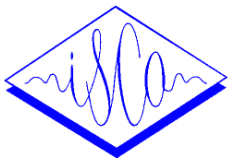

The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

AUTOMATIC INFERENCE OF A SYLLABIC PROSODIC MODEL

M.Giustiniani(**), A.Falaschi(*), P.Pierucci(**)
(*) Universita' di Roma "La Sapienza", INFO-COM Dpt.,
Via Eudossiana 18,00184 Roma, Italy
(**) IBM Rome Scientific Center,
Via del Giorgione 159, 00147 Roma, Italy

ABSTRACT

A syllabic strucutre model is proposed, as a tool for defining a detailed phonemic transcription. The microprosodic features (duration and loudness) of the defined phonological units are estimated by statistical measures on an automatically segmented speech database. The use of these units in a text-to-speech concatenative synthesiser is discussed.

I. Introduction

In order to obtain a good synthetic speech quality, the whole processing chain of the synthesiser should be able to capture the relevant linguistic features of natural speech. A direct way to achieve this goal is to cope with the various aspects in a sequential manner, from the linguistic levels up to the acoustical ones [1], [2]. Object of this work are two basic and stricly related blocks, dealing with the syllabic microprosody knowledge acquisition and its utilization in order to properly concatenate the acoustic segments.

The microprosodic knowledge acquisition is dealt by means of a Syllabic Structure Model (SSM), whose goal is to predict microprosodic variations on a phonological basis. The SSM identifies, for each phoneme, a set of Functional Allophonic Units (FAU), which are considered as different units representing different phonological events according to the SSM adopted. To each FAU duration and loudness typical values are associated, as derived by the collection of statistics gathered by the analysis of a Phonetically Segmented Data Base (PSDB) automatically processed. In the following, after having defined the SSM and the so derived FAU, we will illustrate the procedure for the PSDB analysis, giving also some results. It is then described the use of the loudness and duration FAU parmateres in the text to speech synthesiser currently under development at the IBM Rome Scientific Centre.

II.The Syllabic structure model

In the last years, a syllabic interpretation of many phonetic phenomena is attracting a lot of attention in different laboratiories involved in speech processing [3], [4], [5]. Athough the syllable could seem an ambiguous matter, its phonological definition as an elementary articulatory event constitutes the basis for a very good sovra-segmental theory. In particular, adoption of a SSM allow to integrate phonemic-level segmental characterization of speech in the framework of the immediately subsequent prosodic structure, i.e. the syllable. The SSM here adopted is defined on the basis of the automaton skeleton depicted in Fig.1. The figure reports a set of states, identifying the allowable functional roles played by the phonemes in the

same syllable, and a set of transitions connecting such states, describing the phonotactical constraints on the phonemic sequences and funcionalities.

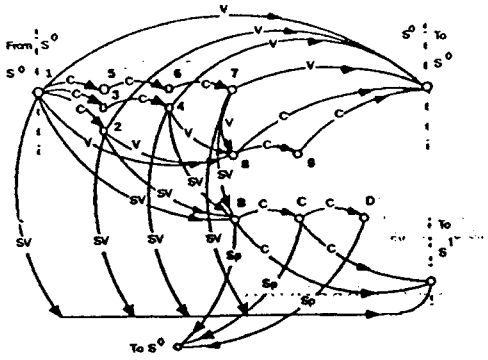

Fig. 1

```
esemplifikazjOne_ <- Phonemes
000000000000001110 <- Stess
122233344556667771 <- Syll #
```

Fig. 2

Syllabic structure model transition diagram Example of FAU transcription

II.1 - The Functional Allophones Transcription

Once a syllabified phonemic transcription of text is available, it can be parsed according to the state indexes defined by the SSM, which acts as an acceptor automata. This analysis brings us to the definition of a FAU transcription of text, in which each phoneme is associated with a set of two indexes, as given in Fig. 2. The first index is assigned according to the position of the syllable containing the phoneme with respect to the lexically stressed one, as after-stress syllables are generally produced less accurately than others.

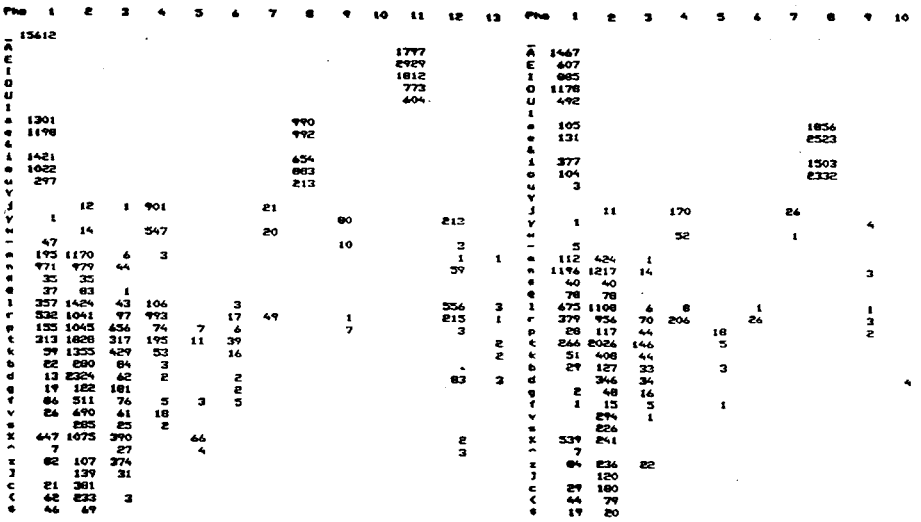

Table 1 Occurence frequencies of the FAUs

The second index is the state number to which the phoneme has been assigned after SSM parsing. In such a way, a vowel is differentely classified according to whether it is stessed or not, is placed in open or closed syllable, or at word end. Consonants will be differentiated on the basis of the length of the cluster in which they occur. After examination of a 9000 words

text database, 217 different FAU had been observed, as reported in Table 1. together with their corresponding frequencies.

