A RULE-BASED ITALIAN TEXT-TO-SPEECH SYSTEM

S. Barber(*), R. Carlson(**), P. Cosi(***),
M.G. Di Benedetto(****), B. Granström(**) and K. Vagges(***).

(*) INFOVOX AB, Box 2503 s-171 02 Solna, SWEDEN
(**) Dept. of Speech Comm. and Music Acoustics, KTH, Stockholm, SWEDEN
(***) Centro di Studio per le Ricerche di Fonetica - C.N.R., Padua, ITALY
(****) Infocom Dept., University of Rome "La Sapienza", Rome, ITALY

ABSTRACT

An Italian text-to-speech system based on the INFOVOX architecture will be described. With respect to earlier versions of the system, the complete set of Italian vowels and diphthongs was taken into consideration. A set of specific rules for the assignment of word stress, entirely based on statistical considerations, as well as a complete set of grapheme-to-phoneme rules were implemented. Appropriate parameter definitions for the realisation of Italian phonemes were formulated and, on the basis of preliminary results on the study of Italian prosodic structure, appropriate phonetic changes were introduced. Finally, the results of perceptual experiments which were carried out in order to evaluate the system will be described.

INTRODUCTION

The present paper will describe some of the recent work on the Italian language version of the text-to-speech system developed at the Royal Institute of Technology in Stockholm and commercially available through INFOVOX AB in various languages [1]. The general structure of the system, with particular emphasis on some of the most recent improvements, will be outlined. Results on a vowel determination test and on a really severe consonant identification test which were carried out in order to validate the system at the segmental level will be reported.

TEXT-TO-SPEECH SYSTEM COMPONENTS

Figure 1 shows the basic configuration of the system. Previous papers presented in various international conferences could be considered as references [2], [3].

The number rule component (DIG),implements a simple grammar to convert strings of numbers in words.

If the word under inspection is not a string of numbers, a search is first carried out to find if the word belongs to the exception lexicon (LEX). In such a case the word is immediately presented to the phonetic component (FON), otherwise the grapheme-to-phoneme component (GRAF) is applied.

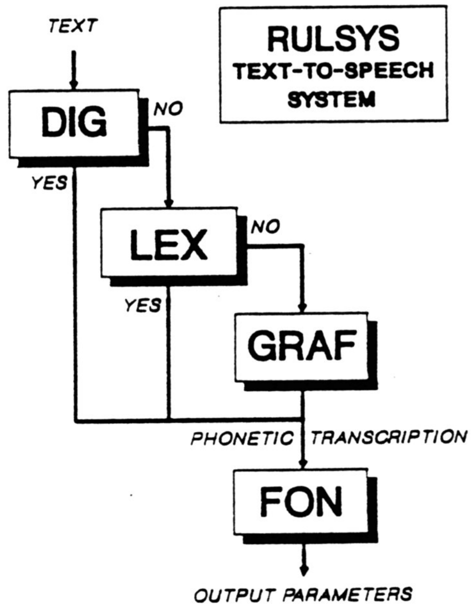

Figure 1. General structure of the text-to-speech system.

The grapheme-to-phoneme module essentially consists of two main sub-components: the stress assignment and the grapheme-to-phoneme conversion. Since the vowel quality and diphthong transcriptions depend on stress position, word stress assignment is the first factor to be taken into account. The stress assignment module is entirely based on statistical considerations focusing on word endings and exceptions. The stress assignment rules were tested on a 10000 most frequent Italian word corpus [4] achieving approximately a 4% error rate. Preprocessing input text obtained with a morphological analysis module could be applied to identify word grammatical category in order to facilitate both word stress assignment and the definition of prosodic rules [5].
As for the grapheme-to-phoneme conversion the Italian language can be considered rather simple. Only few cases need special attention; in particular the graphemes E and O which can be open or closed, the

diphthongs, the graphemes S and Z which can be transcribed as either /s/ or /z/ or /ts/ or /dz/ respectively. A set of rules was implemented [6] and their productivity has been tested on a 10000 most frequent Italian word corpus obtaining an error rate of approximately 8%. An appropriate exception lexicon(LEX) which contains all the cases not correctly transcribed by the rules was automatically created. A user lexicon containing abbreviations, frequent foreign words, names and other special words can obviously be included in the main lexicon.

