LINGUISTIC AND PROSODIC PROCESSING FOR A TEXT-TO-SPEECH SYNTHESIS SYSTEM.

Larreur, D.*, Emerard, F.* and Marty, F.**

* Centre National d’Etudes des Télécommunications
BP 40 22301 - Lannion - FRANCE
** French and Computer-based Education Research Laboratory
University of Illinois, Urbana- Illinois- 61801 - U.S.A.

ABSTRACT

The CNET’s commercially available text-to-speech system performs an automatic prosodic parsing, based on the detection of a small number of grammatical words. However, the lack of any other syntactic information is a serious source of errors at the supra-segmental level. In this paper, we present a new prosodic parsing algorithm which overcomes that drawback. First, a recursive left-to-right morphosyntactic analysis assigns a grammatical value to each word in the text and transcribes it phonetically. Second, one hundred and forty hierarchized parsing rules divide the message into a sequence of prosodic groups. Finally, prosodic patterns are automatically assigned to each word by queries to a data base of prosodic events. Preliminary tests indicate that such an approach, combining a linguistic processor and a data base of real prosodic features, yields synthetic speech with a high degree of naturalness.

1. Introduction

Most text-to-speech systems available today include three main components:
- a linguistic component which attempts to analyze the written input,
- a prosodic component which attempts to produce correct intonative and rhythmic patterns,
- an acoustic component for speech signal output.

Current research aims to make synthetic speech a close imitation of pleasing natural speech, thereby making it understandable without undue effort. After a brief review of the CNET’s previous text-to-speech system and of the progress recently obtained in signal processing, this paper will describe an analysis and parsing module which considerably enhances the prosodic output.

2. The CNET’s previous TTS system

The text-to-speech system known commercially as "MULTIVOC" can synthesize any text which is correctly spelled and uses standard diacritics [1].

2.1. Linguistic and prosodic processing

The text-preprocessing and the orthographic-phonetic transcription [2] are executed with about 400 context-dependent transcription rules, and include specific modules for the pronunciation of numbers, abbreviations and acronyms; heterophonic homographs (e.g., "fils") are not processed. The prosodic parsing of the input text [3] does not use a syntactic analysis processor; instead, it relies on a lexicon of 120 frequent grammatical words. The division of a sentence into prosodic groups takes place at the punctuation marks and at the boundaries between a lexical word and a grammatical word. Each prosodic group is assigned a prosodic category depending on the number of lexical words in the group and on the position of the group in the sentence. These categories define, for each group, the potential position of the pauses and generate "prosodic markers" which govern the duration and the Fo contour of each word.

The duration processing, adapted to each dictionary of acoustic segments chosen for the synthesis, consists in modifying the intrinsic duration of the phones. The melodic processing allows the Fo to produce the traditional "reading" intonation [4] or the intonation referred to as "commercial" [3]. Both versions use a set of specific prosodic markers which refer to corresponding pitch patterns for words or groups of words.

2.2. Acoustic processing

Based on the conventional diphone concatenation method, the Multivoc system can use two dictionaries of acoustic elements: one for a male voice and one for a female voice. Each dictionary contains 1200 diphones. The linguistic, prosodic, and concatenation modules of Multivoc are now independent of the synthesis technique and, therefore, can also function with various signal processing algorithms (as well LPC as the recently developed PSOLA).

3. Recent advances in signal processing.

Three time-domain algorithms using the pitch-synchronous overlap-add (PSOLA) synthesis scheme have been proposed recently. These algorithms are known as PSOLA-MLP [5], PSOLA-FFT [6] and PSOLA-KDG [7]. When combined with a coding algorithm used to reduce the size of the diphone dictionary, PSOLA-KDG creates a system which is comparable in cost to that of conventional LPC synthesis. These three algorithms have been compared with the LPC technique in a pair comparison test [8]. The results show that these techniques (FFT, MPLPC, KDG) were found superior to LPC by 85% of the subjects against 9% who preferred LPC.

4. A new linguistic and prosodic module.

Sentence parsing based entirely on rhythmic criteria as in the Multivoc system has serious limitations; errors in determining the boundaries of the prosodic groups and their relative importance may interfere with the comprehension of the message. In order to solve these problems reliably and generate a coherent prosodic processing, a new sentence prosodic parsing has been developed, based on a linguistic processor.

4.1. Linguistic processing

The present linguistic processing [9] combines the linguistic analysis with the phonetic transcription, and parses the input text into hierarchized "syntactic-prosodic" groups.
4.1.1. Linguistic analysis and orthographic-phonetic transcription.

It is assumed that the input text is correctly spelled and uses standard diacritics; the analyzer accepts all the characters available on a computer keyboard. 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:
- 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).

There are two broad categories of words in the lexicon: 1) those which would not be transcribed correctly by the 553 transcription rules (e.g., "abdomen, amer, portions, est, Caen, Copenhague") and 2) those which by their frequency and grammar category facilitate the linguistic analysis; nearly half the lexicon items are verbs (e.g., "veut") or potential verbs (e.g., "ferme") since these items are more productive in the linguistic analysis than nouns or adjectives.

