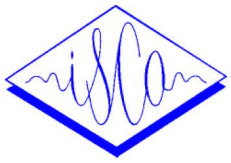

First European Conference on
Speech Communication and Technology
EUROSPEECH '89
Paris, France, September 27-29, 1989

RECOGNITION OF NUMBERS AND STRINGS OF NUMBERS BY USING DEMISYLLABLES: ONE SPEAKER EXPERIMENT.

J. B. Mariño, C. Nadeu, A. Moreno, E. Lleida, E. Monte

Dpto. Teoría de la Señal y Comunicaciones. U.P.C. Barcelona. Spain

ABSTRACT

This communication reports the use of demisyllables for continuous speech recognition in a specific application: the recognition of spanish numbers. After a brief outline of the recognition system, a description of demisyllable syntactic constraints and one-speaker reference generation is provided. Finally, the recognition performance is assessed by means of two experiments: the recognition of integer numbers from zero to one thousand and telephone numbers uttered in a spanish way (strings of integers from zero to ninety nine); in both applications the results that the system yielded were excellent.

I.- INTRODUCTION

Recognition of strings of numbers is an interesting experiment in continuous speech recognition research. In addition to its potentially practical application, recognition of numbers reproduces at reduced scale the general problem of continuous recognition: light acoustic differences may correspond to great semantic differences, numbers exhibit a strong grammatical structure, and there is a relatively important sound variability in the acoustic realization of numbers.

Although in this specific application of speech recognition, words could have been a possible choice as phonetic unit, the system described in this communication is based on demisyllables [1]. This was due to the aim of building a system whose architecture was matched to the general problem of continuous speech recognition, and the conjecture that demisyllable fits properly the characteristics of Spanish language. This conjecture is supported by two facts: a) the syllabification rules for Spanish are well defined [2]; and, b) the inventory of spanish demisyllables is relatively small (less than 750).

In order to define the demisyllable set, every possible syllable was divided by the strong vowel into an initial demisyllable and a final demisyllable. In our definition the prosodic stress was incorporated to the final demisyllable. Accordingly, we distinguished between stressed final demisyllables and unstressed final demisyllables. The main cues of prosodic stress in Spanish are pitch, loudness and syllable length; as pitch and loudness informations are not considered in our system, the main difference between stressed and unstressed final demisyllables will be the length of their references.

II.- OUTLINE OF RECOGNITION SYSTEM

Figure 1 provides a general description of the recognition system architecture. The speech signal is filtered by an antialiasing low-pass filter with a cutoff frequency at fc=3.4 kHz, and sampled at 8 kHz; then the signal is isolated by an end-point detection algorithm and parametrized by a linear prediction filter with 8 coefficients. Preprocessed in that way, the signal enters the recognition algorithm; in essence, this algorithm implements one-stage dynamic programming (described for connected word recognition [3]) driven by a finite state grammar. This grammar affords in a convenient way the demisyllable transcriptions of every item in the vocabulary or language to be recognized. The dynamic programming algorithm computes the bandpass lifted cepstral distance [4] between the demisyllable templates and the test utterance, and decides the grammar-allowed demisyllable sequence that yields the best matching to the speech signal. If necessary, a dictionary provides the semantic meaning of the issued sequence.

This general architecture can be oriented to a specific application by designing the regular

This work was supported by the PRONTIC grant number 105/88.

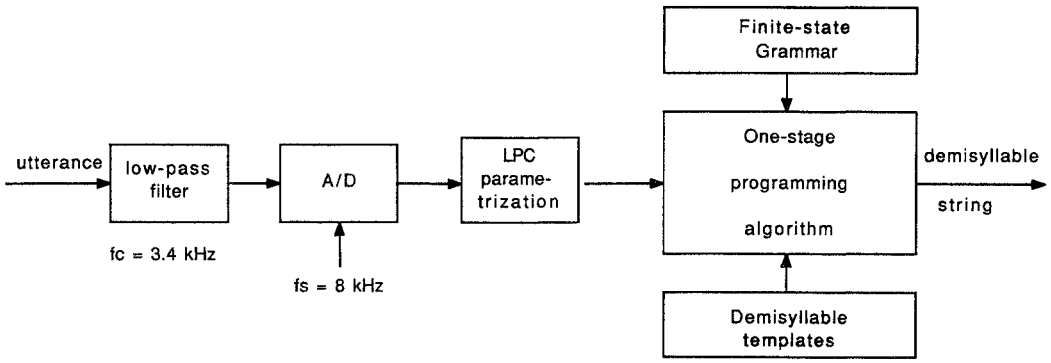

Figure 1.- Recognition system architecture

grammar and training the demisyllable templates. In the sequel, these tasks concerning the spanish integer and telephone number recognition are addressed.