The last module (FON) deals with the final conversion from phonetic transcriptions to effective commands to the synthesizer.
As for the parameters that drive the synthesizer (formants, durations, amplitudes etc.), each phoneme has its own default definition and, depending on the context in which the phoneme appears, several coarticulatory and prosodic rules are taken into consideration.
An example of the synthesis parameters passed to the speech synthesizer is shown in Fig. 2.
The coarticulatory rules take care of problems such as spreading of nasalization, assimilation of lip-rounding, coarticulation for velars and labials and formant transition lengthening for diphthongs. As for the prosodic rules, duration and F0 variations have been examined. Stressed vowel and geminate consonant duration lengthening, duration shortening of vowel and consonant clusters, as well as duration shortening of vowels preceeding tense consonants and of clause final stressed vowel are some examples of durational rules [7],[8], [9]. Obviously, F0 contour depends on the syntactic structure of the language.

The syntactic analysis is not yet implemented in the current system. A first attempt to automatically extrapolate prosodic groups was based on punctuation marks and function words [10] . Function words were manually inserted in the lexicon and marked in order to be retrived by the phonetic module.

SEGMENTAL LEVEL EVALUATION

A vowel evaluation and preference test, and a consonant identification test were carried out in order to assess the segmental level.

Vowel determination test.

Four different sources have been used to determine the default formant values for the 7 Italian vowels: a) INFOVOX default values, b) vowels studied in a previous investigation on Italian speech synthesis [11], c) reference speaker 1, and d) reference speaker 2.
For each of the four different sets no identification test was needed.
The vowels were generated modifying formant values accordingly to the four different sets. As for the evaluation test, each of the 28 vowels (7 vowels x 4 speakers) was repeated five times and presented in a randomised order. There was a 4 seconds interval between stimuli and 10 seconds every 10 presentations. Five phonetically trained subjects were asked to rate the stimuli on a 5 points scale. As for the preference test, each vowel was paired with itself and with the remaining three productions of the same vowel in both orders, that is, V_1V_2 and V_2V_1, for a total of 16 pairs per vowel. There were two blocks of trials, each trial having

