme used for the prosodic modification and concatenation operation. Some details are given on the currently implemented TMS version of the TD-PSOLA algorithm which precise the constraints to be satisfied to assure an optimal quality of the synthetic speech.

2.1 Acoustic units inventory coding

A low-bit rate predictive coder has been developed to encode the acoustic units inventory [11]. The coder presents a two stages prediction: a first prediction based on the short-time spectral envelope of speech, and a second long-term prediction based on the periodic nature of the spectral fine structure. In the current implementation, multipulse excitation has been chosen to encode the innovation (residual signal) [12].

As the coding process is performed off-line, our goal was to optimize the coder efficiency without trying to reduce its computational cost. Pitch-synchronous weighted covariance analysis has been chosen to optimize the short-term predictor estimation (the pitch-marks having been automatically set by the algorithm described in 2.3.2) Closed-loop estimation of the long-term predictor has been retained to improve the speech quality.

The overall bit-rate is variable as LPC frame update is adapted to the local speech characteristics. Split Vector Quantization is applied to encode LPC parameters. Efficient coding procedure of the excitation (including vector quantization of pulse amplitudes) has been used in order to keep the bit-rate low.

Nearly transparent coder operation is performed for an average bit-rate of 30 kbps at 16 kHz sampling frequency (16 kbps/sec at 8 kHz). The memory requirement to store the diphone inventory is therefore reduced to 700 kbytes at 16 kHz (400 kbytes at 8 kHz) including the side informations needed to retrieve pitch-marks (cf 2.3.2).

2.2 TD-PSOLA synthesis principles

The TD-PSOLA synthesis scheme involves the three following steps: an analysis of the original speech waveform to produce an intermediate non-parametric representation of the signal, prosodic modifications brought to this intermediate representation, and finally the synthesis of the modified signal from the modified intermediate representation [1-3]. We briefly review here the salient features of this algorithm.

The waveform $x(n)$ is decomposed into a sequence of short-term overlapping signal $x_m(n)$. These ST-signals are obtained by multiplying the signal by a sequence of analysis window $h_n(m)$:

$$x_m(n) = h_m(t_m - n)x(n)$$

The successive instants t_m, called pitch-marks, are set at a pitch-synchronous rate on the voiced portions of the signal.

The stream of analysis ST-signals is processed to produce a stream of modified synthesis ST-signals $\hat{x}_q(n)$, synchronized on a new set of synthesis pitch-marks t_q. The algorithm determines simultaneously the synthesis pitch-marks according to the pitch-scale and time-scale modifications factors, and the mapping $t_q \rightarrow t_{\phi(q)} = t_n$ between the synthesis and analysis pitch-marks. This mapping specifies which analysis ST-signal is to be copied to obtain any given synthesis ST-signal and the stream of synthesis pitch-marks indicates the delay to be used between the synthesis ST-signals.

$$\hat{x}_q(m) = x_{\phi(q)}(m + t_{\phi(q)} - t_q) = h_{\phi(q)}(t_q - m)x(m + t_{\phi(q)} - t_q)$$

Generally, this mapping is not one-to-one and results in either a duplication or an elimination of the analysis ST-signals. The synthetic speech is obtained by overlap-adding the stream of synthesis ST-signals [13-15]:

$$\hat{x}(n) = \frac{\sum_q \alpha_q \hat{x}_q(n) h_q(t_q - n)}{\sum_q h_q^2(t_q - n)}$$

The additional normalization factor α_q is introduced to correct the energy variations due to the pitch modification procedure.

2.3 Constraints for real-time high-quality implementation

2.3.1 Properties of the TD-PSOLA pitch modification scheme

For convenience and without loss of generality, we use a model of stationary voiced sound consisting in the superposition of a deterministic periodic signal with a wide sense stationary stochastic signal (WSSS). The deterministic signal takes into consideration the harmonic component of the speech signal which must be affected by TD-PSOLA modification. The stochastic signal models the irregularities which can be observed on the speech signal from one period to another. In order to avoid artefacts, this component should not be affected in a significant way by TD-PSOLA modifications. We assume that:

(1) the deterministic signal is periodic with period P. The stream of analysis pitch-marks are set at a pitch-synchronous rate so that: $t_n = nP$

(2) the pitch modification factor is a constant β and is equal to the time scale modification factor. There is a one-to-one mapping between analysis and synthesis pitch-marks so that: $t_n = n\beta P$

(3) the analysis windows are equal $h_m^2(n) = h(m)$ and verify the normalization condition $\sum h(t_q - m) = 1$.

