MULTILINGUAL PSOLA TEXT-TO-SPEECH SYSTEM

D. Bigorgne, O. Boeffard, B. Cherbonnel, F. Emerard,
D.Larreur, J.L. Le Saint-Milon, I. Métayer, C. Sorin, S.White

FRANCE TELECOM-CNET, Department TSS/RCP
Route de Trégastel, 22301 Lannion-FRANCE

ABSTRACT

This paper describes recent developments of the CNET PSOLA text-to-speech system towards the availability of high-quality multi-lingual versions using the same PSOLA synthesizer. A new system architecture (SYC) has been specifically designed for assuring real-time, multi-channel and interactive functioning

The data control and transfer methods within the system are now fully independent from the content of the modules, allowing for an easy adaptation of the system to multi-lingual functioning. A complete version has been developed for German and "mixed" systems have been elaborated for Italian and English in collaboration with other laboratories using the same PSOLA synthesizer as for French. The overall quality of the French version has been improved both in naturalness (optimization of the linguistico-prosodic processings) and in articulation accuracy (use of longer units to be concatenated). An automatic segmentation procedure has been developed for the rapid building of new repertories of speech units from recordings by new speakers, i.e. for the creation of new synthetic voices.

INTRODUCTION

This paper describes the main results recently obtained at CNET towards the realization of high-quality multilingual text-to-speech systems well adapted to industrial requirements, in particular for their use in interactive voice services over the telephone network. The developments reported here correspond to various extensions of previous work on concatenation-based Text-to-Speech (TTS) systems that led to the specification of the PSOLA/TD synthesis method [1,2] and to the realization of a TTS system for French generating high-quality synthetic speech : CNETVOX/PSOLA [3]. The main characteristics and overall structure of the multi-lingual system remain the same as for French A first set of (language-dependent) "high-level" modules transcribe the input text into a sequence of phonemes supplied with prosodic values. A second set of "low level" modules (the language-independent synthesizer) reconstruct the speech signal by concatenating speech-units waveforms extracted from a (language-dependent) repertory and applying the PSOLA process in order to respect the prosodic requirements. The current version of this CNET PSOLA/TD synthesizer is no longer combined with low-bite rate speech coding as described previously [3]. Therefore it doesn't require anymore a DSP board. A purely software version of the complete system runs in real-time under DOS or UNIX.

The following sections report on the works which have been carried out along 4 main lines :

1 - a new system architecture (SYC) has been specified and implemented for assuring, under various kinds of operating systems, real-time, multi-channel, interactive functioning as well as versatility of the whole system (in particular easy substitution of one language-specific set of modules by the corresponding set of modules for another language).

2 - using the same PSOLA synthesizer as for French, a complete TTS in German has been realized at CNET (ALLVOC/PSOLA) and "mixed" systems have been developed for Italian and English in collaboration with other laboratories : these "mixed" systems incorporate the language-dependent "high-level" modules elaborated, independently, by our partners and can use either their (language-specific) speech unit repertories or the CNET ones.

3 - the overall quality of the CNETVOX/PSOLA TTS system for French has been improved in two ways : first, by optimizing the linguistico-prosodic processings (improved naturalness) and secondly, by allowing the current PSOLA synthesizer to process longer units, such as triphones and quadriphones, that are now included in the speech unit repertory for French (improved articulation accuracy).

4 - an automatic segmentation procedure has been developed for allowing the rapid building of new repertories of speech units from recordings by different speakers, i.e. for the creation of new synthetic voices.

