# THE MULTIVOC TEXT-TO-SPEECH SYSTEM

Danielle Ribot, Frédéric Le Diberder and Pierre Martin

Cap Sogeti Innovation
33, Chemin du vieux Chêne
ZIRST 38240 - Meylan FRANCE

ABSTRACT - MULTIVOC is a real-world text-to-speech system geared to the French language. The full system is described including the technical view and the main application of the product up to now as a basic component of a telephone-based mail service.

## INTRODUCTION

The text-to-speech MULTIVOC system is the result of a technology transfert from a research institute (CNET LANNION France), which conceived the basis of the system, to an industrial company (Cap Sogeti Innovation France) in charge of making the system a commercial product.
Starting from an ordinary French text, MULTIVOC generates in real-time high quality speech using a synthesis-by-diphone method.

The originality of this product resides in the following points.
- Its processing includes elaborate **phonetization** and **prosody** techniques. The combination of these two techniques allows MULTIVOC to generate continuous speech, including the liaisons specific to French, and providing the correct intonations and rhythm fluctuations for the input text.
- MULTIVOC provides several **run-time parameters** that allow the user to adjust the following speech characteristics:

  1- the style of prosody, *reading-style* corresponds to the usual way of reading a text while *advertising-style* is dedicated to short commercial messages.
  2- the gender of voice (*male* or *female*),
  3- the tone of the output voice (*from low to high in the range of 50-350 Hz*),
  4- the speech speed (*from 1 to 10 syllables per second*).
- MULTIVOC has been conceived as a **driver** which can be easily integrated into any application that needs spoken French outputs.
- MULTIVOC is a **real-time** component, running in a micro-computer environment on IBM-PC based system.

The application possibilities of MULTIVOC are numerous, including computer-aided education, alarm systems, data base interrogation and mailing services. We will present in the second part of this paper one of the most important applications developed up to now: MULTIVOC has been integrated as a basic component in a phone-based mail service developed for the French Telecommunications.

## DESCRIPTION OF THE MULTIVOC PROCESS

The overall processing is organized as a pipeline set of transformations applied to the input text, as shown in figure 1.
Each process takes as input the results of the preceding one and fills specific attributes of the objects composing the internal representation of the text.

- The **pre-processing** (or lexical processing) is a text-to-text transformation, which expands some non-word terms like numbers (*1987* -> "*mille neuf cent quatre-vingt sept*"), administrative codes (*A4/B5* -> "*A quatre B cinq*") , abbreviations (*Mr* -> "*Monsieur*") *, or acronyms (CSINN* -> "*Cap Sogeti Innovation*").

- The **phonetization** process transforms the pre-processed text into phonemes according to predefined rules stored in a user-modifiable rule-base.

- Then comes the steps corresponding to the prosody support. An automated prosody process [Aggoun 1987] has been developed to deal with the different voices "style" (reading, advertising) that can be used in MULTIVOC.

The prosody process includes two different levels

- macro-prosody related to the syntactic and semantic structure of the sentence and determining the general intonation and rhythms fluctuation on the sentence afterwards repercuted on each word. This level is supported by the **prosody marking** process.
- micro-prosody treating the interaction between the consecutives phonemes. This process includes two different steps. First the **rhythm marking** process computes the duration associated to each phoneme. Then the **frame generation** process computes the pitch value associated with each phoneme while the LPC frames corresponding to the input text are produced according to the different run-time parameters specified.

- These frames are then sent to the LPC interpreter of a speech synthesis device (not described here).

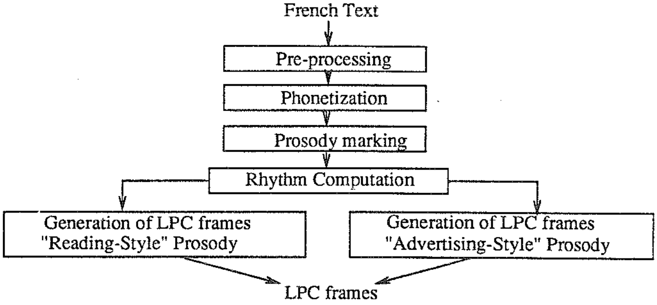

French Text
Pre-processing
Phonetization
Prosody marking
Rhythm Computation
Generation of LPC frames "Reading-Style" Prosody
Generation of LPC frames "Advertising-Style" Prosody
LPC frames

figure 1: MULTIVOC processing

PRE-PROCESSING

The purpose of this first step is to recognize the words corresponding to each sentence after the decomposition of the input text into a set of sentences.

