THE PRINCIPLES OF PHONETICAL STRUCTURING OF VOCABULARY FOR SPEECH RECOGNITION SYSTEM

Valeria Kuznetsova

Department of Philology, Moscow State University
Moscow, USSR, 119899

ABSTRACT

In the present paper the problem of decoding the results of the first stage of speech recognition into vocabulary units is discussed. The open syllable is proposed as the basic element for such decoding. The final decision is made with consideration both lexic and phonetic context. The context function is carried out by specially organized vocabulary module in the system.

INTRODUCTION

Lately the problem of mapping the results of preliminary acoustic analysis onto linguistic units draws great attention of different researchers. This problem is very important both for the description of the model of human speech perception and for developing the system of automatic speech recognition and understanding.

The purpose of the present paper is to suggest the solution of this problem in relation to the system of automatic speech recognition. Not discussing in detail the problem of human speech perception, we adopt the following starting-point hypothesis:

I. The decision about signals phonetic content is made for elements corresponding to syllables.
2. Until the content is correlated with the semantic meaning of the unit it is considered to be preliminary and is represented by a limited set of variants or by generalized phonetic content.
3. To arrive at the final interpretation of the signal (to correlate it with some vocabulary unit and to define its phonetic composition more accurately) multifold strategy is implemented on the basis of the information supplied by phonetic and higher levels of analysis.

The present paper deals with the problem of the phonetic structuring of vocabulary module, so without taking into consideration higher levels of linguistic analysis we'll describe some possible model of transition from signal representation (in terms of the first stage alphabet) to vocabulary units.

PHONETIC SYLLABLE RECOGNITION

The basic element of our recognition model is an open syllable. The selection of this unit is supported both by the acoustic-phonetic literature data regarding it to be the minimal unit of speech perception and production /I/-/5/ and by the possibility of automatic segmentation of the results of the first stage recognition into elements corresponding to open syllables /5/, /6/.

The results of the first stage recognition presented in /6/ were used to test the model's reliability. Signal, corresponding to syllable, automatically having been singled out and recorded in terms of the first stage alphabet (FSA) is compared with syllable sample (SS) from the system's memory. Each SS is correlated to phonetic syllable. Thus the result of the first stage recognition goes into the input of the given submodule while in the output there are syllables in phonetic transcription.

These SS were designed on the basis of analysis of the results of the first stage recognition of the definite system with regard to possible within-syllable coarticulation and the duration of syllable's constituents. Thus SS are in their nature idealized, generalized concept of the results of the first stage recognition and are recorded in terms of FSA. The SS set is determined by the requirements put on the recognition system vocabulary. It is rather small in case of limited vocabularies. For evolving systems of automatic speech recognition with extensive and unlimited vocabularies the SS set must be compiled with regard to syllable statistics. The existing syllable statistics for Russian speech /7/, /8/ do not fully answer the requirements of this problem as they are received on the basis of idealized transcription of written texts. Contrary to the statement in /I/ syllables constituting these statistics cover no more than 60% of different type oral texts, as it was shown in our experiment. Thus taking the statistics presented in /8/, as the starting point we are now compiling a fuller statistics that would comprise up to I000 open syllables revealed from the recordings of different types of oral texts. This statistics would supply the basis for SS set of the speech recognition system with extensive vocabulary in which every syllable would get SS representation. The model was tested with SS set of I00 syllables and vocabulary of 200 words.

The syllable corresponding to signal is selected by means of comparing the entering signal to each SS and is determined by minimal istance between them. The distance is measured y means of consecutive comparison of each signal constituent to elements of the sample with he help of the matrix of phonetic distance MPD) stored in the system's memory. MPD comp-ises conventional distances between elements f FSA constituting the signal and SS.

We used the following technique for creating MPD. Each element of FSA corresponds to a certain set of acoustic features. It can be characterized by presence/absence of some feature and the strength of its manifestation (e.g. absence of fundamental frequency is characteristic of unvoiced consonants; by different degrees of its manifestation along with several other features,voiced obstruents, sonorants and vowels are distinguished).