1 - ARCHITECTURE OPTIMIZATION

In the previous version of the CNET TTS system [3], the various transducer modules communicated by files. For guarantying real-time processing and shortest reaction time, assuring easy multi-lingual functioning (real-time switching from one language to another, for example), allowing really interactive user commands (like spelling, replay etc...), handling simultaneously several channels with a single system, accepting mixed inputs (text and audio files, for example) and guarantying independence vis a vis computer operating system (and allowing board implementation), a specific architecture (SYC : SYnthesis Control) has been designed. This SYC architecture relies on the specifications of an Synthesis Module Programmation Interface that insulates every processing transducer module from the rest of the system. It includes a set of constraints relative to module declaration, specification of the nature of the databases to be accessed (rules, speech units, prosodic tables etc...) and a set of

0-7803-0946-4/93 $3.00 © 1993 IEEE

primitive functions for I/O operations, activity signalization (output of significant data, no input, error) and service actions (error messaging etc...). The description of the data streams doesn't include any detailled constraints on the processed items but allows for the transfer of "mixed" data (such as partially processed data, audio files etc...). The respect of this interface authorizes free combinaison of SYC compatible modules within the system without any modification of the source code. Different versions of the executive kernel have been developed allowing either mono-channel implementation (with or without interactive functioning) or multi-channel implementation. All these versions are Operating System independent: they runs under UNIX, DOS or on PC boards.

In addition to the fulfilment of industrial requirements for introducing TTS in telecommunication applications (real-time, interactivity , multi-channel), the genericity of this architecture allowed other European partners to quickly adapt their language-specific TTS modules for interfacing them with the CNET PSOLA synthesizer (see section 2 below). The development of this SYC architecture allows to demonstrate the feasibility of building industrial multi-lingual TTS systems sharing an unique synthesizer (the PSOLA concatenation-based synthesis module) but using language-dependent "high-level" modules and speech unit repertories developed independently in different laboratories.

2 - PSOLA-TTS IN GERMAN and prototypes of "mixed" systems in other languages

2.1. PSOLA TTS in German : ALLVOC/PSOLA

ALLVOC/PSOLA is a complete TTS in German language that combines the CNET PSOLA synthesizer software used for French, a specific repertory of 2750 German subword speech units and a set of "high-level" processing modules specific to German language, ALLVOC.

For the elaboration of the speech unit repertory, a corpus of words and logatomes (non-sense words) has been used, that allowed to extract two sets of units : 2310 diphones and 440 additional "special" units containing glottal stops (either /vowel-glottal stop-vowel/ or /consonant-glottal-stop-vowel/). These later units appeared to contribute significantly to the naturalness of the synthetic voice. The whole corpus has been recorded by a carefully chosen male speaker. The segmentation of the speech units has been performed, at first, by a phonetician expert. A second version of this speech units repertory has been generated using the fully automatic segmentation procedure described in section 4 below (see section 4 for the comparison between these two segmentations). The pitch-marking of these units has been done by an entirely automatic procedure : this pitch-marking procedure uses two complementary pitch detection algorithms (auto-correlation function and cepstrum) and includes various procedures (smoothing, detection and corection of Fo discontinuities, choice of the optimal placing of the pitch marks on the waveform) that makes it particularly robust, accurate and well adapted to the subsequent PSOLA processing.

The "high-level" processing module set for German, ALLVOC, includes 5 main components : pre-processing, morphological analysis and grammatical labelling, syntactico-prosodic parsing, orthographic-phonetic transcription and prosodic values generation.

The pre-processing module [4] performs the normalization of the input text by converting numerical expressions, abbreviations, diacritics etc... into sequences of graphemes. It also marks the nouns which are detected through their initial capital letters.

The morphological analysis and grammatical analysis module [4] carries out the segmentation of derived and compound words, the lexical stress assignment and the grammatical labelling of each word. This is done by combining rules and lexicon look-up. A large lexicon of roots (8000 entries) is used as well as 3 dictionaries of about 80 prefixes, and 50 suffixes and 370 function words.

The syntactico-prosodic parsing [4] allows to determine the boundaries and types of prosodic groups in the sentence and to insert a limited number of syntactico-prosodic markers. Primarily, the parser operates by comparing the sequences of grammatical labels to the sequences of labels defined within hierarchized parsing rules, similarly structured to those used for French [5]. The orthographic-phonetic transcription module uses the results of the morphological analysis and includes some 400 rules and 10 sets of exceptions [6, 7].

