USING DIPHONES IN LARGE VOCABULARY WORD RECOGNITION

C. Vicenzi - D. Sciarra

Central Research Department
Elettronica San Giorgio, ELSAG S.p.A.
Via Puccini, 2 - 16154 Genova - ITALY

In this paper we present a large vocabulary, speaker dependent, isolated word recognition system with diphones as basic units, so that the training session is much faster and useful for any application. The system, tested on a vocabulary of 910 words on one speaker, gave a word recognition rate of 78%, slightly lower than an Itakura recognizer with whole word templates (WRR=85%).

INTRODUCTION

In a template-matching recognition system for large vocabulary applications, speaker dependence still seems to be an essential requirement for a satisfactory performance. On the other hand the classical and most common approach to isolated word recognition, the whole-word template matching, presents a serious drawback. In fact a training session where the whole vocabulary has to be uttered, even only once, becomes time consuming. Moreover, one or more repetitions of each word will be necessary for the extraction of reliable templates. The only practical solution to the problem is to use some kind of sub-word units to represent words. We chose the diphones, that, in our definition, include transitions between two phonemes, small portions of steady-state sounds and some longer transitional elements embracing three phonemes [1,2]. These units provided good performance in speaker-dependent connected speech recognition experiments with small and medium size vocabularies [3]. Moreover the diphones proved to be robust and economic units, as they are quite invariant with the context and a set of about 300 of them (corresponding to less than 400 templates) is sufficient to cover the whole Italian lexicon.

ISOLATED WORD RECOGNITION USING DIPHONES

The use of diphones is particularly appealing in speaker-dependent applications, as the training session for a new speaker, consisting in the utterance of a set of meaningful sentences, is only few minutes long. By means of an automatic technique [4], a diphone template inventory suitable for any application in the Italian language can then be derived from the collected speech material.

In the language model with diphones as basic units we assume that time warping may be allowed only during stationary diphones; templates for these units consist of a single spectral state, and appropriate lower and upper duration bounds ensure the time alignment capability. No warping is allowed on transitional diphones, whose templates consist of a sequence of spectral states of specified duration.

The model of a word consists of a lattice of diphones, where appropriate duration bounds are associated to each diphone. Alternative paths are present in order to deal with different possible pronunciations or phonetic variations [1]. Building up a word prototype as a lattice of diphone templates gives an accurate representation of the word, that is expected to work as well or better than the relevant whole-word template, as was shown in experiments on small-vocabulary connected word recognition [3]. As an example, similar words should be better discriminated as their representations coincide except for the actually phonetically different portions.

However, in the recognition of isolated words with no syntactic constraints, the use of a general lattice model becomes unsuitable, as the computational load and memory requirements of the decoding strategies may sensibly grow when the vocabulary size increases, making it hard to achieve a real-time performance. A compromise solution may be obtained if we consider that, in a classical isolated word matching, faster strategies can be implemented; in fact, as the speech model within a word consists of a regular lattice of spectral states, the same transition rules can be applied to any state.

Our approach then makes use of a diphone description of word templates in order to minimize the storage requirements, but, during the recognition phase, a spectral state description is recovered to speed up the matching. When building a word template, its lattice representation is translated into as many single path prototypes as needed, each one composed of a sequence of diphone labels and associated duration bounds. Each diphone label is then a pointer to the beginning of the spectral description of a diphone template in a common area containing the inventory. In the current implementation each spectral state description consists of 12 LPC Cepstral parameters computed every centisecond on a 25.6 msec portion of a 10 KHz sampled signal.

When a word prototype has to be matched in the recognition phase, its diphone label sequence is used to fetch the appropriate sequence of spectral states and to build in a work area a synthetic prototype according also to the duration bounds of each diphone. The input word, isolated by an end-point detection algorithm, can then be matched against each expanded prototype using an isolated word recognition approach and producing a cumulative distance score.

In a preliminary stage of our work, two Dynamic Programming algorithms were tested, obtaining essentially the same results. The former is derived from classical Itakura D.P. equations where weights are attached to skip and duplication transitions; the duration of stationary diphones is adjusted to the value that approximately gives the estimated duration bounds for that sound when the 2:1 warping of Itakura's equations is applied. In this way a sort of synthetic whole-word template is built, and the matching strategy loses any information

about the diphones that originated it. The latter algorithm (the one used to carry on the experiments) is more closely related to our diphone language model, as it allows time warping to be performed on stationary diphones only, giving a broader range of compression ratios than the usual 2:1 . In this matching strategy the transition portions of the reference pattern, as well as the mimimum duration portions of stationary diphones, are always completely traversed (no duplication, no skip), while skipping to the next diphone is only permitted on stationary diphones when their minimum allowed durations have already been reached.
The implementation of this technique has shown to be very efficient and less time consuming than the conventional ones; dynamic programming choices are not made at every frame of the reference pattern, but only on limited portions of it, corresponding to the variable length part of stationary diphones.