The difference between FSA elements regarding each feature was estimated by assigning certain marks to them. The results of our analysis /9/, /I0/ of reliability of these features in recognition were taking into consideration. The distance between reliable features was given a higher mark, while the distance between less reliable ones was given a low mark, that is the scales of distances were not linear. Thus the scales were made not for elements of alphabet,but for the features by which these elements are characterized. The summarized distance between the constituents of the compared features of FSA elements was put into MPD. In the process in a number of cases frequent substitution of elements of the alphabet in the signal or complete absence of such substitution was taken into consideration.

The technique described above can be presented in the following way: if M and N are elements of FSA, and M is characterized by the set of features $/x_1, \ldots, x_i/$ while N - $/x'_1,\ldots,x'_i/$ then

$$R_{M,N} = r_{(x_1,x'_1)} + \ldots + r_{(x_i,x'_i)} + k_{M,N}$$

where $R_{M,N}$ is the distance included in MPD, r $/x_i, x'_i/$ - the distance in the scale for each feature, $k_{M,N}$ - correction coefficient of substitution frequency of elements in the definite recognition system.

We have distinguished and scaled the following acoustic features:

I. Fundamental frequency
2. Presence of formant structure and degree of its manifestation
3. Intensity
4. Main area of energy concentration
5. FII frequency
6. FI frequency

These acoustic features are highly analogous to syllable contrasts described by L.Bondarko /I/. The main difference here is the absence of durational contrast in our scales. It is impossible to introduce this feature into MPD because the decision is made about each time segment of the signal, and not about segment corresponding to some phonetic unit (whether sound or syllable). The coefficient of comparison between signal duration and sample duration is introduced into algorithm for calculation of distance between the result of the first stage recognition and SS. It seems interesting to compare our data with those obtained on the basis of /I/. By means of the technique described above we have constructed MPDI on the basis of the scales corresponding to syllable contrasts description in /I/. Naturally the absolute distance rarely coincide as singled out features do not match completely, although some general tendency in the sequence of elements of the alphabet which are arranged according to the degree of closeness to each element can be observed. We are planning to compare the efficiency of the matrix in the recognition system.

The results of comparing the signal with SS set allow us to put forward a preliminary hypothesis about some syllable corresponding to the certain signal. As it was mentioned above such decision is represented by either a set of syllables with minimal distances from the signal (in our case 3 minimal distances were taken into consideration) or by symbolic recording of the syllable, reflecting generalized phonetic content (e.g. TA - a syllable consisting of unvoiced stop and non-front vowel). Whether a set of variants or a generalized content would be selected for syllable recording depends on the signal's character (the degree of manifestation of features that allow us to define some concrete sound with greater or lesser precision) or on its distance from the sample. Such attitude seems quite reasonable as not always in the signal there are acoustic cues that would allow us to correlate it with some definite sound, syllable or even word /II/, /I2/.

THE STRUCTURING AND USE OF THE VOCABULARY

As the result of the program for comparing the signals of the first stage recognition with SS each word is represented in form of open-syllables' string. This fact determines the character of phonetic description of the vocabulary. The constructing of the vocabulary can be divided into 2 stages.

On the first stage lexical units in the form close to idealized phonetic transcription are recorded as strings of open syllables. On the second stage pronounciational variants and the most frequent substitutions in recognition being singled out in the preliminary analysis are included into transcribed word recording. If limited vocabulary is used on this stage it is advisable to set apart possible quasihomonyms such as [KAPÁT'-KATÁT'] (in this example the vowels of initial syllable are identical while the consonants at the beginning of the second syllable are phonetically very similar and practically undistinguishable in the process of recognition or are distinguished irregularly).

Each transcribed recording is correlated with corresponding word or words and in the case of reliable syllable recognition we get spelling of the words on the vocabulary output.

The program for syllable joining compares all possible strings of syllable-candidates with those recorded in the vocabulary and corresponding to real words. These equivalents are then recorded into spelling and sent to the output. Variants of input strings of syllables that do not correspond to any vocabulary unit are eliminated. This program imitates the role of the lexical context in phonetic recognition. In some cases it's possible that a whole group of syllable strings would correspond to vocabulary units, thus we'll get 2 or more words at the output. During program approbation such cases were rather few and the number of words at the output didn't exceed 3. This can be explained by the small size of the vocabulary. Theoretically the number of variants for the selected number of syllable candidates /3/ is 3^X, where X stands for number of syllables in the given word. We suppose that in such cases the elimination of extra variants is possible on a higher level of analysis and it corresponds to the role of syntactic, pragmatic and semantic context in speech perception.

