THE MULTIVOC TEXT-TO-SPEECH SYSTEM

Olivier M. Emorine and Pierre M. Martin
Cap Sogeti Innovation
Grenoble Research Center
Avenue du Vieux Chene, ZIRST
38240 Meylan, FRANCE

ABSTRACT

In this paper we introduce MULTIVOC, a real-world text-to-speech product geared to the French language.

Starting from a ordinary French text, MULTIVOC generates in real-time a high quality speech using a synthesis-by-diphone method. The processing is divided into 3 main transformations (phonetization, automatic prosody and rhythm marking, and generation of LPC frames).

This paper provides a full description of MULTIVOC including not only the technical view but also some applications of the product within the real world.

1. PRESENTATION OF MULTIVOC

The text-to-speech MULTIVOC system is the result of a technology transfer from a research institute (CNET Lannion, France), which developed the basis of the system, to an industrial company (Cap Sogeti Innovation, France) which made the system a commercial product. Generating Linear Prediction Coding frames from ordinary text written in French, the goal of MULTIVOC is to give any standard applications the ability to produce (in real time) low-cost and high-quality speech output.

MULTIVOC is shipped as a complete software system which aims to provide a sophisticated driver enabling applications to directly send French spoken text. The software package consists of the kernel of the driver itself and a set of dictionaries used by it. Several tools in the package allow an advanced user to tailor his own MULTIVOC driver to specific usage. Beside this static configuration facility, MULTIVOC also provides several run-time features. By submitting specific requests an application can change the following parameters:

- The sampling frequency for generated frames. Three different frequencies are available: 8 kHz, 10 kHk and 16 kHz. This parameter will characterize the quality of the output voice, a frequency of 16 kHz providing the best results.
- The tone of the output voice can be adjusted in the range 50-350 Hz.
- The speech speed may be set from 1 to 10 syllables per second.
- Two styles of prosody are provided. The "reading-style" corresponds to the usual way of reading a text, while the "advertising-style" is dedicated to short commercial messages like jingles.
- One can also choose between a female or a male voice.

The method used for the synthesis produces Linear Prediction Coding (LPC) frames generated from a diphone dictionary. Such a dictionary is specific to the sampling frequency used (8, 10 or 16kHz) and also to the style of voice (Female or Male). For this purpose, MULTIVOC provides 6 differents diphone dictionaries.

The overall processing is organized as a pipelined set of transformations applied to the input text. At the higher level, one can distinguish the following functions:

The **pre-processing** (or lexical processing) is a text-to-text transformation aiming to expande some non-worded terms like numbers (1987 --> "Mille Neuf Cent Quatre-Vingt-Sept"), administrative numbers (A4/B5 --> "A Quatre B Cinq") 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 base.

The **prosody marking** process scans the phonetized text and generates appropriate marks to reflect the prosody of the text using built-in rules based on the different punctuation signs and the grammatical type of words.

The **rhythm marking** process computes the duration associated to each phoneme.

Last, the **frame generation** process produces the LPC frames which correspond to the input text according to the different parameters specified and can be sent directly to the output device.

In this overall processing, we have deliberately avoided a time-consuming syntax analysis, to enable MULTIVOC to run in real time. This choice has made MULTIVOC a commercially viable product providing a high-quality speech at low cost and which has been sold to serve as a basic component for several industrial applications.

MULTIVOC is available on IBM-PC based systems.

2. THE MULTIVOC PROCESS

As explained in the previous section, the input text provided by an application is processed in a "pipe-line" through five processes (see figure 1).