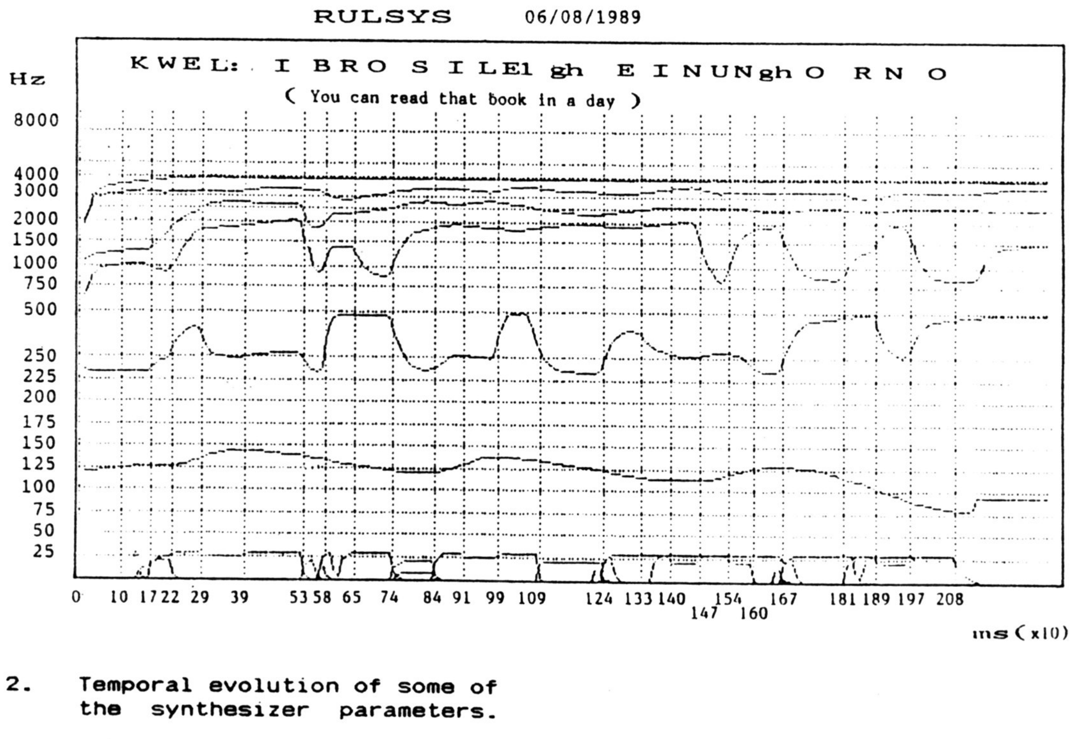

Figure 2. Temporal evolution of some of the synthesizer parameters.

112 pairs (16 pairs x 7 vowels). The vowels within a pair were separated by 1 second, there was a 5 seconds interval between pairs, and a 10 seconds interval every 14 pairs. Five phonetically trained subjects were asked to write whether they preferred the first or the second vowel or whether the vowels in the pair were the same. The responses of the subjects for the vowels /a,i,u/ showed consistency in the two tests, whereas this was not the case for the vowels /ɛ,e,ɔ,o/.

The combined results of the evaluation and preference test gave rise to the final vowel group.

Consonant identification test

The subjects partecipating in the consonant identification test were a non homogeneous group of 14 listeners composed of phoneticians, language therapists and students. Some had previous experience listening to synthetic speech. They were not paid for their services. The stimuli were 63 VCV syllables constructed from the 21 Italian consonants and the vowels /a,i,u/. The two vowels were the same in each syllable. Three different tapes were constructed, one for each vowel. Each tape contained 3 randomized repetitions of the 21 stimuli. The stimuli were recorded with a 4 seconds interstimulus interval and 10 seconds every 7 presentations. The subjects were informed that they would hear synthetic speech and the responses were given by pressing a key on the PC keyboard corresponding to the consonant they heard. Figure 3 shows the correct identification rate for each consonant.

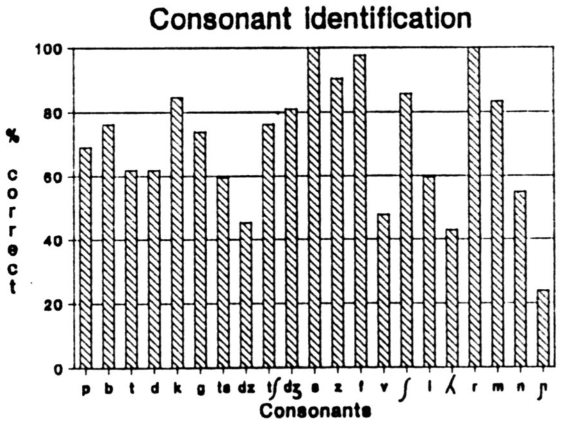

Figure 3. Results of the consonant identification test.

Figure 4 refers to the correct identification rate for each consonant class. It is evident that, even if few problems remain for some affricates, palatal liquids or nasals, the confusions are within the same category class.

In other words the results, in terms of consonants identification percentage, were satisfactory since there were few confusion between consonants characterised by different manner of articulation. Further work will be needed to avoid confusion between consonants belonging to a set grouping consonants with the same manner of articulation, especially in the case of liquids and nasals.