Within this process the non-word terms are expanded to uniformize the subsequent processing.

- numbers, digital dates and time templates are expanded according to French rules
  *ex: '10:30:20' --> 'dix heures trente minutes vingt secondes'*
- abbreviations and acronyms are translated according to a user-defined lexicon. The corresponding translation can be
  - either empty the word is then spelled,
    *ex: 'MIT.' --> . (which will produce 'M I T' [EM EE TAY in French])*
  - either a full text string which will replace the matching word,
    *ex: 'MIT.' --> 'Massachusetts Institute of Technology' (in French!...)*
  - either a phonetic string if its pronunciation does not correspond to the French phonetization standards.
    *ex: 'MIT.' --> 'AI"MAYTI'. (better)*

This last facility can be very useful for company or product names or to treat telex-style messages.

Furthermore each single word is affected a lexical attribute used to characterize grammatical words (pronoun, determiner, prepositions) and the usual auxiliary verbs referenced in a predefined file. This dictionary is rather small (300 entries) but constitutes a sufficient and powerful basis for the following phonetization and prosody steps.
The potential advantages of a complete lexical analysis of the text would in fact not be significant as the size of the corresponding dictionary would imply a very time-consuming process preventing MULTIVOC to work in real-time. Anyway all the facilities to include a more complete analysis

exist so it will be very easy to include such a process when considered as quite efficient.

## PHONETIZATION

The phonetization process is carried out using a dictionary of rules which are applied to the input text to transform a sequence of characters into a list of phonemes.
This transformation is carried out by five sets of rules that are applied successively to determine first all the possible liaisons in the text, plural end word pronunciation, phonetic translation "sensus stricto", an elision process to deal with the so-called French "e-muet", and finally the suppression of the non correct liaisons between words.

Each rule has the following form:

```
   [<LC>] <MS> [<RC>] --> <PS> .
```

where

<MS> is the Matching Sequence of characters in the input text
<LC> and <RC> are the respective Left and Right contexts of the Matching Sequence
<PS> is the sequence of Phonetic Symbols to be generated

and has the meaning:

"Replace <MS> by <PS> if <MS> is preceded by <LC> and followed by <RC>.

Each context specification (<LC> and <RC>) can be empty, in which case the rule is applicable with no conditions, or can be expressed as a logical combination of elementary context:

```
  context == elementary.context AND context
          | elementary.context OR context
          | elementary.context
```

An elementary context is either a sequence of characters or a class of sequence of characters (e.g. consonants or vowels).
During interpretation, if several rules are applicable, the one containing the longest Matching Sequence is chosen: thus, the interpreter goes from the particular case to the general case. If more than one rule satisfies this criterion the first one is chosen and if no rule is applicable, a character is popped from the input and pushed to the output before the process start again.

*Example of rules:*

```
        [ _LORS | _PUIS | _QUOI ] QUE_ []  -->  <K><EU>_ .
        [] QUE_ []  -->  <K><E>_ .
```

*With the following characters playing a special role:*
*- the character '_' (underscore) denotes a blank character*
*- the character '|' denotes the logical operator OR*

## PROSODY MARKING

Only the macro-prosody support is addressed in this section.
In the first step each sentence of the input text is decomposed by words into rhythm and intonation units called prosody groups.
In a second step each word within a group is characterized with respect to the others and to the characteristics of the group.

### Prosody-group categorization

The input sentences are decomposed using a set of rules to determine the boundaries of a group and its associated category. The main criteria involved in this decomposition are the punctuation marks and the grammatical types of the words.
These few criteria are powerful substitutes to a complete syntactical analysis of the sentence defining the exact syntactic units within the sentence (subject group, etc...).
The resulting sequence of groups is then processed in order to adjust their categories. Here again, the process is governed by rules based on the following information:

- the length of the group (the number of words it contains),
- the number of syllables of each word within the group,
- the number and the length of non-lexical words,
- the category of the adjacent groups

*As an example of rule:*

```
    IF there exist a sequence (S) containing 3 groups of category '5'
      without a pause already established for one of them,
      AND if one of them (G) begins with one of the following determinant ('AU' or 'AUX')
    THEN give a category '4' to G
      and give it a short pause except if its pause is already long.
```

For instance, 50 rules of this kind allow a complete categorization of the groups.