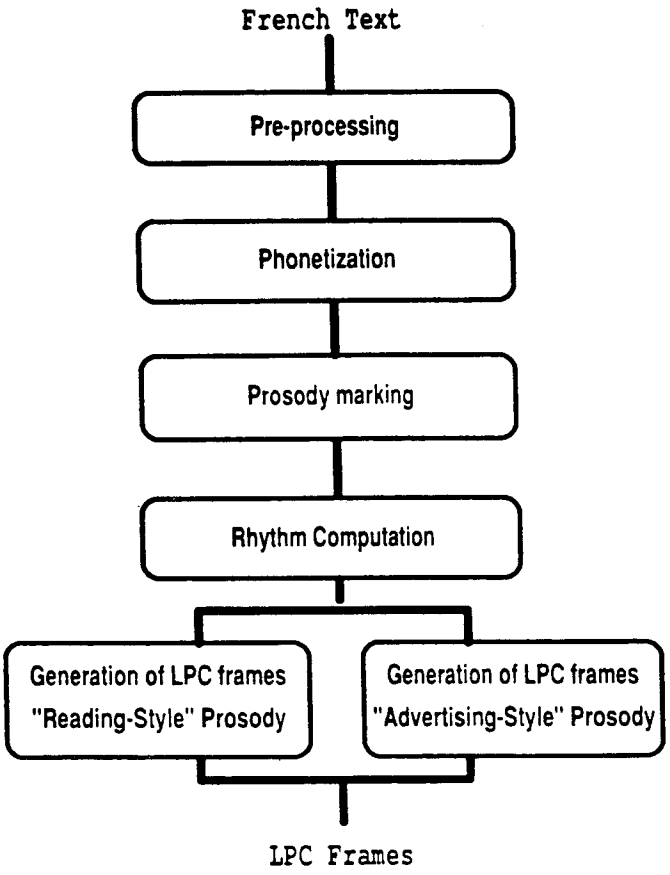

Figure 1: The MULTIVOC processing

Each process takes as input the result of the preceding one and fills specific attributes of the objects composing the internal representation of the text. The final result, a list of LPC frames, is then sent to the LPC interpreter of a speech synthesis device (not described here).

■ PRE-PROCESSING

The main purpose of this first step is to decompose the input sentences into a list of words and to set the lexical attributes of each word. In order to allow ordinary-written text to be correctly processed, some patterns are translated into a sequence of words:

- numbers are expansed according to the French language rules. The words generated are tagged to permit a correct prosody marking for numbers.

• digital dates, time templates (not exhaustive) are matched against corresponding patterns in a set of rules which define the transformation to be applied. Patterns corresponding to the matching part and the transformation format are expressed using a UNIX-scanf/printf-like syntax.

• abbreviations and acronyms are translated according to a user-defined lexicon. The translation part associated each entry of the lexicon can be:

- empty to specify that the recognized word is to be spelled

  ex: 'MIT.' --> . (which will produce 'M I T' [EM EE TAY in French])

- a full text string which will replace the matching word

  ex: 'MIT.' --> 'Massachusetts Institute of Technology' (in French!...)

- a phonetic string if the pronunciation is very different from the lexical form. This function is particularly useful for company or product names ...

  ex: 'MIT.' --> 'AI"MAYTI'. (better)

• mathematic symbols are also translated

The process then checks if each word can be pronounced, according to a dictionary of the French sequences of pronounceable letters, and if it cannot the word is spelled.

Finally, an attribute is associated to each word describing the grammatical nature of the word (pronoun, determin, preposition, ...). This dictionary is rather small (300 entries) and does not contain most verbs but does contain the usual auxiliaries.

A complete analysis of the sentences would provide a better prosody but, due to the size of the corresponding dictionary, could not be processed in real-time. The resulting prosody is nevertheless judged very natural, albeit in some few cases somewhat strange.

■ PHONETIZATION

This process transforms the sentences into a sequence of phonetic symbols. This transformation is carried out by five set of rules. The sets are applied successively to the input text.