Using the results of the syntactico-prosodic parsing, prosodic markers patterns are then attributed to each word (13 prosodic markers) The prosodic marker assignment and the specification of corresponding prosodic patterns (Fo and duration values for each phoneme) are done entirely automatically by a set of rules and tables, similarly structured to those previously used for French [8].

Adapted to the SYC interface, the complete ALLVOC/PSOLA software runs in real-time under various kinds of operating systems. It provides notably improved sound quality in comparison to commercially available systems for German. However work remains to be done, mainly at the linguistico-prosodic level, to reach the same naturalness as the one obtained currently for French with CNETVOX/PSOLA. A tape will be demonstrated at the Conference.

2.2. "Mixed" systems for other languages

The availability of precise specifications of the interfaces between the language-dependent components of the TTS system ("high-level" processing modules and speech unit repertories) and the language-independent PSOLA/TD synthesizer allowed us to collaborate with other laboratories for the rapid prototyping of "mixed" systems in other languages. For example, a complete PSOLA-TTS system respecting the SYC interface has been developed with CSELT (Italy) for the Italian language : this "mixed" CSELT-CNET system combines the "high-level" processings developed independently by CSELT for the Italian language [9], the CNET PSOLA/TD synthesizer and either of the 2 Italian diphone repertories built separately by CSELT and CNET (roughly 1000 diphones each). Here again the resulting sound quality is notably improved in comparison to other available systems for Italian. A tape will be presented at the Conference.

Similar collaborative work is currently underway for English and Spanish.

3. OPTIMIZATION OF THE CNETVOX/PSOLA SYSTEM FOR FRENCH

The overall quality of the CNETVOX/PSOLA system for French has been improved in two ways : first, at the linguistico-prosodic level for improving the naturalness of the synthetic speech, secondly at the synthesizer level for improving the articulation accuracy.

3.1. Linguistico-prosodic optimizations for French

The CNETVOX high-level processings for French have been optimized for producing a synthetic speech naturalness better suited to the expectations of the users in automatized man-machine dialogue applications.

With respect to the previous version of the system [5], the main improvements concern the phonetic transcription of proper names (the use of new rules and lexicons leads to an average proper name pronounciation error rate of 7 % on 1000 telephone directory utterances including one patronymic, one street and one city name each, i.e. 3000 proper names in total) and the prosody [10]. The only modification introduced in the syntactico-prosodic parsing concerns the addition of a new rule identifying alternative questions. All the other modifications concern the rules for assigning the prosodic markers and the specification of the prosodic contours associated to those markers. 27 new prosodic markers (in addition to the 30 previously defined markers) allow to better process interrogative utterances, emphasis, interjection, request for waiting etc..., i.e. linguistic forms that appear frequently in man-machine dialogues.

Formal listening tests have been done with 15 naïve listeners which were asked to evaluate the acceptability of various versions of the CNET "high-level" processing modules for different telephone application conditions (either completely oral man-machine dialogue or unilateral information delivery or touch-tone input/speech output man-machine dialogue). The results show that the listeners do, in fact, largely prefer the new CNETVOX version for situations corresponding to interactive spoken dialogues : the average acceptability score, measured with the magnitude estimation technique developed for assessing the performance of TTS synthesis sytems [11], increased by 10 % [10] with respect to the previous CNETVOX version.

3.2. Synthesis optimizations for French : use of longer units

In concatenation-based synthesis, the choice of the units to be concatenated plays a key-role: the acoustic differences between the stored segment and the requested one as well as the acoustic discontinuities at the boundaries between adjacent segments have to be minimized. Currently, this choice still require to maximally exploit a priori phonetic knowledge, even if progresses have been made towards automatic identification and selection of optimal speech units for Synthesis [12].