III.- GRAMMAR DESING

Let us consider the vocabularies in Table 1. Vocabulary D corresponds to the digits excluding the zero; vocabulary T includes the zero, the integers from ten to nineteen and multiples of ten; finally, vocabulary P is built by the prefixes necessary to form the integers from twenty one to twenty nine (2...), from thirty one to thirty nine (3...) and so on. The vocabulary I of the integers from zero to ninety nine can be obtained by the following combination of the three previous vocabularies:

$$I = (D + T) + (P * D)$$

where symbol + means union of vocabularies and symbol * denotes cartesian product (every item in P can be followed by every item in D). The grammar G corresponding to I is set up from the grammars of D, T and P by using the tools described in reference [5]. Finally, the grammar of the vocabulary formed by all possible strings of elements of I (telephone numbers are a subset of such a vocabulary) can be derived by joining the initial and final states in grammar G into one only state (self-chaining of grammar G)

In order to describe every item in vocabularies D, T and P as a string of demisyllables a standard phonetic transcription was performed, including the most frequent allophonic variations; thus, one item was allowed to be represented by several different demisyllable sequence. Furthermore, the combination P*D of vocabularies P and D and the self-chaining of grammar G bring new combination of sounds, whose coarticulation has to be considered. Our tool for designing regular grammars is well suited for solving the problem in case. It can either change a subsequence of demisyllables by another (if the coarticulation effect is considered to happen every time that combination of sounds

D = {uno (1), dos (2), tres (3), kωatro (4), θiŋko (5), sejs (6), sjete (7), otʃo (8), nωeβe (9)}

T = {θero (0), djeθ (10), onθe (11), doθe (12), treθe (13), katorθe (14), kinθe (15), djeθisejs (16), djeθisjete (17), djeθjotʃo (18), djeθinωeβe (19), bejnte (20), trejnta (30), kωarenta (40), θiŋkωenta (50), sesenta (60), setenta (70), otʃenta (80), noβenta (90)}

P = {bejnti (2...), trejnta i (3...), kωarenta i (4...), θiŋkωenta i (5...), sesenta i (6...), setenta i (7...), otʃenta i (8...), noβeŋta i (9...)}

Table 1.- Basic vocabularies used to build the spanish integers from zero to ninety nine.

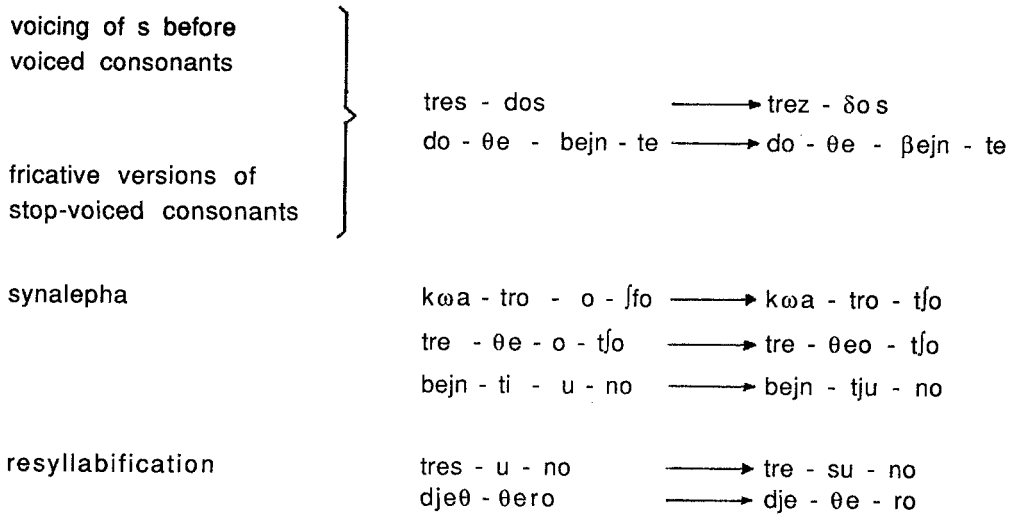

Table 2.- Examples of coarticulation effects considered in the present applications (the hyfhen separates syllables)

appears) or add a new subchain of demisyllables. In Table 2 the main coarticulation effects considered are illustrated with some examples. The final grammar was constituted by 93 states, that accounted for the different contexts of a total of 64 demisyllables.

Following a similar procedure, the grammar of spanish integers from zero to one thousand was set up. For this application we needed 67 demisyllables and a grammar with 118 states.

IV.- TEMPLATE TRAINING

The template training was carried out from a set of 20 telephone numbers and 20 integers designed in such a way that every demisyllable appeared three times at least and the occurrence of demisyllables were balanced with the frequency of its appearance in the integers from zero to one thousand. These 40 speech signals were segmented by recognition with demisyllable templates taken from another experiment; this segmentation was supervised and when necessary corrected by hand.

