DEFINITION OF RECOGNITION UNITS THROUGH TWO LEVELS OF PHONEMIC DESCRIPTION

M.Cravero, R.Pieraccini, F.Raineri

CSELT - Centro Studi e Laboratori Telecomunicazioni S.p.a. - Via G. Reiss Romoli 274 - TORINO (Italy)
Tel. + 39 11 21691 - Telex 220539 CSELT

ABSTRACT

In this paper a development system allowing the definition of different recognition unit sets is described. It takes into account acoustic, phonetic and phonologic knowledges. Such a system can be easily used to transcribe large lexicon into recognition units, starting from the ortographic form of the words. In the following a detailed description of the formalism used is given, along with some experimental results obtained by our unit set.

1. INTRODUCTION

A recognition unit set must include a certain number of informations belonging to different knowledge sources. Our recognition system, developed within a speech understanding project partially supported by ESPRIT Project No.26, takes into account the following:

a. Acoustic knowledge, i.e. the knowledge needed to hypothesize, recognize or verify an acoustic event by observing a set of features extracted from a speech segment.
b. Phonetic knowledge, that is the ability of dealing with the acoustic events and their relation to defined phenomenon classes (i.e. phonemes).
c. Phonological knowledge, namely the capability of transcribing each higher level segment (word, sentence) by means of the abstract categorization defined at the phonetic level.

In our system, the acoustic level is implemented by means of Hidden Markov Models (HMM); it means that each unit is described by an HMM in terms of number of states transition and emission probability matrices that are estimated with the Forward-Backward algorithm [1].

The other two knowledges are used to represent whatever Italian word in terms of basic units by means of a rule system that includes main phonetic and phonological variations. That interface between the acoustic knowledge (HMMs of units) and the lexical one is realized by a system based on two levels of description; the first one is the standard phonemic form of words along with additional forms accounting for inter-speaker variations. The second level is a description of each phoneme (the Underlying Phonemic Structure or UPS) by means of smaller units; they are mainly stationary segments and transitions [4]. Besides, a set of contextual rules handles the final transcription of a word in terms of stationary and transitional units.

This development system was designed to define an optimal unit set whose performance was experimentally evaluated within a recognition system. The optimal set proved to be a trade-off between phonemes and diphones; when the transition between two sounds is considered significant for the recognition of the two sounds themselves (i.e. plosive followed by sonorant), the corresponding diphone is included in the set, otherways the transition model is realized appending the two phonemic models.

2. PHONETIC TRANSCRIPTION

A module involved in the task of transcribing a lexicon into the corresponding defined elementary units must first translate an utterance from the ortographic form into the corresponding phonetic one. Italian language[2], as many others, has not an ortography faithful to the phonetics, in the sense that to each grapheme can correspond more than one phoneme, and some phonemes can be indicated by two graphemes (for istance the ortographic sequence "gl"can represent the unique phoneme ʎ of the IPA alphabet, or can be pronounced as the plosive "g" followed by the lateral "l").Besides that ambiguity inherent in the language, other problems arise: people coming from different italian regions pronounce some words in different ways (i.e. the phoneme "s" of the word "casa" (house) is pronounced as a voiced phoneme by the northern people and as an unvoiced one by southern people).Moreover each speaker has their habits in the pronounciation of some words (for istance a schwa can be added or not to a word ending by consonant). These considerations suggested the idea of implementing a semi-automatic transcription: in the phase of lexicon creation, the operator introduces the new words one at a time; if an ambiguity is pointed out, all the possible trascriptions of the utterance are created and the manual intervention is required in order to decide if all these sequences are representative of the word (different pronounciations) or if some of them must be excluded being wrong.

3. UNDERLYING PHONETIC STRUCTURE

As said before the lower level of phonetic description consists in the so called Underlying Phonetic Structure (UPS); the idea is to transcribe each phoneme into a sequence of elements (Underlying Phonetic Elements or UPE) which show roughly uniform acoustic characteristics. Incidentally the alphabet used to describe UPS is the same as the phonetic one: while at the higher phonetic level each symbol represents a whole phoneme, at the lower UPS level a symbol represents a phoneme portion. To associate a UPS to each phoneme we use a set of rewriting rules as shown in Table 1, where the plus "+" symbol has the meaning of transition from the preceding or to the following phoneme; so the rule a=+a a a+ means that the phoneme a (on the left of the production) can be translated into a left transition (+a), a stationary portion (a) and a right transition (a+). In Table 1 a complete UPS for the Italian phonetic system is reported (the semicolon indicates geminate consonants). Notice that unvoiced plosives are translated as silence "-" plus transition to the following sound while voiced ones as stationary portion (the voicebar "b") plus transition.

ʃ = ʃ	l = +l l	b; = b; b;+
ɛ = +ɛ ɛ	m = m	dz; = dz; dz;+
ɔ = +ɔ ɔ	n = n	ts; = ts; ts;+
~ = ~	o = +o o	s; = +s; s;
ŋ = n	p = - p+	k; = - k;+
ʎ = ʎ ʎ+	s = +s s	t; = - t;+
ɱ = n	t = - t+	tʃ = tʃ tʃ+
ɲ = ɲ ɲ+	u = +u u	l; = +l; l;
r = +r r r r+	v = +v v v+	m; = m;
a = +a a	w = +u u u+	v; = +v; v; v;+
b = b b+	z = z	tʃ; = tʃ; tʃ;+
d = b d+	ʎ; = ʎ; ʎ;+	ʃ = ʃ
e = +e e	d; = b; d;+	dʒ; = dʒ; dʒ;+
f = f	n; = n;	r; = +r r; r; r+
g = b g+	ɲ = ɲ; ɲ;+	dʒ = dʒ dʒ+
i = +i i	f; = f;	dz = dz dz+
j = +i i i+	g; = b; g;+	ts = ts ts+
k = - k+	p; = - p;+	

Tab.1 - UPS for the Italian phonemes

Each phoneme is represented by means of a single UPS which is constituted by a sequence of UPE.In this

way segments of different phonemes showing acoustic similarities can be treated by the same statisical model, as the voicebar of the voiced plosives.

