SOME CONSIDERATIONS ON THE DEFINITION OF SUB-WORD UNITS FOR A TEMPLATE-MATCHING SPEECH RECOGNITION SYSTEM

Anna Maria Colla

Elettronica S.Giorgio - ELSAG S.p.A..
Via Puccini 2, 16154 Genova ITALY

Some considerations on the definition of sub-word units suitable for speech recognition are exposed. An example of a kind of units particularly well-suited to syllable-timed languages is presented, together with some hints for the definition of similar units for different languages. Some experimental results are supplied.

FORMAL DEFINITION OF A SPEECH RECOGNIZER

A Speech Recognition System can be considered as a formal system $\mathcal{G}(\mathcal{U}, \mathcal{R}, d)$, where $\mathcal{U}$ is a set of PHONETIC UNITS, $\mathcal{R}$ is a set of RULES for representing each utterance of a given task language $\mathcal{L}$ by means of elements of $\mathcal{U}$, and d is a SIMILARITY OR DISSIMILARITY MEASURE between any "segment" of an utterance and any element of $\mathcal{U}$. More formally we have:

$$d : T \times \mathcal{U} \longrightarrow [0,\infty)$$

$$d\ (t, u) = x$$

where $t \in T$ is a segment of an utterance of approximately equal size as the units, $u \in \mathcal{U}$ is a phonetic unit and x is a non-negative real number. The recognition system can also be considered as an operator acting on a given set S of utterances and yielding for each $s \in S$ an interpretation i(s) in the set $\mathcal{I}$ of all the permitted sentences of the task language $\mathcal{L}$:

$$\mathcal{G}(\mathcal{U}, \mathcal{R}, d) : S \longrightarrow \mathcal{I}$$

$$(\mathcal{G}(\mathcal{U}, \mathcal{R}, d))\ (s) = i(s)\ .$$

Actually $\mathcal{R}$ is a function $\mathcal{R}(\mathcal{L}, \mathcal{U})$ of the language $\mathcal{L}$ on which the system operates and of the set $\mathcal{U}$ of phonetic units. $\mathcal{U}$ also can be regarded as $\mathcal{U}(\mathcal{L})$.

In a Template-Matching system each unit $u \in \mathcal{U}$ is represented by one or more templates expressed in a convenient form (e.g. as vectors or matrices of appropriate acoustic parameters).

DEFINITION OF THE SET $\mathcal{U}$ OF PHONETIC UNITS

Needless to say, the correct choice of the set $\mathcal{U}$ of the phonetic units (hence of $\mathcal{R}$) is of paramount importance for the efficacy of the whole recognition system.

The elements of $\mathcal{U}$ can be words: in this case the definition of the templates is quite natural and $\mathcal{R}$ simply is the set of grammatical rules apt to represent the permitted word sequences in sentences belonging to the task language $\mathcal{L}$; d can be any distance measure between the vectors of parameters (Mel-Based Cepstrum, LPC, and so on) chosen to acoustically represent input and templates. The calculation of d is made more complex by the need of achieving some time alignment between the input sentence and the templates.

In the Speech Recognition System described in [1] the language representation is an HTN [2]; $\mathcal{U}$ is a set of diphone-like units, which we simply have named "diphones"; $\mathcal{R}$ has also to comprise a set of rules to "translate" each word into a net of diphones and to specify the durations of the related events, and d is an Euclidean distance measure between the LPC-Cepstrum vectors respectively representing each time interval of an utterance and of the diphone templates.

The good results obtained in our diphone-based S.R. system are mostly due to the properties of the adopted set $\mathcal{U}$. Quite naturally, the dictionaries of diphone-like units have been designed taking the characteristics of the Italian language into account. The rhythm of Italian is syllable-timed, that is, syllables are pronounced in approximately the same space of time. Therefore units related to syllabic rhythm are particularly well suited to represent the Italian language.

Basic Hypotheses

The complete set $\mathcal{U}$ of the phonetic units we propose for the Italian language has been derived according to the following hypotheses:

- the transitory parts of speech must be as adequately represented as the stationary ones (whilst generally more emphasis is given to steady-state parts, which are longer);
- the units must be short in order to be fairly insensible to coarticulation (hence economical);
- the units must be related to syllabic rhythm (as Italian is a syllable-timed language);
- the duration of "transitory" units must to some extent be related to articulatory time constants.

The Diphones and Their Properties

According to the above hypotheses the diphones [1,2] are very short units: each stationary sound consists of one spectrum, while each transition is represented by a sequence of very few spectra (5-9). This indeed implies a fair insensibility to coarticulation between adjacent units. Therefore each diphone is in principle represented by one template per speaker (notable exceptions are the sounds affected by their position within a word, that is, vowels and sonorant consonants). The set $\mathcal{R}$ of rules is simply deducible from the phonetic strings corresponding to words, by means of a standardized procedure [3] consisting of 4 steps: orthographic-to-phonetic transcription, generation of the diphone sequences, context study for the choice of multiple templates for sonorants, definition of the duration rules.