### Word Marking

According to the category of the group it belongs to, its length, its grammatical nature, each word of a group is then marked and, if necessary, a pause is placed at the end of the word.
The set of rules used depends on the style of prosody required by the application ('reading' or 'advertising'). For the 'reading' style process only the last word of each prosody group is "marked" with respect to the others.
For the 'advertising' style we thus differentiate six types of words, with associated micro-marks used to adjust the initial pitch of the word.
*For example:*

```
    IF the group contains exactly 2 non-lexical consecutive words,
      AND the first one has one syllable
      AND the second more than one,
    THEN give the first word the mark '6+' and give the second the mark '4-'
```

Although some attempts have been made to express the prosody-marking rules in a declarative way [Sorin, 1984], [Aggoun, 1987], based on the logic paradigm, the efficiency criteria and the real-time objective we have defined for this product led us to represent them in a procedural way rather than in a production-rule form.
At the end of this process, some words remain unmarked. In the next processes, we consider a sequence of unmarked word terminated by a marked one (a prosody-word) as the basic entity to deal with.

## RHYTHM COMPUTATION

The third process involved in MULTIVOC consists in the computation of the duration to be associated to each phoneme. This duration is computed according to the different attributes attached to each word and to each phoneme, which are:

- the kind of phoneme (vowel [a], plosive [bang], fricative [french], liquid [long]),
- the mark associated to the corresponding word
- the number of syllables in the word
- the position of the phoneme within the word and a set of rules using this information.

*As an example of such rules:*

```
    IF the last phoneme of the word is a vowel
      AND the mark of the word is '5'
      OR if a pause is associated with the word,
    THEN give a coefficient of duration of '1.4' to this phoneme
```

*[Note: the default coefficient duration of every phoneme is '1.0'. This coefficient is combined to the speech speed parameter to obtain the duration of the phoneme.]*

## PROSODY GENERATION

Within this step the reading-style prosody differs from the advertising-style prosody by the fact the macro-melody and micro-melody are not dissociated. We develop in this section the advertising-style prosody facets. To every word-mark corresponds a macro-melody schema. This schema enables us to determine the variation of the pitch along the word.
Three basic functions are used to express the pitch variation:

- constant: the pitch remains unchanged
- linear interpolation
- exponential variation, namely $F(t) = F(to) * e^{-p(t-to)}$

where F(t) denotes the value of the pitch at the time 't', to is the initial time and p is a constant

(p = 0.68).
Every macro-melody schema begins at $F_{beg}$, the fundamental frequency of the speaker. $F_{beg}$ is set to 240 Hz for a Female voice and 120 Hz for a Male voice. This fundamental is adjusted if the word has a micro-mark '+' or '-'.
Then a set of rules determines when these functions should be applied to a word.
*As an example:*
*For words with mark '1' and containing more than four syllables:*
*- apply constant from the beginning until the middle of the second vowel,*
*- apply exponential with p/2 until the beginning of the first 'voise' phoneme of the last syllable (point A),*
*- apply constant* $F_{beg}/2$ *from the end of the last vowel (point B) to the end of the word,*
*- interpolate from A to B*
Then a set of micro-prosody rules is applied on the vowels ('fine tuning').
*Example:*

*IF a vowel is not in the last syllable of a word*
*AND followed by an unvoiced consonant*
*THEN the pitch of the last LPC frames of the vowel is adjusted in a decreasing way.*

Finally microfluctuations are introduced within the frames to overlay the "singing" effect resulting of a constant value from the pitch on consecutive frames.

At these step in the process, all needed information has been computed (pitch, duration) and MULTIVOC generates an LPC structure after having accessed a dictionary of diphones to get the coefficient of the lattice filter for each phoneme. This dictionary is specific to the style of voice (Male or Female) and to the sampling frequence used (8, 10 or 16 kHz) which characterize the quality of the output voice (the 8 kHz value corresponds to the pass-band used for French telephone, the 16 kHz value is the one providing the best results).