II.2 - Microprosodic inference

The above exposed sillable model has been usefully utilized in the implementation of the microprosody control of the experimental text-to-speech synthesiser currently under development at the IBM Rome Scientific Centre. By microprosody control we mean the way to assign the intrinsic duration and energy of each functional allophone, accordingly to the acoustical context. These parameters are computed as exposed in the following. As a by-product of a large vocabulary automatic speech recognition system development, a set of phonemic HMM is obtained. These models can be used as a tool to automatically segment speech into phonemic elements, given its linguistic representation, thus avoiding the tedious and time-consuming manual labeling of data. The result of this procedure generates the PSDB; the HMM phonemic labels are then converted to the FAU transcription.

The voice database actually used is constituted of about six thousands balanced words, uttered by a single speaker. The speech is sampled at 10 kHz; each sample is represented by a 16 bit linear coding. The phonetic transcription of the speech is avaliable. A phonetic alphabet of 39 units has been used; to each phoneme two indexes representing their syllabic position are assigned, resulting in a set of 217 different units.

Phone \ Syllabic State	Open Syll. 01	Closed Syll. 08	Closed Syll. 08	Open Syll. 11	Closed Syll. 18
a	68	74		62	103
i	54	51		47	80
u	63	61		63	77
A			176	193	
I			104	138	
U			103	134	

Table 2.a

Phone \ Syllabic State	Open Syll. 01	Closed Syll. 08	Closed Syll. 08	Open Syll. 11	Closed Syll. 18
a	11.6	10.4		8.0	1.8
i	7.5	5.2		4.3	0.0
u	6.5	5.7		4.6	2.2
A			14.0	13.7	
I			10.5	10.0	
U			10.1	9.7	

Table 2.b

Estimated microprosodic features for three vowels, both stressed or not. A) mean durations in msec; B) mean loudness in dB, relative to the weakest vowel

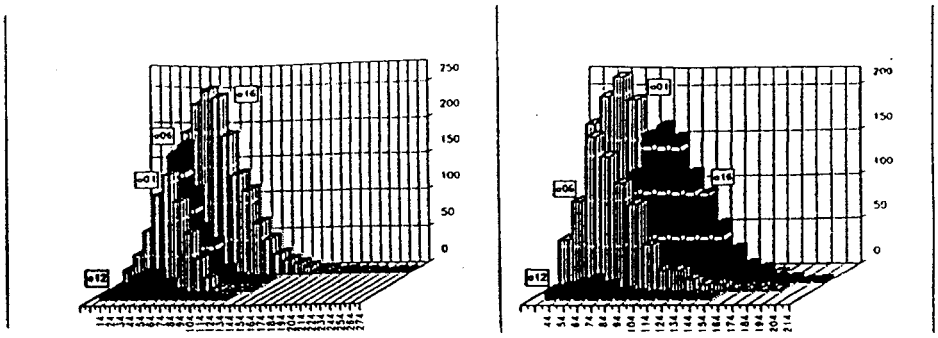

Fig. 3

Duration distribution histograms of vowels 'a' and 'e'.

The data base has been used to automatically compute intrinsic duration of the functional allophones. To this aim, all the durations and loudnesses of the realizations of each functional allophone have been measured; means have then been computed by direct automatic measures on the whole corpus. For each functional allophone the resulting histogram have been evaluated; a sample result is depicted in fig.2 for the vowels 'a' and 'o'; duration distributions are usually roughly gaussians; a chi-squared test performed at 5 per cent of significance on all the functional allophones with more tha 30 occurrences has shown that for more than 71 per cent of the units the measured duration distributions can be considered gaussian. We have not observed bimodal distributions, that could account for a not accurate syllable structure modelling. Steady states length has to be adjusted so that the overall duration fit the theoretic intrinsic durations. This strategy has not been used for triphones, that is for prerecorded sequences of three not divisible phonemes, actually used in some context by the synthesiser; the need of triphones occurs only when the central phoneme does not present a clearly identifiable steady state, so that duration control cannot conceptually be applied. As an example of the estimated microprosodic cues, Table 2 reports the mean durations and loudnesses for three vowels, for the different functional roles they can assume.

References

[1] Klatt D., Allen J., Hunnicutt M. S., From Text to Speech: The MITalk System, *Cambridge University Press*, 1987.

[2] Quazza S., Varese G., Vivalda E., Syntactic pre-processing for high quality text-to-speech, *Proceedings of Eurospeech, Paris, France*, 1989.

[3] A. Falaschi, A functional based phonemic unit definition for, statistical speech recognizers, *Proceedings of Eurospeech, Paris, France*, 1989.

[4] Randolph M., Syllable-based constraints on properties of speech, sounds, *PhD Thesis, MIT*, 1989.

[5] Moulines E. et al, A real-time french text-to-speech, system generating high-quality synthetic speech, *Proceedings of ICASSP, Albuquerque, NM, USA*, 1990.