The acoustic representation of the diphones is obtained by bootstrapping [3] the template(s) for each unit from a rather small training set by means of a forced recognition. The templates have to be well representative of the lexicon, and moreover should not become inadequate because of intra-speaker variability, which can be

serious especially for steady-state sounds. The latter problem can be tackled by a definition of the diphone templates in accordance with a sort of probabilistic approach, where the prototypes are regarded as "average" or "modal" values of a distribution. One "average" template is derived for each steady-state sound and for each different prosodical context of each sonorant. These average templates are used in the same way as normal ones, regardless of the implicit variance.

The diphones have proved to be very effectual for the Italian language. In fact the representation they supply is:

- economical (at most 307 units, with about 350 templates per speaker);
- flexible (that is, apt to deal with pronunciation and duration variability both inter- and intra-speaker);
- automatically deducible for any word from its orthography, including template bootstrapping [3] (this makes the system easily trainable);
- Connected-Speech oriented (straightforward treatment of word coarticulation);
- yielding high scores of correct interpretation (ranging from 82% on the top candidate in a medium-large vocabulary I.W. recognition task, up to 99.5% in a Connected Digit recognition task [1]).

DIPHONE-LIKE UNITS FOR FOREIGN LANGUAGES

It can be hypothesized that, by rules similar to the above ones, adequate sub-word units can be defined also for languages other than Italian, and, in particular, that the units representing similar acoustic events in different languages, being only phoneme-dependent and not context-dependent, can be represented by means of the same templates.

The extension of the recognition system to languages other than Italian can be performed by adapting the different steps in the generation of the diphone representation to the peculiarities of the new language.

In particular the orthographic-to-phonetic transcription must be redesigned for any language, as the phonetic systems of various languages, although partially overlapping, are quite different from one another for a number of reasons: for instance an higher number of phonemes is generally required than for Italian, and above all the orthography is generally much more complex than the Italian one.

On the other hand the rules for the diphone lattice generation need only to be slightly modified, provided that the rhythm of the new language is syllable-based, that is, syllables are entirely pronounced or only their final vowels are not uttered (such as for instance in Spanish, French or German). For languages that are not syllable-timed, that is, languages whose rhythm is governed not by the syllable sequence(s), but by the sequence(s) of strong stresses (such as English or Swedish), the rules for deriving units like the diphones according to our definition are not so straightforward. The use of longer units, spanning the more complex phonetic events pertaining to these languages, is likely to be more appropriate.

EXPERIMENTAL RESULTS AND CONCLUSIONS

The correctness of the above hypothesis has been tested by trying to extend our definition of "diphones" to a language other than Italian and quite dissimilar from it, that is, German. A set of experiments on Connected German Digit recognition has been performed by the same recognition system used for Italian [1]. The test set is made up of 130 1-to-12 digit sentences generated at random, 662 words as a whole. The experimental results are shown in the Table below by the Word and Sentence Recognition Rates (WRR and SRR). Almost as satisfactory results as in the tests on Connected Italian Digits have been obtained both by using entirely new diphone templates (N), and partly re-using "old" diphone templates (O) previously derived for corresponding Italian events (e.g. "AI", "NO", and part of the steady-state sounds). The performance has been improved by submitting the rules of the German diphone lattice generation to some slight refinement, especially about duration of steady-state sounds. By the use of "average" templates for the stationary diphones a further better performance has been achieved (A).

EXPERIMENT	N	O	A
W.R.R.	97.3	96.2	98.2
S.R.R.	85.4	83.1	90.0

Summarizing, satisfactory results have been obtained in a Connected German Digit recognition task by a Template-Matching S.R. System based on diphone-like units as the ones which had proved to be so effectual for Italian [1]. These results show that the extension of such units to languages other than Italian is feasible.

Two are the crucial problems in the definition of diphone-like units for languages other than Italian: 1) the phonetic transcription; 2) the possible need of longer units for stress-timed languages. Moreover a context study is likely to be necessary in order to decide if the same rules for the selection of multiple templates are valid as with Italian.

REFERENCES

[1] A.M. Colla, C. Scagliola & D. Sciarra, "A C.S.R. System using a Diphone-Based Language Model", Proc. ICASSP 1985 (31.9), Tampa (1985)

[2] C. Scagliola, "C.S.R. Without Segmentation: Two Ways of Using Diphones as Basic Speech Units", Speech Communication, 2 (2-3), p. 199 (1983)

[3] A.M. Colla & D. Sciarra, "Automatic Generation of Linguistic, Phonetic and Acoustic Knowledge for a Diphone-Based C.S.R. System", in: R. DeMori & C.Y. Suen (Ed.) "NEW SYSTEMS AND ARCHITECTURES FOR AUTOMATIC SPEECH RECOGNITION AND SYNTHESIS", NATO ASI Series n. F16, Springer-Verlag (1985)