## IMPLEMENTATION OF MULTIVOC

The MULTIVOC software was developed in C on MS-DOS 3.2. This product is available either as a running package (binary form) for IBM-PC compatible computers or as an adaptable package (source form) for specific usage.
On the IBM-PC, the speech synthesis device used are the IBM-PC pluggable boards based on a Texas Instruments TMS320/20 processor. The ones we use are the OROS-AU20 supplied by OROS (France) and the IBM VCO (Voice Communication Option).
The MULTIVOC driver is implemented as a memory-resident program which applications can address via an interrupt mechanism. Doing this, any application can easily send text to be pronounced in real time.
A Microsoft Windows application has been developed to demonstrate the facilities offered by the system. Users can enter text using a built-in editor and can send all or mouse-selected text to MULTIVOC according to the parameters set previously.
MULTIVOC has also been ported to UNIX BSD 4.2 on a SUN-3 but the driver specific aspects have not yet been developed because of the lack of speech synthesis devices.

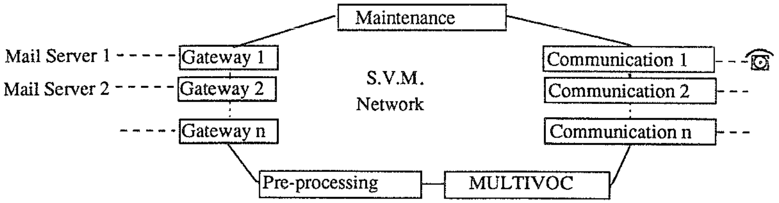
## S.V.M. PHONE-BASED MAIL SERVICE

The S.V.M system ("Système Vocal de Messagerie") allows using a French push-button telephone by the way of the vocal synthesis facilities to access the functionalities of (written) mail server. These include consulting of messages, answering to a given message, and automatic call of the mail server on prioritary messages.

The S.V.M. environment is constituted of the following components connected on a network:

- the application components in charge of managing the phoning communications and dialogue with the user and executing the appropriate commands.
- the gateway components between the S.V.M. and the different mailing servers accessible on the system.
- the pre-processing component in charge of restoring the accentuation of the messages transferred by the Mailing server.

- the MULTIVOC text-to speech component
- the maintenance component for the whole system.



Using the buttons on the phone, the user may dial commands to consult the different text input messages by the way of a mailing server.

Within the S.V.M system MULTIVOC provides the following interactive facilities:

- support of the dialogue messages with the user at all levels: introductory messages, help messages, answer messages.
- performance of the synthesis of a message referenced by an address
- interruption of the synthesis of a message and skip to the next one
- interruption of the synthesis of a message
- go back to the previous sentence already pronounced
- spelling of the last word that have just been pronounced.
- control of the voice parameters: voice tone, volume and speech speed parameters.

## CONCLUSION

Although based on a simple mechanism using only a local lexical analysis, avoiding expensive syntactic or semantic analysis, the results obtained with MULTIVOC are impressive. In particular, the output speech has very natural prosody. Finally, the performance achieved by MULTIVOC makes it a real-time Text-To-Speech system that will be widely applied in industry.

Research issues include the handling of other languages (English, German, Italian), knowing that some important parts of MULTIVOC have been dedicated to French for reasons of efficiency and therefore will have to be re-written. More valuable results are foreseen by applying our company's experience in natural language processing [Lancel, 1986], [Decitre, 1987] to the input phase of MULTIVOC.

## REFERENCES

Aggoun A. (1987) *Le système Synthex: Traitement de la prosodie en synthèse de la parole*, Technique et Science Informatiques, vol. 6, no. 3, 217-229

Decitre P., Grossi T., Jullien C., Solvay J.P., (1987) *Planning for Problem Formulation in Advice-Giving Dialogue*, 3rd Conference of the European Chapter of the Association for Computational Linguistics, Copenhagen (Denmark).

Emerard F., (1977) *Synthèse par Diphones et Traitement de la Prosodie*, Thèse de troisième cycle, Université de Grenoble.

Guidini A., Choppy C., Dupeyrat B., (1981) *Application de Règles au Calcul Automatique de la Prosodie. Comparaison avec la Prosodie Naturelle*", Symposium Prosodic, Toronto.

Lancel J.M., Rousselot F., Simonin N., (1986) *A Grammar Used for Parsing and Generation*, Proceedings of the XIth International Conference on Computational Linguistics, Bonn (FR Germany), 536-539,.