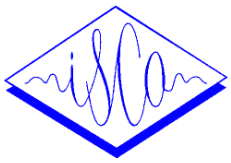

The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

A DATABASE FOR DIPHONE UNITS EXTRACTION

P.Pierucci, G.Ferri, M.Giustiniani
IBM Rome Scientific Centre
Via del Giorgione 159 00147 Rome Italy

ABSTRACT

In this paper we present a speech database for speech synthesis applications. The speech database is tailored for text-to-speech synthesis acoustic units definition, and results in the generation of a database of diphone-like acustic segments. The system architecture integrates the resulting acoustic units database in an LPC diphonic speech synthesizer allowing fast and reliable diphonic units definition and refinement, as well as a general scheme for speech synthesis performance assessment.

1. INTRODUCTION

Speech databases are widely used within the speech community for research and benchmark purposes. Initially speech databases were collected in order to provide a consistent basis for ASR systems performances evaluation. Up to date speech databases extend their applications in many fields of speech processing research, ranging from speech and speaker recognition to speech synthesis systems performance assessement, and their structure is growing in size and complexity. This paper is a report on the use of a general speech database structure for the definition of acoustic-phonetic units for a diphone based text-to-speech synthesizer. The developement of a text to speech system requires the analysis of a large corpus of speech data uttered by a reference speaker. The analysis procedure is made in order to obtain an estimate of the adopted speech production model's parameters. Whether the speech synthesis method is based on a formant scheme or a diphone concatenation approach, the reference speech corpus is collected in order to represent the main acoustic-phonetic events of the language as well as coarticulation effects. It is important that each event would be represented with statistical significance, i.e. more than once, in order to obtain consistent model's parameters estimation. The rest of the paper is organized as follows : section 2 defines the structure of the database system, in terms of hw,sw and available tools. Section 3 describes the diphone units extraction procedure. Section 4 shows the inter-operation of the database with a text-to-speech synthesizer, in order to show the developement capabilities offered by the system.

2. SYSTEM STRUCTURE

The data The structure of the system constitutes a complete speech database with some features added to make it suitable for the particular application. Speech corpus is a phonetically balanced 6180 words set; 10 ms. frame synchronous phonetic labelling is avalaible on a phonetic alphabet for the italian language. Each of 10 speakers provides a repetition of the speech corpus. Speech is sampled at 10 KHz and quantized using linear 16 bit coding. Database structure The data is organized in a relational database making use of the following tables :

Phonetic Alignement Table			
phonetic transcription	sampled data filename	phoneme start pointer	phoneme end pointer
syllable number	word number	take number	speaker id

Speakers Table					
speaker id	name	sex	age	signal conditioning	speech acquisition date

Diphones Table			
phonetic transcription	sampled data filename	origin sampled data id	unit start pointer
unit end pointer	phonetic borders	V/UV regions borders	

The system can be seen as an extension of a general speech database with phonetic transcription; the addition of the DT table makes it suitable for diphone units extraction and management.

Hardware achitecture The relational database structure is implemented on a mainframe computer. This allows the integration of the system in a more general speech database valuable for speech synthesizers and speech coders performances evaluation. As far as the sampled data management is concerned we a dopted a distributed data approach. A number of PC workstations are connected with the mainframe computer via a token ring network. Each workstation manages sampled data speech using an optical disk peripheral. The workstations are equipped with digital to analog conversion boards. Signal conditioning modules are shared among the workstations via a mixing consolle. Figure 1 shows this architecture.

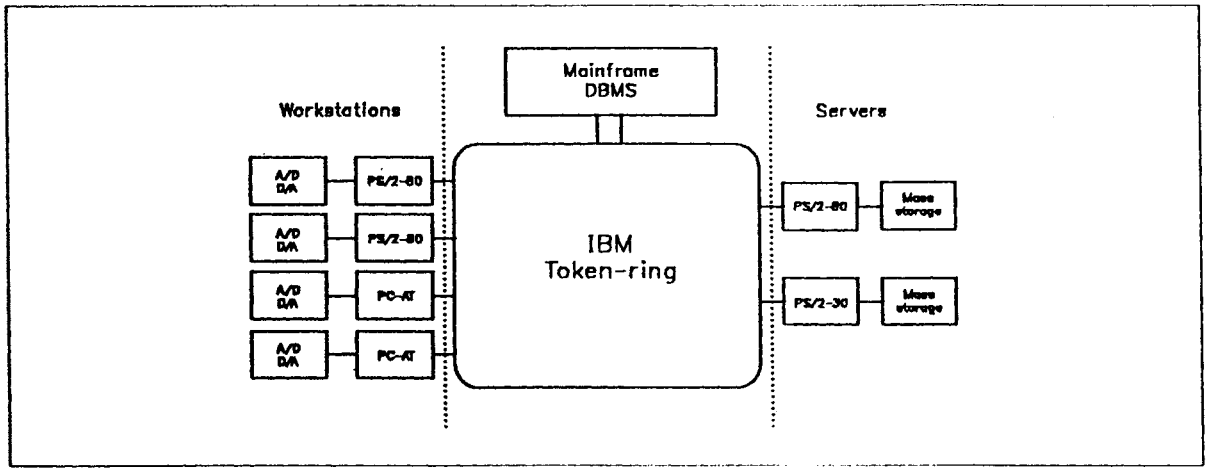

Figure 1.