The grapheme-to-phoneme transcription is provided either by a direct look-up into the lexicon or by the transcription rules. The linguistic analysis 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 (e.g., "les, court, bien, son, ensemble, fait") belongs to several grammar categories. 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: four slots to the left and four slots to the right to resolve the grammatical ambiguities posed by homophonous homographs, five slots to the left and four slots to the right (plus a semantic analysis module when necessary) to resolve the grammatical ambiguities posed by heterophonous homographs and choose the correct transcription. Upon completion of the linguistic analysis, each word has been assigned an unambiguous grammar category and a single phonetic transcription.

4.1.2. Syntactic-prosodic parsing.

The task of the parsing module is to establish syntactic-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 a prosodic boundary, admitted in French, can be derived, in most cases, from the grammar category of the word and its left and right neighbours. To achieve this goal, 140 parsing rules are used:
- Each rule is composed of a sequence of grammar category sets; each set may consist of one (e.g., the first part of the negation) or of several categories (e.g., all nouns plus all adjectives plus allverbs). In our system, each set refers to a number; for example, rule 60 is read as:

\[
559(j) \quad 536 \quad 571 \quad 054
\]

which means that a prosodic boundary is set after the word (j) if word j belongs to set 559 (a sub-group of nouns or adjectives); and if word j+1 belongs to set 536 (a sub-group of determiners); and if word j+2 belongs to set 571 (a sub-group of nouns, acronyms, numbers or names); and if word j+3 belongs to set 054 (any category except punctuations).

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 parsing rule number is inserted after the boundary word.

For example, the sentence: "Il écoute docilement les consignes de ses experts en communication et en publicité et il écoute avec le même sérieux les avis de sa fille cadette." will be processed in the following way:

I- The input text is changed by the linguistic module into:

\[
\text{IL } \text{g78 } \text{EI} \text{"KOU'T } g5 \text{ DOSILM} \text{AN } g68 \text{ LEI } g42 \text{ KON'SINY } g10 \text{ DEU' } g56 \text{ SEI' } g42 \text{ ZL} \text{"AI'KS } \text{KP } \text{SAI'RI'G } g10 \text{ AN' } g46 \text{ KOMUNIKASYON'} \text{g12 } \text{EI' } g105 \text{AN' } g46 \text{PUBLISTEST' } g12 \text{ EI' } g105 \text{IL } g78 \text{EI} \text{"KOU'T } g5 \text{ AVAI'K } g63 \text{ LEI' } g39 \text{ MAI'M } g24 \text{ SEI' } \text{REYE' } g9 \text{ LEI' } g42 \text{ ZL} \text{"AVI' } g10 \text{ DEU' } g56 \text{ SA } g30 \text{ FIY } g12 \text{ KADAI'T } g24.
\]

in which:

\[
g(9,10,12) = \text{noun}
g(18,24) = \text{adjective}
g(25,43) = \text{determiner}
g(1,7..15..17) = \text{verb}
g(78) = \text{subject pronoun}
g(68..73) = \text{adverb}
g(46,50..66) = \text{preposition}
g(105) = \text{coordinating conjunction}
\]

II- In the next step the 140 parsing rules are applied to each word of the transcribed text and three matches are found: rule 23 after the word "experts", rule 17 after the word "publicités", and rule 60 after the word "sérieux". The output text becomes:

\[
\text{IL } \text{EI'KOU'T } \text{DOSILMAN' } \text{LEI' } \text{KON'SINY } \text{DEU' } \text{SEI'} \text{ZL} \text{"AI'KS } \text{KP } \text{SAI'RI'G } /\text{23/ } \text{AN' } \text{KOMUNIKASYON'} \text{EI' } \text{AN'PUBLISTEST'/17/ } \text{EI' } \text{EI'KOU'T } \text{AVAI'K } \text{LEI' } \text{MAI'M } \text{SEI' } \text{REYE' } /\text{60/ } \text{LEI' } \text{ZL} \text{"AVI' } \text{DEU' } \text{SA } \text{FIY } \text{KADAI'T' }.
\]

III- Once all the words of the sentence have been examined, a hierarchical analysis is carried out on each detected word 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. A pause is introduced only to separate prosodic groups containing more than 20 syllables.

The parsing module is very reliable (over 95% of the boundaries are correctly detected), but two kinds of problems may occur:

(a) The linguistic analysis may assign the wrong grammar category to a word. This, for example, may happen when a word which can be an adjective or a past participle occurs between a noun and a preposition. In the following example, only the situational context determines the grammar category of "couronnees":

EUROSPEECH '89, Paris, France, September 1989 1511
A wrong analysis would lead to a wrong prosodic parsing since the boundaries are in different places:
- boundaries before the past participle:
  ..."à Wimbledon, parmi les têtes couronnées dans les tribunes, on reconnaissait..."
- boundaries after the adjective:
  une marche forcée// pour habituer les combattants...
(b). The input text is incorrectly spelled or punctuated; this leads to various kinds of errors:
- incorrect spelling may express a meaning other than the intended one. For example, using "prévenants" instead of "prévenant" changes the meaning, the analysis, and the prosodic parsing:
  ...les évoyés// prévenant le président... (the employee// warning the president)
  ...les employés prévenants// le président... (the considerate employees// preside it)
- since the misspelled words are not recognized by the lexicon or the tables, the linguistic processor may not have enough information to execute a correct analysis.
- the punctuation usage is not well respected. If the comma is omitted in the sentence:
  "Las d'attendre, le joueur double la mise."
the needed prosodic boundary is not set and the meaning is changed:
Las d'attendre le joueur// double la mise.

4.1.3. Processing for liaisons and schwas.
Once the prosodic boundaries are set, the rules for the liaisons and the schwas can be activated.
A set of 67 rules governs the addition of the liaison consonants. These rules operate only when, in a given syntactic context, a final consonant grapheme from the set (d,g,n,p,r,s,t,x,z) is followed by a vowel or semivowel grapheme. In order to improve the intelligibility of synthetic speech, the schwa rules represent a rather formal delivery; the schwa is always retained in the nine monosyllabic function words (e.g., "me") and when it occurs in the initial syllable of polysyllabic words (e.g., "regardez").

4.2. 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 syntactic-prosodic parsing rule number. Since a particular 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 situated 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 [10],
- 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. This information is contained in tables elaborated after an exhaustive analysis of the prosodic data base [11] described below.

4.2.1. Prosodic data base.
The research which led to the elaboration of this data base focused on the influence which the linguistic environment has on the prosody.
This data base, similar in contents and objectives to the one established for Swedish [12], was compiled from a corpus of 150 enunciative sentences read by two French speakers, one male and one female; The sentences constitute a well-balanced sample of varied syntactic constructions. For each speaker, this 8-minute recording contains over 1,500 words of different lengths and syllabic structures, corresponding to about 2,300 syllables, 6,000 labeled acoustic segments, and 200 pauses. This corpus contains all the phonemes of French and their most frequent contexts.
As shown in Figure 1, a duration is assigned to each segment (phoneme or pause) and each vowel is described by three fundamental frequency values (initial, medial, final). In particular, the following linguistic informations are noticeable:
- the punctuation usage is not well respected. If the comma is omitted in the sentence:
  "Las d'attendre, le joueur double la mise."
the needed prosodic boundary is not set and the meaning is changed:
Las d'attendre le joueur// double la mise.
- Pitch patterns. A melodic table provides frequency patterns adapted to all word lengths and all prosodic markers. Since this melodic table handles only the vowel phonemes, a micromelody must be added to the voiced consonants. This is done by interpolating the final frequency value of the last vowel on the left with the initial frequency value of the first vowel on the right; when the voiced consonant is the initial phoneme of the breath-group, the initial frequency value of the vowel on the right is used as the initial frequency value of the initial consonant. Moreover, for each marker and for each word of a given length, several melodic contours are available; therefore, some melodic diversity may be introduced in synthesizing longer texts.
Column 1: Line number in data base.
Column 2: Duration of current phoneme (in ms).
Column 3: Phoneme left of current phoneme, current phoneme, phoneme right of current phoneme (2 characters; * = V.O.T. of voiceless plosives).
Column 4: Description of current syllable (open or closed).
Column 5: Nature of current word (lexical or grammatical).
Column 6: Position of current syllable in word.
Column 7: Position of current phoneme in syllable.
Column 8: Description of phonemes: left, current, right.
Column 9: Syntactic-prosodic code for current word. For example, 10 refers to a word in final position of pre-verbal syntagm, realized without pause and followed on the right by an end-of-sentence prosodic marker.
Column 10: Initial, medial, final Po of current vowel.
Column 11: Melodic contour of current vowel.
Column 12: Duration of current syllable (in ms).
Column 13: Phonetic coding of current syllable.

CONCLUSION.

The progress recently made in signal processing cannot be fully exploited in text-to-speech synthesis unless parallel progress is made in the linguistic analysis which is indispensable to generate high-quality prosody.

The options described in this paper have resulted in significant progress in four areas:

- a reliable linguistic analysis using less than 640 kO of memory and running in real time is obtained by basing the analysis on the precise data supplied by a 15,000-word lexicon and by tables of roots, prefixes, and suffixes.

- a parsing module which yields a coherent and efficient syntactic-prosodic hierarchy has been derived from the linguistic analysis.

- matching the coded hierarchy of the prosodic data base with the hierarchy generated by the prosodic parsing of the input text produces melodic patterns which emulate natural speech.

- the size of the prosodic data base is such that several melodic patterns are available for a given word and monotony in the synthesized speech can be avoided.

Numerous preliminary tests, using a PSOLA-KDG synthesizer and a partially completed set of prosodic markers and segmental duration rules, have demonstrated that the proposed options yield significant improvements in synthetic speech quality.

REFERENCES