Modifications of the deterministic component

With these assumptions, it can be shown that:

$$\hat{x}(m) = \sum_s w(s\beta P, m) \quad w(n,m) = h(n-m)x(m-n)$$

The analysis window being of finite duration, partial Fourier transform of $w(n,m)$ can be defined with respect to each argument (respectively $W_1(\psi, m)$ and $W_2(n, \omega)$) [21]. Poisson formula, which relates the sum of the signal on a periodic network to the sum of its Fourier transform on the reciprocal network, leaves us with:

$$\hat{x}(m) = 1/(\beta P) \sum_{k=0}^{\beta P - 1} W_1(2\pi k/\beta P, m)$$

$$= 1/(\beta P) \sum_{k=0}^{\beta P - 1} W_2(0, 2\pi k/\beta P) \exp(j(2\pi k/\beta P)m)$$

The synthetic signal is periodic with period βP. The complex harmonics amplitudes are equal to the Fourier transform of the prototype short-time signal $h(-m)x(m)$ sampled at the synthetic pitch-harmonic frequencies

Modifications of the stochastic component

We study the TD-PSOLA modification of the WSSS $x(m)$. Denoting $S_x(\omega)$ its power spectral density (PSD), it can be shown [11] that the synthetic signal PSD is equal to:

$$S_{\hat{x}}(\omega) = 1/(\beta P^2) \sum_{k=0}^{P-1} |H(\omega(1-\beta) + 2\pi k/P)|^2 S_x(\beta\omega - 2\pi k/P)$$

This formula displays the role of the Fourier transform of the analysis window which acts like the transfer function in WSSS filtering. It can be noted that the synthesis signal PSD for a given normalized frequency doesn't depend only on the analysis signal PSD at the same frequency. This emphasizes the non-linear behaviour of the TD-PSOLA tranformation.

This analysis of the TD-PSOLA algorithm defines the set of constraints that have to be satisfied in order to maximize the quality of the synthetic speech. They concern mainly the synchronization of the analysis pitch-marks and the choice of the analysis window.

2.3.2 Synchronization of the analysis pitch-marks

To assure an optimal quality of the synthetic speech, the TD-PSOLA analysis window must be rigourously synchronized with the main excitation instants within each pitch period, namely the instants of glottal closure: incorrect synchronization affects the synthetic signal harmonic phase distribution and results in a distortion of formant amplitudes [16].

A "pitch-marking" algorithm was developped, based on the detection of abrupt changes in the short-time spectral characteristics of the speech signal [17].

This algorithm assumes that the speech waveform, at each pitch period, can be represented by the concatenation of the responses of two all-pole systems, which alternate with closed and open glottis intervals. To determine the instant of transition between the models, we use a sequential maximum likelihood method: for each possible transition time r, two consecutive models (until r and starting at r respectively) are identified. A single model is simultaneously identified under the hypothesis of no change. The first derivative of the likelihood ratio between the alternatives "one change at r" and "no change" is computed and provides a test signal. The detection of glottal closure instant is then performed by a peak-picking algorithm based on dynamic programming, the search path being constrained by the local pitch period.

Provided high-quality recording, this algorithm leads to correct detection of glottal closure instants more than 95 % of the time. The remaining errors (mainly located at the boundaries of voiced segments) are corrected manually. The stream of pitch-marks is set on the acoustic units inventory before its offline coding (see 2.1).

2.3.3 Window selection

As shown above, the spectral resolution of the TD-PSOLA window greatly affects the quality of the synthetic signal. The spectral resolution is conditioned by two factors: the duration and the type of this analysis window.

window duration

Narrow-band analysis condition (the bandwidth of the analysis window being inferior to the instantaneous Fo) has been shown to be inappropriate for TD-PSOLA manipulation [3] [16]. Thus we concentrated only on wide-band (WB) condition, under which the bandwidth of the analysis window is chosen to be (several time) greater than Fo. For classical windows (Hanning, Blackman), this condition is fulfilled as soon as the window duration is less than twice the local pitch period.

Under WB analysis condition, the amplitude spectrum of each analysis ST-signal is a "smoothed" estimate of the true power spectrum, the major discrepancy lying in the bandwidth of formant resonances. As the bandwidth of formant is usually much less than the bandwidth of the analysis window, the shape of the synthetic signal formants depends primarily on the window main lobe. The mismatch is particularily evident for female voices, since the window must be shorter to respect the WB analysis conditions. However, as just noticeable difference limens measured for formant bandwidth are about 40 % for steady vowel (and even larger for continuous speech) [19], the formant bandwith mismatch is generally not perceived.

For a real time TMS implementation, it is also desirable to avoid the division by the factor which compensate for the variable overlap of the windows at the synthesis stage. For that reason, we enforce the energy normalization relation:

$$\sum_q h_q^2(t_q - m) = 1 \quad \forall m$$

which in return constraints the window length.