Software modules An SQL language supports the database query modules implementation on the mainframe computer, under VM operating system. Query modules for diphone-like units are coded in order to search for given phone-sequences in the database. A number of query conditions can be interactively imposed : phoneme duration, position on the word (initial, final, center, stressed/non stressed syllable), and phonetic context (any specified set of phonemes) can be used in order to reduce the search and to obtain selected occurrences in the requested context. Query modules are completed with a set of tools that allows the automatic query formulation for diphone-like units classes (i.e. all the diphones starting or ending with a given phoneme). This set of tools comprise a text editor, an automatic phonetic transcription module, a diphone bank access module in order to restrict the search to not yet extracted units. The results are available to the workstations, running under DOS operating system. At the workstation level the diphone-like unit segmentation from sampled data files is obtained. For each unit to be extracted a list of candidates is available. After segmentation a signal editor module with graphic capabilities and real time D/A conversion from magnetic disk mass storage can be used to score the best candidate. A number of signal processing modules can be invoked in order to ease the scoring procedure. The software is hardware independent, i.e. depending on the workstation processing power the signal processing modules can be made run locally or demanded to the mainframe computer. FFT and LPC spectral tracking, zero crossing, pitch tracking are available. The achitecture is designed in order to allow a concurrent access to the mass storage resources in a multi-user environment. A menu driven interface allows an easy management of the system functions.

3. DIPHONE UNITS EXTRACTION

The diphones unit exctraction procedure starts with the definition of the acoustic units to be used during the synthesis process. We defined the following classes of units :

Diphone-like units			
S .. f	silence to phoneme transitios	f .. S	phoneme to silence transitions
V .. V	vowel transitions	V .. C	vowel to consonant transitions
C/v .. V	consonant or semi-vowel to vowel t.	C .. C	not geminate consonant transitions
C .. C	geminate consonant transitions	f .. L .. f	triphones with liquid phoneme

Each unit is represented as a sequence of vector parameters holding LPC coefficients and a prediction error term σ . No attempt is made to extract units representative of steady state sounds, but an interpolation scheme among ending and starting frames of adjacent diphones is preferred.

For each class the afore mentioned procedure is started. Once the occurrences of the diphone-like unit under consideration has been found by the database query modules the scoring procedure is started. At this level a special tool has been successfully added during the database developement. For the selected speaker a general speech model, referred in the folllowing as EIIMM, is estimated using a phonetically balanced subset of the available speech corpus. This model is able to consistently represent the main spectral events produced by the speaker as well as the phonotactical constraints among them [1], [2]. Using this model it is possible to automatically select the more representative acoustic realizations of the unit and to fasten the scoring procedure.

Once the best diphone-like unit candidate has been found on the list proposed by the query modules, the borders are interactively imposed using the graphical signal editor. Spectral steady state regions are located with the help of the spectrogram and of the EIIMM state sequence track obtained by Viterbi alignement of the speech segment. Finally the voiced and unvoiced regions and the individual phoneme start and ending points are marked. Once the unit characterization is ended a save module can be activated in order to store it in the DT table of the database structure on the mainframe computer.

4. LINK WITH SPEECH SYNTHESIZER

The DT table holds all the necessary informations and files that are necessary during speech synthesis. The speech synthesizer modules are resident on the same mainframe computer, and can directly access the DT table in order to test the effectiveness of the extraction procedure. Sample sentences are synthesized in order to test each diphone-like unit in terms of intellegibility and perceived quality. Figure 2 gives an insight of the synthesizer structure and his interconnection with the speech database.