A more complex case is presented by the situation when some syllables are identified incorrectly and none of the strings of syllables at the input of the vocabulary module corresponds to the vocabulary units. In this case multiple strategy of word search must be implemented. This strategy must be based on some factors that determine identification of the signal with a lexical unit and its segment composition /in other words the strategy is phonetic-context dependent/. The number of syllables in a word, stress position, rhythmical structure of a word as a whole, basic /most reliable in the process of recognition/ syllables, initial syllables, consonant clusters can be named as such phonetic factors here. The type of the selected factors and their number cause the vocabulary structure, the determining of absolutely reliable factors cause in its turn the strategy of word search in general: consequent search beginning with subvocabularies, composed according to absolutely reliable word characteristics /in respect of the process of recognition/ and onto subvocabularies based on less reliable phonetic word characteristics, were only candidates selected with the help of "reliable" subvocabularies are taken into consideration. As the first stage recognition results don't allow us to consider every selected factor absolutely reliable, one has to turn to parallel word search strategy, although it's a rather complicated procedure.

The vocabulary has the following structure. The vocabulary is recorded in the form of its variants /subvocabularies/, which are organized in accordance with the selected factors. Some subvocabulary is derived into parts comprising similar rhythmic structures, another is oriented at the basic syllables and so on. Strings of syllable candidates which have no corresponding lexical elements in the main vocabulary are entered into all these subvocabularies and there are selected every subvocabulary word candidates identical /in the structural characteristic of a certain subvocabulary/ to the entered string. Word candidates are entered into the analyser which in the output delivers words present in all registers. If no such words can be identified word candidates with the highest marks are selected. For this purpose to each of the subvocabularies a certain rank is assigned according to the reliability of the factor reflected in it.

At present we are conducting an experiment aimed at selecting factors used in word recognition and defining the degree of their reliability. For this purpose the results of syllable recognition with false decision were given to a group of experts, who using the words' phonetic features and unlimited vocabulary put forward some hypothesis about lexical correlation of these results. The group consists of 4 linguists who can theoretically ground their decisions. The data thus obtained are of preliminary character but it should be pointed out that the experts pay attention to the words' rhythmical strmcture and to the segment composition of stressed and initial syllables.

VOCABULARY MODULE

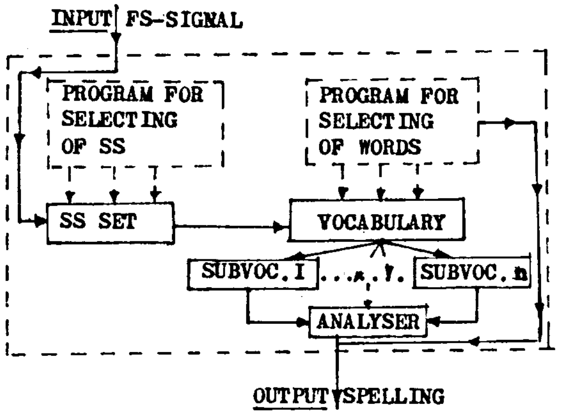

Fig. I

The generalized scheme of the vocabulary module work

With the use of limited vocabulary and pragmatically oriented recognition system the strategy of word prediction can be used /I2/. In this case the reliability of recognition of syllables and the probability of their substitution must be taken into account. The vocabulary must be built in the form of a matrix reflecting the consecutive member of each syllable in a word and the search is conducted through the vocabulary beginning with the most reliable syllable to its possible left or right neighbours. If the supposed neighbouring syllables coincide with the result of the recognition or if they are not contradictory to it /that is, are included into the register of possible substitutions or have the generalized phonetic content/ the search is conducted further on. On the basis of some given text pragmatics semantically and syntactically oriented subvocabularies can be selected to conduct the search while the sequence of entering into each subvocabulary will be determined by the previously recognized words.