The translation of a word from its phonetic form to its description in terms of recognition units starts with the translation of each phoneme into the corresponding string of UPE. For istance, according to table 1, the italian word APPARTIENE, rewritten by the ortographic to phonetic module in the sequence /ap;artjɛne/, can be translated into:

+a a - p;+ +a a +r r r r+ - t+ +i i i+ +ɛ ɛ n +e e

The second step detects where the transitions are possible; the rule to obtain a transition consists in merging two UPE's containing the symbol "+" in adjacent positions into one transitional unit. So, following the previous example, we obtain:

+a a - p;a a +r r r r+ - ti i iɛ ɛ n +e e
a - p;a a r r - ti i iɛ ɛ n e

It must be noticed that defining the UPS of the generic phoneme /x/ as x = +x x+ it comes out the classical diphone definition, while rewriting each phoneme by itself as x = x, we obtain the phoneme definition.

At this point the description of the word can be handled by a set of rules to take into account the possible effects of a particular phonetic context that cannot be catched by the generic UPS.

4. CONTEXTUAL RULES

Contextual rules can be expressed in the following general form:

U1_U2_....._Un=W1_W2_.....Wm

where Ui and Wj are generic recognition units and the production means that the sequence of units Ui(i=1,2...n) is translated into the sequence Wj(j=1,2...m). In our system rules are applied sequentially, in the given order, to the whole word. Table 2 gives an example of a rule set. From the third to the 18-th production, rules to obtain the stationary portion of /r/ only when it is in a non intervocalic context are described. The UPS of /r/ is made up of two consecutive stationary portions (+r r r r+); in fact, being impossible in the Italian language to utter an /r/ between two consonants, these rules make each vowel cutting away an /r/, so obtaining the desired transcription. The rules dealing with /v/ permit to define left transitions only for those /v/ inserted in a left vocalic context.

The rules 1 through 4 make the two vowels /o/ and /ɔ/ be represented by the same symbol /o/ as well as the two vowels /ɛ/ and /e/; this is done because of the acoustic similarity of the sounds and due to the fact that in Italian the use of the two o's and of the two e's depends on the speaker habits.
Finally the rule 17 transforms each geminate into the corresponding singleton as we demand the distinction between them to higher levels of knowledge.

1: #ɔ=#o	9: r_ru=ru	17: #;=#
2: ɔ#=o#	10: a_r=a_ar	18: a_v=a_av_v
3: #ɛ=#e	11: e_r=e_er	19: e_v=e_ev_v
4: ɛ#=e#	12: o_r=o_or	20: i_v=i_iv_v
5: r_ra=ra	13: i_r=i_ir	21: u_v=u_uv_v
6: r_re=re	14: u_r=u_ur	22: o_v=o_ov_v
7: r_ri=ri	15: r_rj=rj	
8: r_ro=ro	16: r_rw=rw	

Tab.2 - Contextual rules

Extending the rules to the previous example it can be easily obtained:

a - pa ar r - ti i ie e n e

This formalism, developed in order to easily transcribe large lexicons into recognition units given different unit definitions (included "phonemes" and "classical diphones"), was implemented by a program whose output is compatible both with the HMM training procedure and with a set of recognition and word verification systems.

5. PERFORMANCE EVALUATION

Recognition experiments [3] suggested that the best set of units is made up of 123 elements, precisely 22 stationary units and 101 transitional units. Hidden Markov models were trained by means of a 989 words vocabulary obtaining an average recognition rate of about 83% in isolated words belonging to vocabularies of monosyllables differing only for one phoneme (e.g. /aba/, /ata/, /aka/, etc.). Table 3 shows the correct recognition rate per phoneme.

b	87	z	93	dʒ	41
d	76	l	96	tʃ	67
g	90	r	77	j	63
t	70	ʎ	83	w	100
k	96	m	25	e	93
p	96	n	67	i	100
ʃ	96	ŋ	70	o	90
f	83	dz	96	u	100
v	41	ts	100	a	100
s	100				

Tab.3 - Correct recognition rate per phoneme.

6. CONCLUSIONS

A formalism was introduced to write a flexible system that permits the definition of a recognition unit set and the corresponding transcription of words and sentences from their orthographic description to a form that directly relates to the acoustic models of the units themselves. That is obtained using two levels of definition; the first one specifies the phonemes that constitute an utterance, while the second one splits each phoneme into stationary and transitional portions. A suitable set of units that relies on that concept was defined and tested obtaining encouraging results.

7. REFERENCES

[1] Baum, L. E., Petrie, T., Soules, G. and Weiss, N., "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains", Ann. Math. Stat., 41, 164-171, 1970

[2] C. Tagliavini, A. Mioni "Cenni di Trascrizione Fonetica dell' Italiano", Patron Ed., Bologna 1983. (In italian)

[3] M. Cravero, R. Pieraccini, F. Raineri "Definition and Evaluation of Phonetic Units for Speech Recognition by Hidden Markov Models", Proc. of International Conference of Acoustic Speech and Signal Processing 1986, April 8-11, Tokyo, Japan.

[4] A.M. Colla, C. Scagliola, D. Sciarra, 'A Connected Speech Recognition System Using a Diphone-based Language Model', Proc. of International Conference of Acoustics Speech and Signal Processing 1985, March 26-29, Tampa, Florida