When TD-PSOLA is used to raise the pitch, these two constraints are simultaneously fulfilled by chosing the window duration to be twice the synthetic pitch period. The same choice can be made when the pitch has to be lowered, at least for moderate pitch-scaling factors: the normalization constraint is respected, but WB analysis condition is no more strictly enforced. For larger pitch-scaling factors (> 1.3), this choice is no more acceptable, as the window spectral resolution deviates clearly from WB-conditions. It appears to be better to choose a window length = $2P$ (P being the local pitch period): the normalization cannot be avoided. In our implementation, we have retained a slighlty different approach: we permute, in that case, LP-filtering and TD-PSOLA operations [1][3] (window length = $2P$, no normalization). This solution allows F_0 to be decreased by a much greater proportion before any degradation becomes apparent.

window type

To avoid spectral distortions, the TD-PSOLA window should check the properties commonly required for spectral analysis window [20]. As mentioned before, the TD-PSOLA window duration is to be adapted to each new synthesis pitch period. For real-time application, the computational burden involved by this operation should be kept as low as possible.

The window is therefore constructed by concatenating two tabulated half-cosine lobes (of width $(\alpha/2)N$) with a short segment of rectangular window (of width $(1.0-\alpha/2)N$). This family of windows exhibits a confusing array of sidelobe levels arising from the product of the two component transforms. However, satisfactory window is obtained when α is less than 0.2.

It appears that less than 20 tabulated half-cosine lobes are sufficient to generate all the windows required by the TD-PSOLA processing.

3 CONCLUSION

The diphone-based TTS system described in this paper includes several improvements at both the linguistico-prosodic and the signal processing levels.

The introduction of a reliable linguistic analysis and the exploitation of a large data-base of prosodic contours allows to generate a very natural, non-monotonous synthetic speech.

The TD-PSOLA algorithm appears to be an efficient method to modify the prosodic parameters, while preserving most of the naturalness of the voice timbre. Moreover, it can be combined in a flexible manner with low-bit rate speech coding in order to reduce the memory requirements to store the diphones inventory.

The complete system runs in real-time on a PC equipped with a TMS320C25 based signal processing board.

REFERENCES

[1] E. Moulines, F. Charpentier, "Diphone synthesis using multipulse linear predictive coding", Proc. FASE, Edingburgh, 1988

[2] C. Hamon, E. Moulines, F. Charpentier, "A diphone synthesis system based on time-domain modifications of speech", Proc. ICASSP Glasgow, 1989

[3] F. Charpentier, E. Moulines, "Pitch-synchronous waveform processing techniques for text-to-speech using diphones", Proc. Eurospeech, Paris, 1989

[4] F. Emerard, "Synthèse par diphones et traitement de la prosodie", Thèse, Univ. Grenoble, 1987

[5] C. Sorin, D. Larreur, R. LLorca, "A rhythm-based prosodic parser for text-to-speech systems in French", Proc. ICPS, Tallin, 1987

[6] D. Larreur, F. Emerard, F. Marty, "Linguistic and prosodic processing for a text-to-speech synthesis system", Proc. Eurospeech, Paris, 1989

[8] L. Danlos, F. Emerard, E. Laporte, "Synthesis of spoken messages from semantic representation", Coling, Bonn, 1986

[7] F. Marty, R. Hart, "Computer program to transcribe French Text into Speech: Problems and suggested solutions", Tech. Report No LLL-T-6-85. Language Learning Laboratory, University of Illinois, Urbana, 1985

[9] F. Emerard, C. Benoit, "Base de données prosodiques pour la synthèse de parole", Journal d'Acoustique, 1988

[10] K. Bartkova, C. Sorin, "A model of segmental duration for speech synthesis in French", Speech Communication, vol 6, 245-260, 1987

[11] E. Moulines, "Contributions à la synthèse de parole à partir du texte", Thèse ENST, Paris, 1990

[12] H. Woo, J. Gibson, "Multipulse-based codebooks for CELP coding at 7 kbps", Proc. ICASSP, Glasgow, 1989

[13] R. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis", IEEE Trans. On ASSP, vol 28(2), 1980

[14] D. Griffin, J. Lim, "Signal estimation from modified short time Fourier transform", I.E.E.E. Trans. on ASSP, vol 32(2), 1984

[15] F. Charpentier, "Traitement de la parole par analyse-synthèse de Fourier à court-terme: application à la synthèse par diphones", Thèse ENST, Paris, 1988

[16] E. Moulines, C. Hamon, F. Charpentier, "High-quality prosodic modifications of speech using time-domain overlap-add synthesis", Proc. XII ième colloque GRETSI, 1989

[17] R. Di Francesco, E. Moulines, "Detection of the glottal closure by jumps in the statistical properties of the signal", Proc. Eurospeech, 1989

[18] M. Portnoff, "Short-time Fourier analysis of sampled speech", IEEE Trans. On ASSP, vol 39(3), 1981

[19] O. Ghitza, J.L Goldstein, "Scalar LPC quantization based on formant JND's", IEEE Trans. on ASSP, vol 34(8), 1986

[20] J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transfom", Proc. Of IEEE, vol 66(1), 1978

[21] M. Portnoff, "Time-Frequency representation of digital signals and systems based on short-time Fourier analysis", IEEE Trans. on ASSP, vol 28(1), 1980