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> .
```

Note: several characters play a special role:
- the character '_' (underscore) denotes a blank character
- the character '|' denotes the logical operator OR
- the character '&' denotes the logical operator AND

One of the set of rules is dedicated to the determination of the correct liaisons between words.

■ PROSODY MARKING

The synthetic speech produced by mere concatenation of diphones is comprehensible but not very natural. To provide it with an acceptable quality, it is necessary to operate a prosody processing.

Prosody facts are of two kinds (Emerard, 1977), (Guidini, 1981), (Sorin, 1984):

• macro-prosody, related to the syntactic and semantic structure of the sentence,

• micro-prosody, treating the interaction between two consecutive phonemes.

A study of a set of phrases and the diversity of the voice "styles" (reading, advertising, ...) has provided an automatic prosody generation system (Aggoun, 1987). In the first step, this process decomposes the sentences in a set of so-called prosody-groups, and associates to each of them a group category. In the second step, each word within a group is marked and a pause is associated with it.

Prosody-Group Categorization

A prosody-group is by consecutive words. A set of rules determines the boundaries of a group and its associated category. The main criteria involved in this decomposition are:

• the punctuation marks (including the end of a sentence), each of them defining a different category

• the grammatical natures of two consecutive words.

For example, a group ends after a lexical word (noun, non-auxiliary verbal form) followed by a grammatical word (determinant, pre-position, ...). In that case, the category of the group depends on the second word.

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.

[Note: some of them are simpler !]

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, possibly, a pause is placed at the end 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-'

It should be noted that the set of rules used depends on the style of prosody required by the application ('reading' or 'advertising').

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-Srule 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 associate 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 (plosive [bang], fricative [french], liquid [long]),

• the mark associated the word

• the number of syllabin of 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 duration of '1.4' to this phoneme

[Note: the default duration of every phoneme is '1.0']

■ PROSODY GENERATION

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 $F(t) = F(to) * e^{-p(t - t0)}$
  where F(t) denotes the value of the pitch at the time 't', t0 is the initial time and p is a constant (p = 0.68)

Every macro-melody schema begins at F_{deb}, the fundamental frequency of the speaker. F_{deb} 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 $F_{deb}/2$ 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 the following manner:

let C = [F(LF - 3) - 7/12 * F_{deb}] * 100 in
F(LF - 2) = F(LF - 3) - 10 * C
F(LF - 1) = F(LF - 3) - 15 * C
F(LF) = F(LF -3) - 20 * C

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.

3. IMPLEMENTATION OF MULTIVOC

The MULTIVOC software was developed in C on MS-DOS 3.2 and is compatible with UNIX BSD 4.2. This product is sold 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 comes from the OROS Company (France) and is featured as an IBM-PC pluggable board (OROS-AU20) based on a Texas Instruments TMS320/20 processor. The MULTIVOC driver is implemented as a memory-resident program which application can address using an interrupt mechanism. Doing this, any application can very easily send text to be pronounced in real time.

A Microsoft Windows application has been developed to demonstrate the facilities offered by MULTIVOC. Users can enter text using a built-in editor and can send all or mouse-selected text to MULTIVOC. A form (Dialogue-Box) allows the different parameters of MULTIVOC to be set to user specified values.

MULTIVOC has also been successfully 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 for such machines.

4. APPLICATIONS OF MULTIVOC

We give below three examples of concrete and real-world applications of MULTIVOC in an industrial context:

- The first one was to use MULTIVOC to pronounce TELEX-style messages. This has been realized by defining an appropriate lexicon for the numerous abbreviations and acronyms used in such messages. The sources of MULTIVOC have not been modified.
- The second application, or class of application, is to adapt MULTIVOC to low cost and small home-computers to develop a new generation of product for this market (Computer aided education software, for example). This is conducted by two customers who bought the sources of MULTIVOC and are now producing a restricted version of the product.

• The third application is to use MULTIVOC as a basic component in a sophisticated application. We are now running a project for the French Telecommunications (DGT) to develop phone-based mail services. Using a standard French phone, any user will be able to call the mailing service and dial commands to hear the different messages he has received. Several user-friendly features will enable to hear again part or all of a message or to change MULTIVOC-like parameters (deeper voice, slower, ...). For the purposes of this project MULTIVOC will not be changed.

5. FURTHER WORK

The work planned around MULTIVOC is of two kinds: the more research issues and the more commercial/industrial ones.

Research issues will include the handling of other languages (English), 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.

As a commercial issue, we will continue to sell the MULTIVOC software system and to collaborate with our customers. In the industrial field we think that a component like MULTIVOC will be a much-appreciated complement to many common applications. To prepare that, we envisage to install MULTIVOC on other machines and other operating systems and this should not cause any trouble.

We will also adapt MULTIVOC to different speech synthesis devices based on the linear prediction technique. Finally, we will investigate the use of other synthesis technics (synthesis by formants for instance).

6. CONCLUSION

Although based on a quite 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.

7. REFERENCES

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

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

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

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

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

Sorin C., Stella M., Aggoun A., Barthkova K., "Règles Prosodiques et Synthèse de la Parole 'MULTI-STYLE', Symposium Franco-Soviétique sur le Dialogue Homme-Machine, Pouchino, 1984.