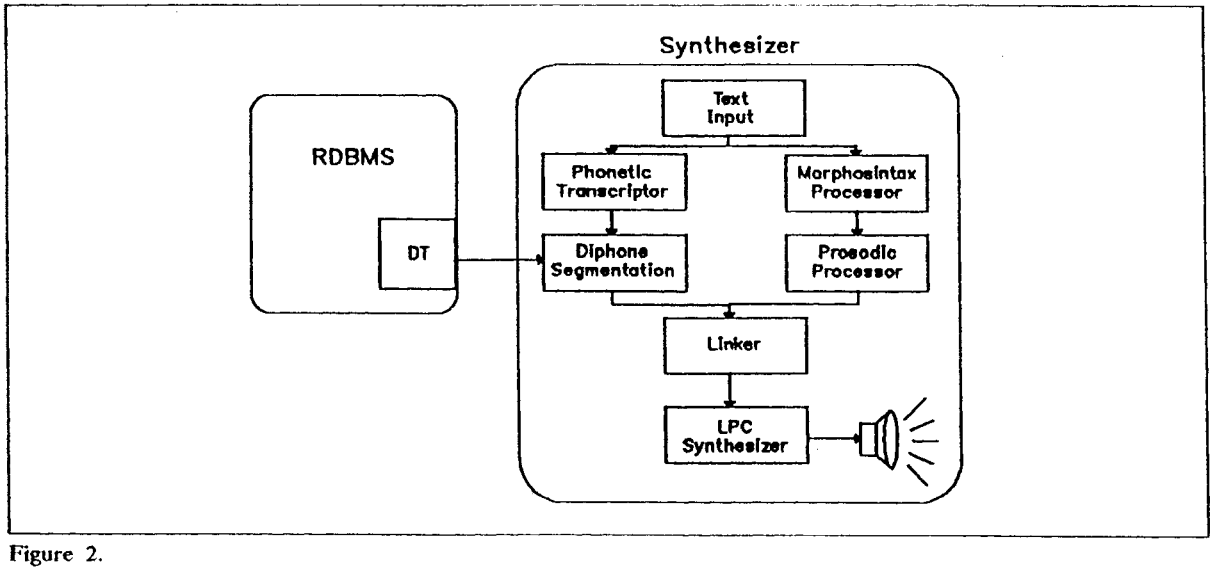

Figure 2.

The proposed structure allows a fast and reliable speech units refinement procedure. As an example of the possibilities offered by the system we will describe the individual phonemes energy contour modelling. The amplitude contour we observe in natural speech is very important for the intellegibility of phonemes like affricates. In general the preservation of the original energy contour of a diphonic unit has benefic effects on the perceived quality of the reconstructed synthetic speech. A phoneme is reconstructed by concatenation of the output transition from a diphone unit with the input transition of the next one, using a suitable spectral and energy interpolation scheme. As far as the spectral interpolation scheme is concerned we adopted a log-area-ratio approach with good results.

The energy interpolation scheme is as following. Using the phonetically labelled speech database we calculate the average RMS energy E_f of each phoneme /f/ of the adopted phonetic alphabet in different syllabic positions, following the approach outlined in [3]. From text input, for instance the CVC /č // a // k /, the corresponding diphone-like units are determined. During synthesis we want to match the energy of phonemes to the average phoneme energy in the same syllabic position, while preserving the energy contour between them. Let's consider the sequence of prediction errors contained in the diphones /č // a / and / a // k / that are relative only to the /a/ phoneme. We calculate the factors k_1, k_2, k_m such that :

$$k_1 = \sum_{i=P_{ca}}^{len_{ca}^{v}} \sigma_{ca}^{i}; \qquad k_2 = \sum_{j=1}^{P_{ak}} \sigma_{ak}^{j}; \qquad E_a = k_m \times (k_1 + k_2)$$

where P_{ca} and P_{ak} are the hand marked borders of phoneme / a / respectively in diphonic unit /č // a / (starting point) and / a // k / (ending point), and len_{ca}^{v} is the total length of diphonic unit /č // a /. The σ of each vector component of phoneme /a/ is rescaled accordingly. The same procedure is applied to other phonemes.

Figures 3 and 4 compare a natural utterance of the speech sounds /č // a /. and a synthetic reconstruction via the outlined method. The amplitude pattern is preserved, giving to the reconstructed speech a more natural behaviour. This procedure gives best results in terms of preceived quality and intellegibility of phonemes like affricates, as confirmed by preliminar informal listening tests. The link between speech database and synthesizer allows the location and refinement of diphone-like unit's parameters in a very straightforward way, and represents a valuable tool to assess the synthesizer performances.

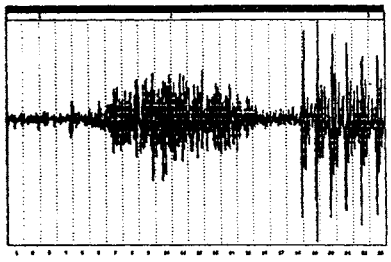

figure 3

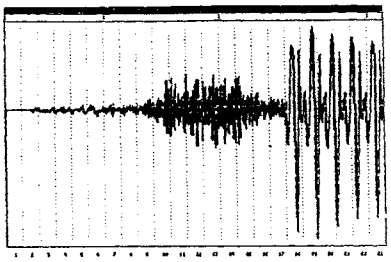

figure 4

REFERENCES

[1] A.Falaschi,M.Giustiniani,P.Pierucci **A finite states Markov quantizer**, *Proceedings of ICASSP, Albuquerque,* 1990

[2] P.Pierucci,A.Falaschi,M.Giustiniani **Phonetic units and phonotactical structure inference by EHMM,** *Proceedings of NATO/ASI workshop, Cetraro,* 1990

[3] A.Falaschi,M.Giustiniani,P.Pierucci **Automatic inference of a syllabic prosodic model,** *Proceedings of ESCA workshop on speech synthesis, Autrans,* 1990