In this way, a set of samples for every demisyllable was obtained. From this set a template was trained to represent the corresponding demisyllable. The training was accomplished as follows: once the centroid of the set was determined, every demisyllable sample was aligned to it by dynamic time warping and then all the samples were averaged to provide one unique template.

After some previous recognition experiments, it was realized that some minor modifications of this procedure were necessary. It was observed that the length of templates in general did not match the average of demisyllable duration in the test material; and moreover, the templates of some demisyllables did not work well in certain contexts. The most important cases were the following: a) the final demisyllables of open syllables when occurred in the absolute final end of the utterance, and b) the initial demisyllables of syllables starting by a fricative (s, θ) or nasal (n) consonant when the syllable was the first one in the utterance.

To cope with these two problems, the following strategies were developed. First, the average duration of spanish sounds were established from statistical studies that have been carried out by phoneticians [6, 7]; afterwards, the demisyllable duration was determined by adding up the individual duration of each sound in the demisyllable; thereby the length of demisyllables was standardized for a natural articulation rate between 5 or 6 syllables per second. This synthesis procedure provided values for the demisyllable lengths in good agreement with the test material; furtermore, one reason that partially explained the template mismatching in the indicated contexts was found, i.e.: in such situations the duration of a certain sound in the demisyllable is very dissimilar to that of the rest of cases.

The second strategy relates with the sample averaging. The demisyllable samples corresponding to the mismatched context were collected in one set; the rest of samples made

up another set. Once these sets were defined, the template generation procedure went on as previously described, so that one template for each context was trained. Only 7 demisyllables needed this additional template.

V.- RECOGNITION PERFORMANCE

The system performance was assessed during four recognition sessions in which both applications were tested. The speech signals were directly obtained from the speaker by a microphone model vr-230 of Shure (the same microphone was utilized to record the training material in a previous and independent session) in a quiet room; the recognition was performed on line. Fifty telephone numbers and one hundred integers were uttered in each session; the corpus of numbers to be uttered in both applications was chosen following two criteria: one half was selected in order to check the coarticulation effects considered in the training phase, and the other half was obtained randomly; in every testing session the corpus uttered was different. The articulation style of the speaker was assessed for two listeners in each session; the style was qualified as natural as in human beings conversation. The articulation rate of telephone numbers varied from 5 to 7 syllables per second (no intent was made to adopt a particular rate); the articulation rate of integers spanned from 4 to 7 syllables per second, as a consequence of the different length of utterances (the larger is the number of syllables of an integer, the higher is the articulation rate).

The recognition scoring was excellent. No errors occurred when testing integer recognition; the only way to issue misrecognition was to alter the natural articulation rate, either by slowing in excess the articulation rate or by mispronouncing at a very fast articulation rate. As far as the telephone number application is concerned, no errors happened either but when the combination of sounds "...tres (3) trejnta...(3...)" appeared; this concatenation of numbers was confused with "...tres (3) sesenta... (6...)" less than a ten per cent of times. This error can be related with the conjunction in the utterance "trejnta" of two fenomena: an incomplete closure of the vocal tract before the first stop consonant t, and a relatively long vocalic residue between the same t and the vibration r. Perhaps, the use of energy and spectral time variation as additional features would eliminate this source of error.

VI.- CONCLUSION

This communication reports an experiment that supports the use of the demisyllable for continuous-speech Spanish recognition. It has been shown that the recognition of numbers can be accomplished by using this recognition unit, at least for speaker dependent systems. In order to check further the usefulnes of the demisyllable as a recognition unit and appraise its suitability for speaker independent systems, we are currently developing a new version (based on hidden Markov models) for the system described in this communication.

VII.- REFERENCES

[1] A.E. Rosenberg et al., "Demisyllable based isolated word recognition", IEEE Trans. ASSP-31, pp 713-726: June, 1983

[2] J. Alcina and J.M. Blecua, "Gramática española", Ed. Ariel: 1983

[3] H. Ney, "The use of an one-stage dynamic programming algorithm for connected word recognition", IEEE Trans ASSP-32, pp 263-271: April, 1984

[4] B. H. Juang et al., On the use of bandpass liftering in speech recognition", IEEE Trans ASSP-35, pp 947-954: July, 1987

[5] J. B. Mariño et al., "Finite state grammar inference for connected word recognition", Proc. EUSIPCO'88, pp 1035-1038: Sept, 1988

[6] T. Navarro Tomás, "Cantidad de las vocales acentuadas", RFE vol-3, pp 387-407: 1916

[7] T. Navarro Tomás, "Diferencias de duración entre las consonantes españolas", RFE vol-4, pp 367-393: 1918