EXPERIMENTAL RESULTS

The complete approach was tested on the recognition of isolated words from the vocabulary of 910 names beginning with the same consonant "B". In a whole-word template training session made by a cooperative speaker it would take about 4 or 5 hours (with no breaks) to collect a single repetition of the entire vocabulary. Stress effects were not considered.
In our experimentation, a female speaker uttered a set of 36 meaningful sentences in a connected way, which constituted the training speech material for the extraction of the speaker-dependent diphone inventory. This session lasted ten minutes only.
An automatic bootstrapping procedure was then applied to extract the diphone templates: a forced recognition step was employed to determine the boundaries of each diphone occurrence in all the training sentences; the first occurrence of each transitional diphone was chosen as a template, while for each stationary diphone a clustering technique was applied to choose among all its occurrences one or more "representative" ones as templates. Generation of the templates for the words in the vocabulary was then automatically obtained by translating their orthographic forms into corresponding diphone sequences.
Two repetitions of the 910 words of this vocabulary were also collected from the same speaker, and an end-point detection procedure was applied to each word; we will refer to them as SET A and SET B.
In the first experiment a classical Itakura isolated word recognizer was run using in turn sets A and B as test or reference patterns (tests I1 and I2). Both of these experiments, as shown in Fig. 1, gave a Word Recognition Rate of 85%; in both cases, also, in 97% of the times the correct word was classified within the tenth position. These numbers were used as reference scores for the following experiment, where the diphone based isolated word recognizer was tested. Using SET A as test patterns, the diphone based templates gave a WRR of about 78% (see Fig. 1, test D) which is significantly lower; anyway, correct classification score within the N top candidates rapidly converges to that of I1 and I2 tests, indicating that the adopted approach should still be refined in order to achieve a better discrimination among similar words.
In fact, a qualitative inspection of the classification errors occurred, convinced us that, while the diphone language model seems to be adequate, in most cases misrecognition has to be ascribed to local confusion generated by diphone templates for some particular classes of sounds (such as liquids). We believe that a more accurate generation of the template inventory will yield more satisfactory WRR results.
This problem will be the focus of future work, together with the implementation of a sub-system that should restrict the number of word prototypes to be matched by means of a gross preclassification algorithm based on classes of diphones.

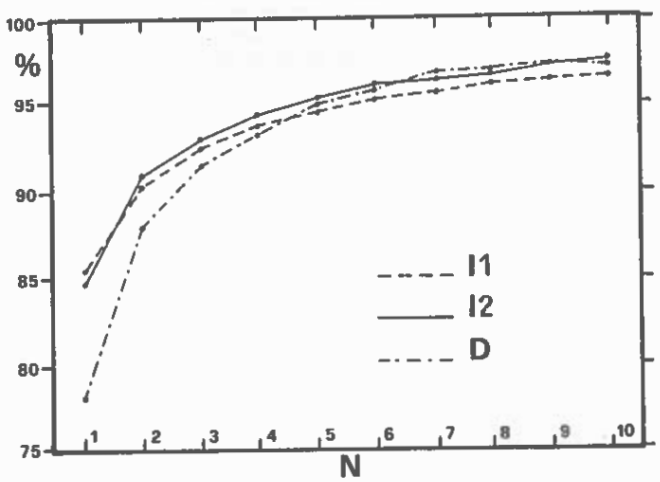

Fig. 1: Word recognition rates within the N (N= 1,..., 10) top candidates in the experiments I1, I2, D (see text).

REFERENCES

[1] A.M. Colla, C. Scagliola, D. Sciarra, "A Continuous Speech Recognition System using a Diphone-Based Language Model", Proc. ICASSP 1985 (31.9), Tampa (1985)
[2] A.M. Colla, "Some Considerations on the Definition of Sub-Word Units for a Template-Matching Speech Recognition System", Proc. Montreal Symp. on Speech Rec., Montreal (1986)
[3] C. Scagliola, "Language Models and Search Algorithms for Real-Time Speech Recognition", Int. J. Man-Machine Studies. Vol.22, n.5, pp. 523-547 (1985)
[4] A.M. Colla, D. Sciarra, "Automatic Diphone Bootstrapping for Speaker-Adaptive Continuous Speech Recognition", Proc. ICASSP 1984 (35.2), S. Diego (1984)