The results presented here correspond to an intermediate step towards the realization of a PSOLA synthesizer (and associated speech unit repertory) able to optimally and automatically choose the nature of the units to be concatenated, as a function of the linguistic and phonetic context [13].

The current CNET diphone repertory for French includes 1290 diphones which are extracted from recordings of carefully defined non-sense words ("logatomes"). For each diphone, the left and right phonetic contexts within the logatome have been chosen in order to minimize the coarticulation effects of the surrounding phonemes on the diphone constituting phonemes.

However, several exhaustive intelligibility tests [14] showed severe remaining problems for specific phonemes,.especially in consonant clusters. This concerns particularly the semi-vowels [w,j] and liquids [l,r] which are highly coarticulated and pose segmentation problems.

Therefore, a supplementary corpus of triphones and quadriphones has been defined where these sounds are stored (and "protected") within larger units whose boundaries can be more easily segmented. The choice of these triphones and quadriphones takes into account the statistics of occurence of sequences of phonemes observed in spoken French [15] : only the most frequent sequences are stored. A total of 1047 new units have been added to the previous corpus of 1290 diphones : 153 triphones including [y], 54 triphones including [w], 544 triphones including either [l] or [r] (same constituting rules), 288 quadriphones including [y] and either [l] or [r] and 8 "special" triphones and quadriphones.

These 1047 triphone and quadriphone units are currently processed by the same version of the PSOLA/TD synthesizer developed for the processing of diphones. They are segmented in specific diphones, designed as "crypto-polyphones", which are added, with specific marks, to the diphone repertory. At the synthesis stage, the original triphone or quadriphone is reconstructed by concatenating the crypto-polyphones corresponding to this specific unit,.: the choice between the use of diphones or a "crypto-polyphone" is made by simple identifying the longest available unit through a left-to right look-up of the phonetic string from the current phoneme.

Preliminary intelligibility tests (12 listeners) on 170 monosyllabic isolated words including these fragile phonemes [w,j,l,r] showed that the addition of the triphones described above already reduce the intelligibility error rate by 20 %, with respect to the diphone version. More exhaustive intelligibility tests with the full diphone, triphone and quadriphone repertory [i.e,.2337 units] are currently underway. The results will be presented at the Conference.

4. AUTOMATIC SEGMENTATION OF SPEECH UNITS

Multi-voice synthesis is an important challenge for industrial exploitation of TTS. While algorithmic voice modifications are the true solution for achieving this goal, an intermediate solution is to speed up the procedure of creating new speech unit repertories by automatizing the unit segmentation process after the unavoidable phase of recording items (logatomes, words or word sequences) pronounced by the new speaker.

An entirely automatic segmentation procedure has been recently developed [16]. The recorded items (logatomes or words) are first segmented into phonemes using HMMs trained on the whole recorded corpus without any prior manual labelling. In a

second phase, optimal concatenation points are identified using a spectral distance measure.

A first evaluation on the French diphone repertory showed that the difference between manual and automatic segmentation is inferior to 30 ms for 90% of the diphones[16].

This method has been also applied to the building of the new German units repertory described in Section 2. The difference between manual and automatic segmentation marks assigned to the boundaries between 2 phonemes ("middle" of the diphone) is inferior to 20 ms for 82% of the units. It should be noted however that the segmentation accuracy is notably higher for the units which are extracted from non-sense words (logatomes) than for those extracted from real words. For the units extracted from logatomes (75% of the German corpus), the differences are less than 20ms for 89% of the units while this remains true for only 68% of the units extracted from real words (25% of the German corpus).

This observation emphasized the need for a rigorous control of the phonetic structure of the items from which the units have to be extracted. A non-homogeneous choice of these unit contexts may lead to audible distorsion at the synthesis stage but also prevent the use of fully automatic segmentation procedures for the building of new speech unit repertories.

Exhaustive intelligibility tests are currently underway for comparing the automatic and manual versions of the repertories on perceptual bases. The results will be presented at the conference.

