A Functional Based Phonetic Units Definition for Statistical Speech Recognizers

Alessandro Falaschi

La Sapienza University of Rome, INFOCOM Dept, Via Eudossiana 18, 00184 Roma, Italy

ABSTRACT

A functional approach to the description of speech in terms of phonemic units is exposed. It is based on a structural analysis of phonetically transcribed and syllabified texts, which takes into account the word stress and the syllabic structure of the language. As a consequence, the defined units utilize syllabic microprosodic cues and are present in a limited number of contexts. The method is easily applicable to continuous speech, and substantial improvements in recognizers performance can be expected. Finally, the methodology can be advantageously applied to other speech research areas.

I - INTRODUCTION

Statistical speech recognizers are based on the description of the linguistic units allowable phonetic realizations by means of Hidden Markov Models (HMM) of speech [1] [2]. As the vocabulary size becomes larger, a set of sub-word units must be utilized in order to obtain HMMs of each word in terms of a concatenation of these models. Often such units are defined as the phonetic alphabet of the language, augmented by the explicit description of particular events as phoneme clusters [3] or contextual allophonic variations [4]. The choice of these additional units is generally based on the analysis of the recognizer errors, as a general trial-and-errors schema. More pragmatic approaches, based on the definition of a different phonetic model for each phonemic context [5], or for the syntactic class to which the word belongs (word-dependent phonetic model) [6], results in a quite large inventory, giving some problems for training size and practical utilization of the resulting units, letting partly unresolved the continuous speech problem.

Here is presented the definition of a phonetic inventory based on functional criteria, i.e. on the explicit consideration of the role played by the phoneme inside the syllabic an stress structure of the language to be recognized [7]. The assessment of the utility of such a knowledge basis has been proved in previous experiments [8], in which it was utilized as a structured language model for a phonemic recognizer [9].

Three main advantages can be foreseen in utilizing a phonemic inventory based on the syllabic structure. The first is the possibility of an explicit utilization of microprosodic cues, as duration constraints [14]. Second, the considered structure reduces the number of allowable contexts for the defined units, giving some contextual allophonic discrimination as a by product of the structural analysis performed. Third, continuous speech is easily dealt after having defined the word boundary effects in the framework of the syllabic and stress structure considered.

The procedures that will be described are relative to the Italian language; it has to be noted that extension to other languages is quite straightforward [10], if the needed phonological rules are known.

Let us now anticipate how the paper is organized. Sect. II will treat the symbolic pre-processing necessary to obtain a phonetic, syllabified text from some written text, together with the technique utilized for dealing with continuous speech. Then, sect. III will illustrate the considered syllabic and word stress structures. The discussion about the utilization of the proposed phonetic units definition for speech recognition will be undertaken in sect IV. Finally, sect. V will outline the foreseen future application of the structured phonetic analysis proposed.

II - LINGUISTIC PREPROCESSING

This section will illustrate the procedures needed for having a phonetic, syllabified transcription of some written text and which are composed by three consecutive modules, as depicted in Fig.1: a lexical analyzer, a phonetic transcriber and a syllabifier. Such a technique is an almost common one as first stage of text-to-speech synthesizers [11] [12]. Here is suggested its utilization also for recognition systems, with the scope of obtain a transcription of the dictionary in terms of the proposed phonetic units.

By first, as illustrated in Fig. 1, the written text is examined by a lexical analyzer, in order to create a file which will be correctly processed by the subsequent module. The main task of this stage is the stress placement for exception words. As it is known, 'regular' words in Italian have the

![Fig. 1 - Linguistic preprocessing](image-url)
stress placed on the last syllable but one; each word is thus checked with a list of exceptions, by a dichotomyc search, and its orthography modified if the word is not regular, marking the correct stressed vowel. The same list is utilized also for dealing with usual abbreviations, foreign words and initials, which are substituted with the orthographic transcription that will give the correct pronunciation when submitted to the phonetic analyzer. No effort has been spent in this phase of work for dealing with homograph, i.e. word with the same written form but different pronunciation, which are very few in Italian and are usually dealt by means of some syntactical analysis module.

The phonetic alphabet utilized is shown in Fig.2; the set of rules utilized for the phonetic transcription process are given in [13], as it is also for the detailed description of the syllabification procedure. Here it is to be noted that the latter process is strictly tied with the diphthongation one, essential to the correct ascription of the syllabic vowel.

The continuous speech pronunciation is represented by means of apostrophies between words, which have to be present into the written form of the text. This stands for homogeneity with the traditional use of the apostrophe in Italian, utilized as an indication of a compulsory elision of the ending vowel of function words when the following one begins with a vowel. After having solved the diphthongation process inside the word, it is checked if the continuous pronunciation could give syllables spanning the two contiguous words; moreover, in this phase some cases of syntactic doubling process [15] are dealt, as for the case of vowel - apostrophe - plosive sequences. If a vowel meeting is found between words, the ending one is considered reduced and non-vocalic, giving an anomalous ascending diphthong. The only exception to this procedure is for words beginning with unstressed /l/ or /l/, which will give rise to a descending diphthong with the preceding vowel.

The ability of the described symbolic processing method of dealing with continuous speech allows to the subsequent structural analysis module to take into account word junction phenomena according to the syllabic structure that will be described in the following section.

