SPEECH WORKSTATION FOR ITALIAN TEXT TO SPEECH DEVELOPMENT

G. Abbattista, A. Riccio, S. Terribili

ALCATEL FACE STANDARD, Research Centre
Pomezia (Rome) - ITALY

ABSTRACT - The paper describes the implementation and use of a powerful Workstation suitable for the generation of the acoustic units database and the study and evaluation of the prosodic contours for a Text to Speech system for Italian language.

INTRODUCTION

The development of a text to speech system, based on the segment concatenation approach, requires for the acoustical part, to build up a database of speech elements; these elements have to be extracted from real speech, coded in a convenient technique, and stored together with some additional information necessary for the concatenation. Generally this process is highly time consuming, it requires an expert operator and possibly a good level of interaction to allow corrections, afterthoughts and similar needs; moreover the final quality of the complete system is strongly affected by the appropriateness of this part of the development.
It is obvious that to overcome these problems the optimal solution would be an integrated environment provided with all the necessary tools; over the past years, such an environment was available as stand-alone modules running typically on large main frames; as a result, the grade of flexibility and the level of interaction were poor; nowadays, these limitations can be completely removed, thanks to the availability of fast and powerful DSP's combined with a developer-oriented interface.

SEGMENT DEFINITION

As known from the relevant literature, different choices can be made on the type of segments to be used; for example one could choice diphones, triphones, syllables, morphemes and at least whole words. In any case, regardless of the type of segment, a common problem exists: where exactly to cut the segments in order to be concatenated with the adjacent ones (preceding and following) and how to vary its time duration when the same segment appears at the beginning or at the end of a word.
In other terms this would mean that when extracting the generic segment, the developer should be able to mark in a clear way several characteristic points in the segment to be used subsequently as information for the concatenation.
For example, in our system, the basic segments are diphones, and we found that, besides the indication of the points where a segment starts and ends, it is also needed at least another pair of markers to indicate where to realize the concatenation (when it is required), and these markers depend on the phonetic context.
It is clear that all these markers have to be assigned to each segment, very often iteratively; as the total number of segments increases as a function of the phonetic complexity of the segment itself, a manual procedure is unthinkable.

SYSTEM DESCRIPTION

Hardware

The digital signal processing capabilities are demanded to a specific hardware designed at the same laboratory; it consists of a PC compatible add-on board. Actually the board can support different kind of processing, therefore it is not uniquely adopted for text to speech; in fact, it is

also capable to run speech coding techniques (ADPCM, RELP, LPC) and speech recognition algorithms based on a DTW approach.
Most of the DSP computations necessary for the algorithms mentioned are carried on a TEXAS Instruments TMS 320C25; the board includes other two processors, a MOTOROLA 68000 as CPU and an ALCATEL proprietary custom chip.
The communication between the DSP processor and the 68000 is performed via a shared memory (32 Kword), and all programs to be executed on the DSP are downloaded from the mass memory of the PC; therefore, after the initialization the DSP board acts as a stand alone system. Moreover, since the board is capable to directly interface a telephone line, the A/D and D/A functions and filtering are obtained with an industry standard CODEC; as a consequence the sampling frequency is fixed and equal to 8 kHz.

Software

The complete system can be considered as the combination of SW modules running on the PC under MS-DOS, DSP SW modules running on the board and executed by the TMS 320C25 and a supervisor module running on the board on the 68000 processor.
All SW modules running on the PC are written in Turbo PASCAL ver. 4.0 while the basic DSP modules have been written directly in TMS 320C25 Assembly in order to optimize the timing of crucial routines. Moreover, one of the main characteristic of this software modules is their intrinsic flexibility that offers the developer several options ranging from the choice of an alternative A/D and D/A board up to even more detailed parameters like the number and type of markers to be used for the concatenation. This is a very essential feature, since the amount of knowledge and information one can put in the segmentation process is inherently dependent on the expertise of the developer; with our system the level of details is not fixed, it can be increased and updated at any time.