CONCLUSION

This paper summarizes the work done recently at CNET for developing multi-lingual concatenation-based PSOLA TTS systems, well adapted to their use in Interactive Voice Services.

It shows the feasibility of building multi-lingual TTS systems sharing a unique synthesizer (the PSOLA concatenation-based synthesis module) but using language-dependent "high-level" processing modules and speech units repertories developed, independently, in different laboratories. While the naturalness of the timbre is preserved, due to the intrinsic properties of the PSOLA algorithm, the quality of the resulting synthetic speech depends rather strongly from the choice of the units to be concatenated, from the exploitation mode of the PSOLA method (in particuler, the choice of the pitch-marking procedure) and from the quality of the linguistico-prosodic processings. Speech unit repertory building can now be done entirely automatically, allowing to notably speed up the creation of new voices.

A French version of this system (CNETVOX/PSOLA) is currently in use in France, over the telephone network, in a large-public Interactive Voice Server for catalogue sale.

REFERENCES

[1] PSOLA-TD : French patent N° 8811517, issued September 2, 1988 : "Procédé et dispositif de synthèse de la parole par addition/recouvrement de formes d'ondes".

[2] C. HAMON, E. MOULINES, F. CHARPENTIER (1989) : "A diphone Synthesis System based on time domain prosodic modification of Speech", Proc. ICASSP 89, Glasgow.

[3] E. MOULINES et al. : "A real-time French Text-to-Speech system generating high-quality synthetic speech", Proc. ICASSP 90, Albuquerque, 309-312.

[4] B. SCHNABEL, H. ROTH (1990) : "Automatic linguistic processing in a German Text-to-Speech synthesis system", Proc. ESCA/ETRW on Speech Synthesis, Autrans, 121-124.

[5] D. LARREUR, F. EMERARD, F. MARTY (1989) : "Linguistic and Prosodic Processing for a Text-to-Speech Synthesis System', Proc. EUROSPEECH 89, Paris.

[6] B. SCHNABEL (1988) : "Développement d'un système de synthèse de l'Allemand à partir du texte", Doctoral Thesis dissertation, University of Grenoble III.

[7] B. SCHNABEL, F. CHARPENTIER (1988) : "Multilinguale Sprach synthese", Proc. BIGTECH Conf, Berlin, 97-105.

[8] C. SORIN, et al (1987) : "Text-to-Speech Synthesis in the French E-mail environment", Proc. European Conference on Speech Technology, Edinburgh, 260-263.

[9] L. NEBBIA (1990) : "Text-to-Speech Synthesis Systems for Italian : an overview", Proc. VERBA 90 Conference, Alcatel Face, Roma, 326-333.

[10] J. HOUSE, C. SORIN (1992) : "Evaluation and Respecification of the prosodic output component", SUNDIAL Esprit Project P2218, Deliverable WP7 D8.

[11] C. PAVLOVIC, M. ROSSI, R. ESPESSER (1990) : "Use of the magnitude estimation technique for assessing the performance of TTS synthesis systems", JASA 87, 373-382.

[12] N. IWAHASHI, N. KAIKI, Y. SAGISAKA (1992) : "Concatenation speech synthesis by minimum distorsion criteria", Proc. ICASSP92, San Francisco.

[13] S. NAKAJIMA, et al (1988) : "Automatic generation of synthesis units based on context oriented clustering", Proc. ICASSP 88, 659-662.

[14] F. EMERARD, M. CARTIER (1991) : "Test d'intelligibilité de systèmes de Synthèse à partir du texte", CNET Technical Report CNET, NT/LAA/TSS/427.

[15] J.P. TUBACH, L.J. BOE (1985) : "Un corpus de transcription phonétique : Constitution et exploitation statistique, ENST" Technical Report, ENST-Paris, 85D001.

[16] O. BOEFFARD, L. MICLET, S. WHITE (1992) : "Automatic generation of optimized unit dictionaries for text to speech synthesis" Proc. ICSLP Conference, Banf, 1211-1215.