TEXT-to-SPEECH SYNTHESIS IN THE FRENCH ELECTRONIC MAIL ENVIRONMENT


ABSTRACT

At the present time, one of the most popular application of the text-to-speech (TTS) synthesis technique in the telecommunication area is the "text-to-voice" service allowing E-mail users to access their mailbox through an ordinary touch-tone telephone for receiving a voice output of their messages (BERNEY, 1985). We present here the specific problems encountered in French, and the adopted solutions for adapting the CNET's synthesis system to this specific application.

INTRODUCTION

The way from the laboratory studies to the industrial prototype is often tricky. For example, in the development of a "text-to-voice" facility for french E-mail service, a very specific problem appears, namely the frequent absence of graphical accents on the messages typed on an ordinary keyboard or on a "Minitel"**. TELEX messages, always written in capitals, are an extreme illustration of this problem. This lack of graphical marks has dramatical consequences on the vocal output if no pre-processing module, devoted to the restoration of missing accents, is introduced between the E-mail access and the TTS systems.

The automatic "loud-reading" of E-mail messages presents several other difficulties, such as the correct pronunciation of abbreviations, acronyms, the filtering of unpronounceable tables and the adequate "translation" of the graphical making-up. For these reasons, the introduction of a performant text pre-processing module seems, at the present time, to be one of the "key-points" conditioning the effective use of TTS technique for vocal E-mail retrieval services.

In the first part of this paper, we describe briefly the present structure of the CNET's TTS system, the recent improvements concerning the correct pronunciation of abbreviations and acronyms and the automatic generation of the prosody. In the second part, we describe the text pre-processing module allowing for a correct re-writing of the orthographic text in the case of missing accents. The third part gives a brief description of the architecture and user interface of this "text-to-voice" service in the E-mail environment.

I. THE CNET's TEXT-to-SPEECH SYSTEM FOR FRENCH

The detailed characteristics of the CNET's system (which uses diphones as synthesis units and the linear predictive coding technique) have already been described by STELLA (1983). Since this time, some improvements in the "sound quality" have been achieved:

* Centre National d'Etudes des Télécommunications - 22301 LANNION -
° SYSTEX 26 rue Hipolyte Maindron - 75014 PARIS - France.

** Minitel : a specific terminal distributed by the French Telecommunications for data-base access and text messaging services.
1) by carefully choosing both the speaker's voice from which the diphones were extracted, and the non-sense words used for building the dictionary,
2) by an adequate smoothing, in the spectral and intensity domains, at the boundaries of the concatenated diphones.

At the present time, this system is able to generate two synthetic voices, one male and one female, obtained by the use of 2 different diphones dictionaries. Three versions of each are available according to the frequency sampling (8, 10 and 16 kHz).

After the concatenation of the LP-coded diphones, the generation of the prosody is done in 2 steps:

1) A set of rules, using a small dictionary of function words and some rhythmical constraints allows for an automatic location of the main prosodic boundaries in the sentence (SORIN et al., 1987),

2) Two different prosodic modules are available : the first one tries to mimic the prosody used in reading (EMERARD, 1977), while the second one aims to simulate the peculiar prosody used in commercial announcements (SORIN and al., 1987).

From the parsing obtained in the first step, these modules define the Fo macro-contours for the sentence, the Fo micro-variations for the voiced consonants and semi-vowels (and in some contexts, for some parts of the vowels) and the duration modifications of the phonemes, depending of their word- and sentence contexts.

The conversion of the written text into phonetic symbols is done by a set of about 400 context-dependent letter-to-sound rules (DIVAY and GUYOMARD, 1977).

In addition to this module, other rules have been introduced for the detection of the acronyms (which are very frequent in the E-mail messages), the choice of their adequate pronunciation (as a whole "French word" or spelled), and the adequate translation of the most frequent abbreviations.

This new module offers the user the facility of entering or modifying the pronunciation of a given acronym or abbreviation (the significance of which may be dependent from the application domain).

In this system, acronyms, proper names, numbers and some abbreviations receive a particular prosodic status, because they often correspond to important informations. By these prosodic manipulations, the overall intelligibility of the synthesized message is notably improved.