SOFTWARE TOOLS

Acquisition, segmentation and coding

Generally, the segments necessary for a concatenative speech synthesis are extracted from suitable sentences uttered by a professional speaker; once the speaker's voice has been recorded as audio signal on convenient storage media (open reel tape, DAT), the developer has to convert the audio signal in data files: this stage of the development is called the acquisition; it basically consists in sampling the audio signal and store it as PCM coded files. The developer, using our speech workstation, has several choices, he may use the CODEC on the DSP board or other A/D boards available in the PC; it is also foreseen the possibility that the acquisition process will be performed elsewhere, in this case PCM files will be already available and the user has just to download them.
The segmentation will be realized using the PCM data files; the developer has the facility to display on the screen the waveform of the signal and, at the same time, to listen to the entire sentence or to a portion of it identified by a couple of cursors; to facilitate the task, additional information can be displayed on the screen as for example a formant tracking computed in real time.
On the basis of these information and iteratively listening to the acoustic feedback, the designer can select the portion of speech signal he wants to extract; the segment will be identified by a starting frame and an ending frame; further information, as for instance, the concatenating frames identified by another couple of markers, will notify which frame of the segment has to be used for concatenation, depending upon the phonetic context.
The user can utilize up to ten markers, whose meaning and usage can be defined in a set-up menu.
All these information will be stored in an appropriate header in order to be used by the subsequent stage that is the coding; this module performs the LPC coding of the relevant portion of the PCM files on the basis of the information retrieved from the corresponding header and will repeat this process for all the PCM files and all the headers.

Editing

The necessity to verify the acoustic quality of the segments obtained from the previous stages, together with the possible desire to modify the markers attribution, arose the need of an additional tool able to easily and rapidly edit each segment.
Actually the editing facilities can be distinguished in those only affecting the LPC parameters (pitch, energy and 12 reflection coefficients) and those relevant to the concatenation parameters.
The editing of the LPC parameters is independent on the type of segments; indeed, by means of this facility, entire words or at least complete sentences can be treated.
The tool is able to deal with up to 4096 frames LPC coded, with a frame structure represented with 16 bytes; this limit will vary accordingly to the number of bytes used for a single frame.
In our case we adopted the frame structure reported in fig. 1.

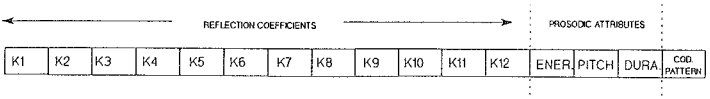

Figure 1. Frame structure.

The definition of the basic frame structure can be simply modified via setup menu.
Normally, each segment comes with its own header, that contains explicit information like phonetic description, original file from which it has been extracted, starting end ending frames, concatenating frames and others as required by the user; it is also possible that the segment is not provided with a specific header; in this case the setup menu can be updated as necessary. Therefore, the capabilities of the tool are not pre-defined, but can be adapted to a large variety of situations.
The basic operations can be performed both at the segment level and at the frame level; in the former case, one can shorten or lengthen the segment, try to concatenate it with other ones, modify the concatenation markers and optionally activate the playback of all these trials.
At the frame level, more insight corrections can be achieved, as for instance the voiced/unvoiced attribute, the energy level, the reflection coefficients; a given number of frames can be deleted and/or inserted in a specific point.
A typical configuration of a portion of the screen is reported in fig. 2 together with some comments.

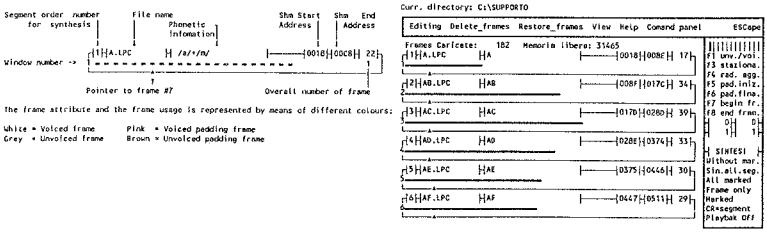

Figure 2. A typical screen layout.

PITCH CONTOURS

Techniques for pitch detection

Accurate and reliable pitch detection is one of the major topics in speech processing; the difficulties comes from different reasons: basically, the waveform corresponding to the glottal

excitation can not be considered rigorously as a train of periodic pulses; furthermore, the characteristics of the vocal tract affect the glottal excitation and finally, when evaluating the actual pitch period, it is difficult to accurately measure the exact beginning and ending of the voiced segments.
Several techniques have been proposed, some based on the time-domain properties of the speech signal, others based on the frequency-domain properties and some more using a mixed time and frequency domains analysis; extensive experiments reported in the literature confirm that the perfect technique is still matter for research and that each approach has advantages and drawbacks.
Therefore, in order to find the best compromise between performances and computational complexity, in respect to our system, we decided to choose the Simplified Inverse Filtering Technique (SIFT). This technique has been extensively tested on our Micro VAX II workstation yielding very good results;.in fig. 3 a block diagram of the SIFT is reported.

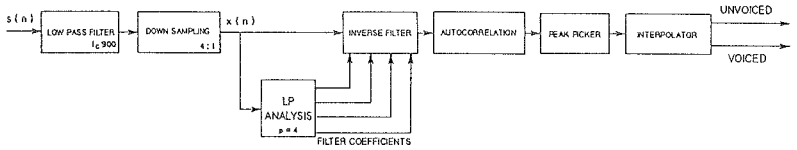

Figure 3. SIFT - Block diagram.

The K-th frame, 320 samples long is low-pass filtered at 900 Hz and then downsampled with a factor of 4; the sampling frequency becomes therefore 2 kHz, that is consistent with the analysis bandwidth.
The speech samples, after the decimation process, are fed to an inverse filter whose coefficients are derived and periodically updated by a linear predictive analysis of 4th order applied on the same samples.
The resulting residual signal is used to evaluate the autocorrelation function; the pitch period is estimated by interpolating the autocorrelation function in the proximity of the peak of the autocorrelation function.
The voiced-unvoiced decision is carried out on the basis of the evaluation of the peak amplitude by using a threshold value that depends on the peak location and on the voiced/unvoiced attribute assigned to the previous frame.
With a sampling frequency of 8 kHz and a downsampling factor of 4, the theoretical range of values for the pitch period is supposed to vary between 64.5 Hz and 400 Hz.
As a matter of fact, experiments done indicated that the reliable range of values has to be restricted to 70 Hz and 350 Hz.

Real time implementation

The most crucial aspects are the voiced/unvoiced decision and, for the voiced frames, the correct location of the peak; moreover, the algorithm has to be able to deal with silence periods that have to be detected and properly handled.
On the other hand, due to the finite accuracy of the DSP used (16 bit, fixed point), sometimes some corrections would be required; however it is possible to reduce these errors by imposing some euristic rule as for example, the pitch variation, as an absolute value, for two adjacent frames cannot exceed a fixed amount; experimentally we found that a typical limit for this variation is that corresponding to an increase or a decrease within an octave; therefore any other value is considered as an erroneous estimation and is automatically corrected.
The real time implementation of the various SW modules involved, in terms of cycles number, computation time, percentage of the frame time and program memory consumption are reported respectively in Table I.

SOFTWARE MODULE	DURATION (msec.)	% WITH RESPET TO A SINGLE FRAME	NUMBER OF WORDS IN PROGRAM MEMORY	NUMBER OF CYCLES FOR A SAMPLE
PCM ACQUISITION	1.382	8.64 %	112	107
LPC ANALYSIS	0.745	4.66 %	972	58
SIFT	1.925	12.03 %	1351	150
TOTALS	4.052	25.32 %	2435	315
FREE TIME	11.948	74.68 %	–	985
	FRAME BASIS			SAMPLE BASIS

Table I. Performances Analysis.

Usage

As known, the final quality of a text to speech system can be greatly improved by the adoption of natural intonation that gives the synthetic voice an attribute of pleasantness and rhythm; sometimes this characteristic affects the user acceptance much more than the acoustical quality itself, especially in real applications.
The SIFT algorithm has been therefore integrated in a very friendly SW module by which an user can display on the computer screen the pitch contour of an input speech signal; up to about 65 seconds of speech can be displayed at each time; the speech signal can be either picked up with a microphone directly from a speaker or, by using a line input from a audio recording equipment or eventually stored as binary file.
The pitch curve can be zoomed, printed and stored; an automatic scaling of the vertical axis takes care to dynamically adapt the frequency resolution. The user can easily evaluate and eventually measure a specific portions of the curve by positioning a couple of cursors.
Some example of the results that can be obtained with this tool are reported in fig. 4 in the case of interrogative sentences for italian. This tool is of course language independent.

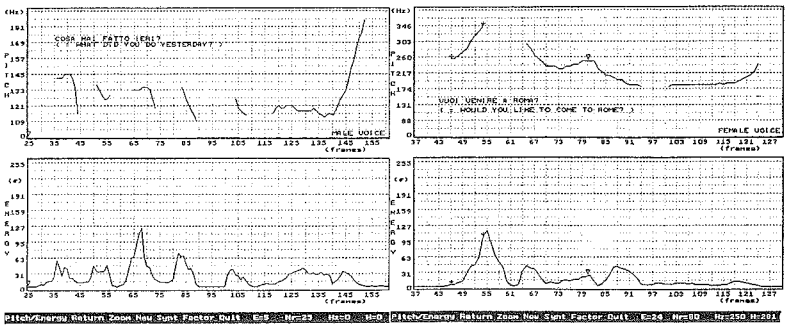

Figure 4. Pitch contours.

CONCLUSIONS AND FUTURE WORK

A complete and powerful workstation for text to speech development has been described; besides the basic advantage of real time operation, compared to the traditional main frames capabilities, the system offers the user a high grade of flexibility allowing the definition of new

parameters and easy addition of further tools for the analysis. These characteristic, combined with the potentiality to track and display in real time the pitch contours, greatly help the researcher and the designer of text to speech systems.
Finally, although the workstation is currently being used for italian language, it can be adopted, without any change, also for other languages.
In the next future other capabilities will be introduced as for example a semi-automatic tool for the verification and correction of the duration of each segment and the realization of a database of prosodic contours.

ACKNOWLEDGEMENTS

The Authors intend to express their thanks to Dr. Enzo Mumolo, now with Sincrotrone-Trieste, who initially started at Alcatel FACE Standard this activity and then continuously supported us with suggestions and discussions.
Thanks also to Dr. Giulio Colangeli and to the whole Hardware Group at our Lab for their efforts to design and made available the DSP board.
This work is partly founded by an ESPRIT (European Strategic Programme for R&D in Information Technology) Project, No. EP 2094, SUNSTAR, Integration and Design of Speech Understanding Interfaces in which ALCATEL FACE STANDARD is involved together with other important European Companies.

REFERENCES

Abbattista, G. & Mumolo, E. (1990) "High Quality Real Time Text to Speech System for Italian Language", Proceedings of VERBA 90 Conference, January 1990, Rome, 343-354.

Hess, W. (1983) *Pitch Determination of Speech Signals - Algorithms and Devices*, (Springer Verlag : Berlin, Heidelberg, New York, Tokio).

Markel, J.D. (1972) "The SIFT algorithm for fundamental frequency estimation", IEEE Transactions on Audio and Electroacoustics, AU-20, 367-377.

Markel, J.D. & Gray, A.H., Jr (1976) *Linear Prediction of Speech*, (Springer Verlag : Berlin, Heidelberg, New York).

Rabiner, L.R., Cheng, M.J., Rosenberg, A.E. & McGonegal C.A. (1976) " Comparative Performance Study of Several Pitch Detection Algorithms", IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-24, No. 5, 399-418.

Rabiner, L.R. & Schafer, R.W. (1978) *Digital Processing of Speech Signals*, (Prentice-Hall signal processing series : Englewood Cliffs, New Jersey).