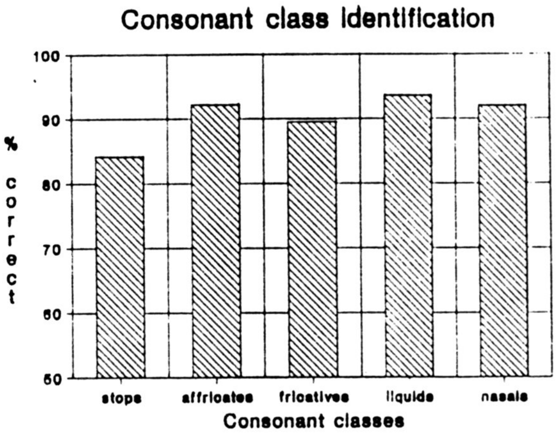

Figure 4. Results of the consonant class identification test.

CONCLUSIONS

The Italian version of the INFOVOX text-to-speech system was described. Main attention was paid to the definition of rules for the following purposes:

- . word stress assignment
- . grapheme-to-phoneme transcription
- . coarticulation and preliminary prosodic structure modeling.

Perceptual experiments, which were carried out in order to evaluate the system at the segmental level, were carried out. The results, in terms of consonants identification percentage, were satisfactory but further work will be needed to avoid confusion between consonants belonging to a set grouping consonants with the same manner of articulation, especially in the case of liquids and nasals.

More complex test, for a complete evaluation of the system will be needed. The study of prosodic aspects of Italian and the formulation of specific prosodic rules will be the main focus of future work.

ACKNOWLEDGMENTS

This work was partially supported by GEPIN S.p.A., Via degli Artificieri, 53 ; 00143 ROME.

REFERENCES

[1] R. Carlson, B. Granström and S. Hunnicutt,"A multi-language text-to-speech module", Proc. ICASSP 82, Vol. 3, pp. 1604-1607, Paris 1982.

[2] R. Carlson, B. Granström, "A text-to-speech system based entirely on rules", Proc. ICASSP 76, pp. 686-688, Philadelphia 1976.

[3] R. Carlson and B. Granström, "Linguistic processing in the KTH multi-lingual text-to-speech system", Proc. ICASSP 86, Vol. 4, pp. 2403-2406, Tokyo 1986.

[4] U. Bortolini, C. Tagliavini and A. Zampolli, Lessico di frequenza della lingua italiana contemporanea, Garzanti, 1982.

[5] B. Granström, P.M. Hansen and N.G. Thorsen, "A Danish text-to-speech system using a text normalizer based on morph analysis", Proc. European Conference on Speech Technology, Vol. 1, pp. 21-24, Edinburgh 1987.

[6] P. Cosi, "A graph-oriented implementation of a grapheme-to-phoneme

mentation of a grapheme-to-phoneme

transcriber for Italian", Speech Comm., Vol. 6, pp. 203-216, 1987.

[7] E. Farnetani and S. Kori, "Italian lexical stress in connected speech", Proc. of the 4th FASE Symposium on Acoustics and Speech, ESA, pp. 57-61, Roma, 1981.

[8] P.L. Salza, "Controllo prosodico nella sintesi da testo dell'italiano" Proc. XVII Convegno Nazionale della Associazione Italiana di Acustica, pp.495-500, Parma, 1989.

[9] R. Delmonte, G.A. Mian and G. Tisato, "A grammatical component for a text-to-speech system", Proc. ICASSP 86, pp.2407-2410, TOkyo, 1986.

[10] S. Barber,B. Granström and P. Touati, "French Prosody in a Rule-Based text-to-speech system", Proc. SPEECH 88, 7th FASE Symposium, Vol. 1, pp. 967-974, Edinburgh 1988.

[11] F. Ferrero, K. Vagges, G. Righini and G.M. Pelamatti,"Un sistema di sintesi dell'italiano: primi risultati", Rivista Italiana di Acustica, Vol. 1-N,1-2, pp. 33-48, 1977.