At the present time, the version of the CNET's TTS system is written in C, for IBM-PC compatible microcomputers running under MS-DOS. The memory size needed is 300 kbytes for the programs and the dictionaries. This software generates frames that can be converted into speech by a software synthesis module written in assembly language on a TMS 32010 from Texas Instruments. The TTS software, called "Multivoc" is available by the french company Cap Sogeti Innovation. The synthesis module can be
easily adapted to any PC compatible board. A version exists on the TI "Speech board" and on the OROS AU-20 board. A reduced version of this TTS system is also available on the PC-compatible board called "Télévox", commercialized by Elan-Informatique. It uses simplified text-to-phoneme and prosodic processings.

II. THE RESTORATION OF MISSING ACCENTS:

An analysis of a corpus of randomly chosen TELEX messages showed that around 13% of the words had to be automatically reaccentuated. Among these words:

- 7% present non accent-ambiguity: for example, "REUNION" always corresponds to the word "rēunion"
- 6% present an accent-ambiguity: for example, "LE MARCHE QUE JE PASSE EST MERITE"

The pronounceable sentence is: "Le marché que je passe est mē­­rite". These ambiguities cannot be solved without knowing, at least, the grammatical status of the word in its context.

Therefore, a new text-preprocessing module has been developed, specifically for this purpose. This module corresponds to an adaptation of the linguistic processing part of the SPIRIT-System, developed by SYSTEX company, for accessing text data-bases in natural language. It includes:

- a full-form lexicon (270 000 forms plus the unaccentuated forms of the possibly accentuated words), where each word is coded into one (or more) of 170 syntactico-grammatical categories,
- a grammatical desambiguation module using succession-matrix, describing the possible sequences of 2 or 3 grammatical categories and some heuristic rules. These matrices were obtained by automatic learning (ANDREEWSKI and FLUHR, 1974): beginning with a grammatical hand-labelling of a small corpus (1000 words), the optimisation of the choice of the grammatical categories and the construction of the succession matrix were obtained iterately, with human corrections, on a larger corpus (70 000 words, juridical texts). This adapted version of the linguistic processor uses no information on the probabilities of the observed sequences.

This module was tested on a corpus of TELEX messages (12 000 words). Without any new training of the succession matrix, it allows for a correct orthographic "rewriting" of 93% of the words presenting a real accent ambiguity. Most of the remaining errors cannot definitively be solved without a semantic analysis of the text, as illustrated by the following example:

"MAIS IL FAUT" to be translated in "Mais il faut",

but

"MAIS A 30 FRANCS" to be translated in "Mais à 30 Francs".

For such cases, arbitrary decisions have been taken, based on simple criterions of statistical occurrences. This module also allows for an improvement of the orthographic transcription of some abbreviations.
III. APPLICATION TO THE E-MAIL "TEXT-TO-VOICE" RETRIEVAL SERVICE:

An industrial prototype of the E-mail "Text-to-voice" retrieval service is under development for the French Telecommunication Administration. It will be able to authorize the access to any compatible X400 standardized electronic mail. This new service will be available in 1989. This prototype will be able to manage up to 32 lines.

The structure adopted in that prototype is composed if a "token ring" of IBM-PC's. One PC (A) is devoted to the gateway with the E-mail through the switched-packet network Transpac. Two other PC's (B and C) are devoted to the general management of the "token ring" and of the service. Another PC (D) is devoted to the text pre-processing operation (this PC needs an additional RAM board of 16 kbytes allowing to process each word in 15 ms, this board will contain the whole French lexicon referred in the part II.). For offering the customer a "real-time access", each telephonic line will be connected to a PC (E) including the TTS software and composed of a specific board with the TMS 32010 processor.

When a customer will access to the service through the PC (E), the message will be transferred through (A) to be pre-processed in (D). Then the whole messages will be transmitted to (E) in two formats (with and without pre-processing to allow both the voice output of the messages and of the spelled words in their initial representation).

The ergonomy of the human/machine dialog was carefully studied. The user-interface is designed in a way allowing the user to manipulate the system as a "performant" tape reader. For example, the user can select a message, play it again, go back to the preceding message, go forward in the received messages, stop the reading at a given place, ask for spelling a word, modify the "listening comfort" parameters (speaking rate, loudness, pitch).

REFERENCES: