

MULTICHANNEL TEXT-TO-SPEECH SYSTEM FOR ELECTRONIC MAIL  
APPLICATIONS

Piero Pierucci, Enzo Mumolo and Corrado Labonia

ABSTRACT

This paper describes a real-time implementation of a text-to-speech synthesizer, based on a diphone concatenation approach and running on a single board hardware which is plugged in a PC slot. The system is based on a 68000 microprocessor and two TMS32010; therefore the system can handle two channels simultaneously. The linguistic processing on the input ASCII string (i.e. text to diphone conversion and prosodic processing) is performed by the 68000 while the TMS32010 performs the actual speech synthesis by means of an LPC-12. The system is provided with a telephone interface, giving the possibility of remote listening of texts.

INTRODUCTION

Automatic conversion of written text to speech is useful in many applications, from information retrieval systems to assistance to handicapped people, from dialogue with expert systems to applications in the area of the office automation. In the latter case, the typical working environment is based on some kind of electronic link between users, in order to make possible a continuous exchange of written messages. It may be therefore useful to have the possibility to listen such messages. If a telephone interface is available, it becomes possible to listen such messages in remote mode.

At FACE-RC, a prototype of a real-time text-to-speech system has been developed. The prototype is built around a single board hardware designed for operating into a PC. The board uses one 68000 microprocessor which communicates with the host PC and two TMS32010 DSPs that are used for the actual speech synthesis; the system can therefore handle two output channels simultaneously. The system works in Italian language and is based on a diphone concatenation method. The diphone vocabulary is made of about 150 elements.

TEXT-TO-SPEECH ALGORITHM

One popular method for text-to-speech synthesis is based on the subdivision of speech in diphones. This approach has been proved to be successful in many languages. Diphone elements are commonly defined as speech segments that start in the steady-state center of one phone and end in the steady-state center of the next phone, and contain a complete transition between the two (ref.1). These elements can be actually extracted manually or via computer-aided methods (ref.2,3).

FACE Research Center, via Nicaragua 10, 00040 Pomezia (Italy)

As a matter of fact, it has been demonstrated that, in most languages, a minimal set of diphones can be extracted such that each utterance of the language can be generated by concatenation of them. In the Italian language, for example, linguistic studies have showed that about 150 diphones are sufficient for intellegibile text-to-speech synthesis (ref.4,5). In this case, the diphone set is made of the following phonetic elements:

- the transitions between each consonant and the following vowel (elements /CV/ )
- the initial part of a geminate consonant (elements /CC/)
- the consonants before other consonants not geminate (elements /C#/)
- the initial and steady-state parts of each vowel (elements /V1/ and /V2/)

The diphone concatenation method requires, of course, a preliminary analysis of the printed text in order to obtain the diphone segmentation. Furthermore, some prosodic informations must be extracted from the input string in order to improve the naturalness of the synthetic speech. The prosodic informations used in our implementation refer to the superficial structure of the phrase, namely the punctuation marks and the position of the stressed syllables. With these informations, the pitch temporal trajectory is automatically generated.

The overall architecture of the system is depicted in fig.1.

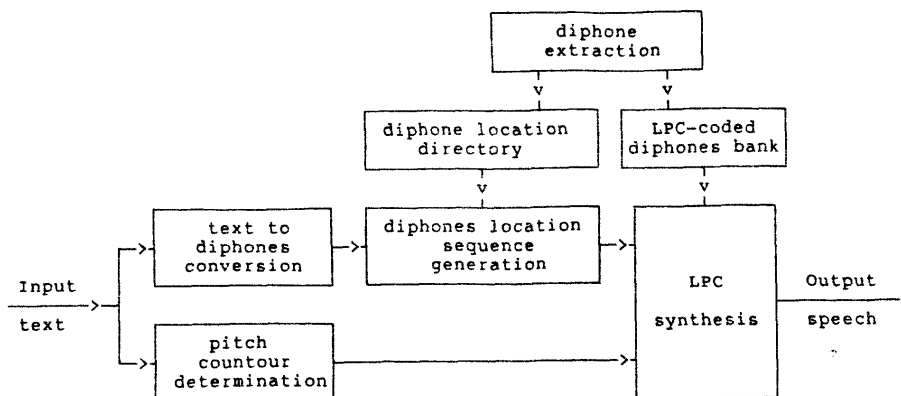


fig. 1

The diphone elements must be extracted from natural speech following some considerations that will be illustrated in the next paragraph. Such elements have been coded according to the speech production model, composed by an all-pole vocal tract filter (ref.6), identified with a LPC algorithm, and a glottal source signal identified with a cepstral technique (ref.7,8). The overall analysis/synthesis algorithm has been intensively tested and optimized and has been proved to be quite robust and to give very good results. The LPC analysis is based on the Durbin algorithm (ref.9). The order of the LPC model is 12; each parameter, namely reflection coefficients, pitch and energy, is coded with 8 bits, giving a total occupation of 14 bytes per frame. The other analysis parameters have been chosen as follows:

sampling frequency 8Khz; Hamming window function; length of analysis frame 300 samples; displacement between frames 150 samples. Assuming that the diphones have been already extracted and located in memory, a directory containing the starting address and duration of each diphone is generated. The output of the text-to-diphones conversion block is the sequence of diphones to be connected in order to obtain the output speech. By looking at the diphone directory table the sequence of start addresses and lengths for each diphone is generated. This sequence is passed to the LPC synthesizer which performs the actual speech synthesis.

#### TOOL FOR DIPHONE EXTRACTION

For each diphone element, the starting position and length must be determined from suitable chosen words. In order to facilitate the identification of the starting and ending points of the diphones, a tool has been developed. This tool, which runs on a PC based workstation equipped with a A/D-D/A conversion board, allows the following operations: complete control of the conversion board in terms of acquisition and restitution of the speech signal; analysis of the acquired

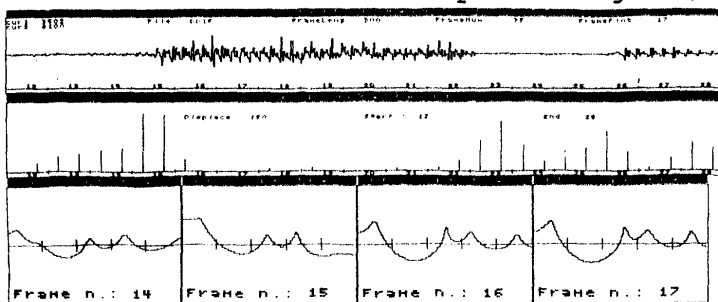


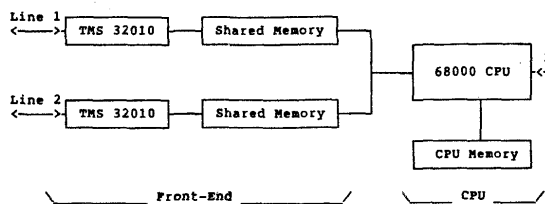
fig. 2

signal in terms of pitch contour and LPC spectrum over consecutive frames; computation of a distance function over selected frames. Fig. 2 shows a typical output obtained with the tool. The word analyzed is "seta" (silk).

The chosen distance measure between frames is the log area ratios distance measure (ref.10), essentially because it allows to compare speech segments independently of energies and pitch, and it can be computed very efficiently from LPC coefficients.

The main utilization of the log area ratios distance function is the identification of the steady-state portion of the vowels. For example it can be seen in fig. 2 that the steady state portion of the sound /e/ starts at frame 17 and ends at frame 21. In this case, the diphone /se/ is taken from frame 12 to frame 17. Each diphone is verified by listening. By using the tool, 150 diphones were extracted and coded in about two weeks.

#### REAL-TIME IMPLEMENTATION



The hardware structure of the system is depicted in fig. 3.

The Front-End section consists of two digital signal processor TMS32010 (one for each of the line that can be handled by the system) which perform the following function: audio line interface management; A/D and D/A conversion; LPC speech synthesis. The Front-End includes the following functional blocks: Codec and PCM interface (the system can interface

both analogue and digital (PCM) telephone inputs); DTMF receivers; PROM/RAM program memory; Shared Memory with control logic. The Shared memory is accessed like an I/O port by the TMS320 using an autoincrement/decrement counter and in the normal way by the CPU. The actual LPC diphone data is contained in the Shared Memory.

The CPU section performs the following functions: handling of the two channels; control of the data transfer with the Host; processing of the input ASCII string to be converted to speech. The CPU contains the following functional blocks: Motorola 68000 CPU running at 10 Mhz; 256 Kwords working memory; Host interface.

#### APPLICATION

To improve the efficiency of an Office Environment, in addition to the traditional telephone subset a number of other telematic features have been introduced (ref.11,12). Communication is the medium that allows contacts among people, usually far apart from each other. Text-to-speech technology can be used in this environment as a tool that allows local or remote listening of texts, such as telex and messages transmitted by electronic mail.

#### CONCLUSION

A real-time system for text-to-speech synthesis has been described. The system is based on a diphone concatenation method and is intended to be used in a office environment. For this reason, the system works with a PC and has telephone interface capabilities for private and public telephone connection. Currently, both an enlargement of the diphone set and an improving of the prosodic processing on the input text are being carried on.

#### REFERENCES

- 1.,2. M Stella, Semi-automatic Constitution of a Diphone Dictionary, in Computer Speech Processing, Prentice Hall, 1985, p.437-439
3. H Kaeslin, A Systematic Approach to the Extraction of Diphone Elements from Natural Speech, IEEE Trans. ASSP, Vol. ASSP-34, No.2, April 1986
4. G L Francini et al, Study of a System of Minimal Speech reproducing Units for Italian Speech, JASA, Vol. 43, 1968
5. C Miotti, S Sandri, C Scagliola, E Vivalda, Unlimited Vocabulary Voice Response System for Italian, IEEE ICC '79
6. J D Markel, A H Gray, Linear Prediction of Speech, Springer Verlag, New York 1976
7. A V Oppenheim, R W Shafer, Homomorphic Analysis of Speech, IEEE Trans. Audio Electroac. Vo. AU-16, No.2, June 68
8. A M Noll, Cepstrum Pitch Determination, JASA Vol.41,1967
9. J Makhoul, Linear Prediction: A Tutorial Review, Proc. IEEE, Vol. 63, 1975
- 10.T P Barnwell, S R Quackenbush, An Analysis of Objectively Computable Measures for Speech Quality Testing, ICASSP 1982
- 11.E Mumolo, P Pierucci et al, Application of Speech Processing to a New Generation PABX, IWASR, May 86, Rome, Italy
- 12.G Colangeli, B Rossi, The DAVIDE Subsystem in a Telecommunication Environment, Globecom '87, Tokio