This model has no computer-program realization, yet we present it here in accordance with the concept that in speech recognition as well as in human speech perception it's impossible to limit oneself to one particular strategy. The final decision can be made on the basis of identification of the word image as a whole, on the basis of the analysis of the factors /phonetical as well as relating to other levels of analysis/ determining this image, on the basis of prediction of syllables and larger units /words and word combinations/ by limiting the communicative vocabulary according to the pragmatic content of the text.

MODEL'S EXPERIMENTAL APPROBATION

Model's partial approbation /syllable recognition and word selection in the vocabulary with the help of the vocabulary of basic syllables/ was conducted on the vocabulary of 200 words and a set of I00 syllable samples. Syllable candidates are obtained as the result of the realization of the program for comparing of the first stage signals with SS. Strings formed of syllable candidates are recoded into vocabulary units. As the experiment demonstrates, for a small vocabulary it's sufficient to introduce I or 2 subvocabularies. 3 operational variants of the program are possible: simple joining of recognized syllables, word prediction by means of the subvocabulary of basic syllables, refusing to make final decision in case of false recognition or absence of the basic syllable in the string. For the purpose of limiting the number of analysed strings the syllables undoubtedly falsely recognized /initial and final syllables that got into middle position, middle syllables that got into initial or final position/ are eliminated. The result appears in the spelling form with an index showing the ratio between the number of correctly recognized syllables and the total number of syllables in the word. Below some examples of different variants of the decision are given:

TRANSMITTED	RECOGNIZED SYLLABLES			THE RESULTS
	I	2	3	
maja	/maI	ja/	-	maja
	/ma	mna/	-	2/2
	(maI)	(mə)	-	
vzaimno	/vza	$\underline{\text{iI}}$	(rvə)	vzaimno
	/za	(/i)	(nə)	2/3
	(za)	(/p'i)	(va)	zaika
				2/3
v des'at'	/vd'eI	s'ə	$\underline{\text{t'/}}$	v des'at'
	(d'eI)	z'iI	f/	3/3
	/d'eI	(/či)	s'/	des'at'
				3/3

/ma - initial syllable
d'eI - stressed syllable
f/ - final syllable
$\underline{\text{iI}}$ - basic syllable
(maI) - syllable eliminated by the program

The experiment was conducted on computer SM-4 with positive results.

REFERENCES

/I/ Бондарко Л.В. Фонетическое описание языка и фонологическое описание речи. Л., I98I.

/2/ Бондарко Л.В. Слог: правила, интуиция, механизмы. - В кн.: Функциональная просодия текста.М.,I982.

/3/ Бондарко Л.В. Акустические характеристики речи. - В кн.: Слух и речь в норме и патологии /вып. I/. Л., I974.

/4/ Уровни языка в речевой деятельности: к проблеме лингвистического обеспечения автоматического распознавания речи./Под ред. Л.В.Бондарко. - Л., I986.

/5/ Белявский В.М., Светозарова Н.Д. Слоговая фонетика и три фонетики Л.В.Щербы. - В кн.: Теория языка. Методы его исследования и преподавания. Л., I98I.

/6/ Белявский В.М. Автоматическая сегментация слитной речи. - В сб.: IX Всесоюзная акустическая конференция. Тезисы докладов. М.,I977.

/7/ Елкина В.Н., Юдина Л.С. Статистика открытых слогов русской речи. - В сб.: Вычислительные системы, I4. Новосибирск, I964.

/8/ Златоустова Л.В. и др. Алгоритмы преобразования русских орфографических текстов в фонетическую запись. М., I970.

/9/ Москаленко Т.А. Акустический анализ согласных звуков в целях автоматического распознавания русской речи. - В сб.: Автоматическое распознавание слуховых образов: Тезисы докл. и сообщ. АРСО-I4. Каунас, I986.

/I0/ Кузнецова В.Б., Смирнова О.Н. Анализ надежности автоматического распознавания фонетических признаков. - Там же.

/II/ Проблемы и методы экспериментально-фонетического анализа речи. Л., I98I.

/I2/ Кузнецова В.Б. О возможном способе формирования словарных эталонов. - В сб.: Автоматическое распознавание слуховых образов: Тезисы I2-го Всесоюзного семинара АРСО-I2. Киев, I982.