III - STRUCTURAL ANALYSIS

This module is a parser of the symbolic string derived as described above, according with the stress and syllabic structures described by the automates whose state transition diagram is shown in Fig. 3 and 4 respectively. As a result of such a process, each phoneme will have associated to it two structural indexes, which give its position into the structural skeleton.

The stress index is a binary one and refers to the states of Fig. 3: index 0 stands for syllables preceding or containing the stress, and index 1 stands for syllables following it. The symbols drawn on the transition diagram represent either a syllable boundary, marked by the <Sp/> symbol, or a word boundary, marked by the <Sp>/ or </> sequences, in the cases of isolated or continuous speech.

Both $S^0$ and $S^1$ have the inner structure reported in Fig. 4, whose states account for the allowable phonetic sequences which can occur inside a syllable. It can be present (or not) an initial consonantal cluster composed of up to three phonemes, followed by a (stressed or not) vowel which directly reach the following syllable or, in the case of closed syllables, is followed by a final consonantal cluster. The $S^1$ structure differs from the $S^0$ one for the missing of the stressed vowel path, allowing the emission of the space.

![Phonetic alphabet](image)

![Word Stress Structure](image)

![Syllable structure](image)
symbol after a vowel occurrence. For sake of clarity, fig. 4 does not show the apostrophized transitions utilized for dealing with continuous speech; here is sufficient to say that they allow to split the common syllable between two words among the states of two different syllabic structures, placing the terminating and initial phonemes of two continuously uttered words into the correct functional syllabic state.

The structural analysis, besides the stress index placement, parses the phonemes to the same syllable according to the state numbers reported in Fig.4. All the semi- and reduced vowels are treated as consonants, letting them to belong to consonantal clusters, making evident their functional equivalence.

It is now mandatory to highlight how the syllabic structure of Fig. 4 could perform an automatic discrimination among all kinds of vowels (stressed, unstressed, in open or closed syllable, on word end) and on the consonantal type (intervocalic, clustered). Such a behavior is the key of the allophonic units definition here stated.

The syllabic structure here exposed slightly differs from the one previously presented in [9] [13] by the explicit discrimination of the initial cluster consonants based on the cluster length, in order to better account for microprosodic timing effects [14].

Fig.5 reports some example of the exposed processing schema: there it is given the original text, its syllabified phonetic transcription and the structural indexes determination as given by the structural analysis. When applied to a corpus of about 6000 different words, the structural analysis gives the identification of about 200 different functional allophones; this figure will rise of about 30 units when words are analyzed as continuously uttered.

IV - UTILIZATION FOR RECOGNITION AND DISCUSSION

The above exposed structural analysis method for phonemic sequences as yet been proved to be very useful in the framework of a phonemic recognizer [8]. In that circumstance, a stochastic, structured source model for the phonetic message has given a phonetic recognition rate improvement of about 3% with respect to the utilization of a stationary phonetic Markov model. In that case, the structural analysis performed associate to each phoneme also the index of the syllable number to which the phoneme belongs; the recognition performance improvements obtained was only due to the phonotactical and morphological constraints considered. In fact, that experiment utilized a set a 40 phonetic HMM to represent all the functional allophones.

It seems that the use of a different HMM for each functional phonetic unit found could give a further recognition rate improvement, once the training material would be adequately augmented in order to contain more occurrences of each unit. Unfortunately, this kind of experiment has not been undertaken up to now, partially a cause of the unavailability of a possible co-author.

The possible improvements could come either from the reduced number of contexts for each unit or from the implicit microprosodic cues utilization; in the latter case some additional improvement could come from the utilization of a a duration density for the accurate modeling of the timing constraints, as shown in [8].

Again, in the framework of a phonetic recognizer, the ability of dealing with continuous speech could be very useful when utilized jointly to the structured phonetic source model exposed in [9], because of its ability of giving a segmentation in words of continuously uttered phonetic sequences [8].

Moreover, also word-based statistical speech recognizers can obtain better recognition scores by utilizing as basic speech representation units the one here presented, due to the yet exposed better modeled phenomena.

V - FUTURE RESEARCH

The recognition experiments utilizing different phonetic HMMs for each of the defined units has to be absolutely carried on, in order to definitively assess the advantages of such an approach.
As yet anticipated, the presence of an higher number of phonetic units could give some problem for the necessary augmentation of the training corpus size, in order to have an adequate number of occurrences of each units. This can be partially solved by means of a procedure able to extract, from a large corpus of phrases, a subset containing all the units in all their allowable contexts, while keeping the total number of phrases minimum. Such a procedure will be based on the structural analysis here exposed, thus directly dealing with continuous speech.

Finally, the author has intention of utilize the described method also for speech synthesis, obtaining a set of functional allophones directly related to the microprosodic structure of speech and thus able to give a very natural synthetic quality. In order to automatically extract the acoustical and duration cues related to the exposed phonetic inventory definition, the recognition and synthesis problems could faced jointly, deriving by first some HMMs for the proposed units, and then utilizing such HMMs for obtain a segmentation of speech into the defined units.

VI - ACKNOWLEDGEMENTS

The author wish to thank Miss Ivana Emanuele, for having compiled the pronunciation exception list utilized into the lexical analyzer; Miss. Maria Gabriella Di Benedetto, for some useful phonological discussion; Mr. Massimo Giustiniani, for his help in solving systemistic problems. He finally thanks Mr. Marco Ferretti, for his opening to the scientific dialogue.

References