A Classification Method of Spoken Words in Continuous Speech for Many Speakers

Toshiyuki SAKAI* and Sei-ichi NAKAGAWA*

Abstract

Speech wave is converted into a time series of short time spectra by 20-channel filter bank and is segmented into four groups: silence, unvoiced-non-fricative, unvoiced-nonplosive, and voiced group. The unvoiced groups are classified into a unit of phoneme by heuristic algorithms and voiced group by Bayes rule.

To normalize the variation of reference patterns among speakers, vowel patterns are learned by the non-supervised learning method.

The optimum matching between a just recognized phoneme string and a phoneme string of a given word in the word dictionary is performed by utilizing the phoneme similarity matrix and Dynamic Programming.

According to the results tested upon 1,500 samples of isolated digits, spoken by 20 male speakers, about 97% were correctly recognized and, in case of the system adapting for each speaker, 98% correctly recognized.

1. Introduction

In automatic word recognition, if the whole input pattern is regarded as a point in the pattern space, the recognizer can avoid the problem of coarticulation. It can also use the linguistic information obtained through the word dictionary, i.e., the redundancy of natural language. Therefore, for limited speakers, it is fairly easy to recognize spoken words in a limited vocabulary.(1)-(3) As the task vocabulary size becomes larger, it becomes necessary that the system be able to identify more words. If each word has to be matched one by one against the acoustic phonetic data, this process will consume a large amount of computational time and memory storage. Therefore, as a unit of recognition, we have to consider smaller units such as phonemes,

This paper first appeared in Japanese in Joho－Shori (Journal of the Information Processing Society of Japan), Vol. 17, No. 7 (1976), pp. 650～658.

* Faculty of Engineering, Kyoto University

syllables, and VCV syllables. In order to normalize differences among speakers, we need some normalizing or learning procedure.

In this paper, we describe a method of word classification on the basis of units of phonemes and a matching method between two phoneme strings.

2. Outline of System [4]

Fig.1 shows the block diagram of the system. The system first analyzes input speech by the filterbank. Primary segmentation is performed on such analyzed speech, now represented by a sequence of short time spectra (We will call each spectrum a 'frame'). Each frame is classified into one of four groups: silence, voiceless-nonfricative, voiceless-nonplosive or voiced; based on the energy and deviation around the low or high frequency of (20-dimensional) spectrum. Each classified segment is recognized as one of phonemic categories. If a part of a sequence of recognized segments is composed of the same successive phonemic categories, these are combined. On the other hand, if a part is irregular, it is smoothed by using rewriting or phonological rules. The output of this algorithm is a sequence of continuous and non-overlapping segments.

A segment which has been regarded as included in a voiceless group is further classified into one of phonemes: /s/, /c/, /h/, /p,t,k/. This more detailed classification is based on the segment duration, the presence of silence preceding the segment, spectral change, etc.

For a segment classified into a voiced group, it is next determined whether each frame included in this segment is stationary, quasi-stationary or transient. Such determination is based on the degree of spectral change between adjacent frames (We call this the secondary segmentation).

The decided stationary and quasi-stationary parts are regarded as a presenting portion for vowels. The most stationary frames are used for partially non-supervised learning of the spectrum patterns of vowels. A portion of voiced consonants is detected as one of the following: 1) rejected by the vowel recognition process; 2) long transient; 3) having weak energy with concave speech level (or dip). The spectral patterns of voiced consonants are gradually trained (or estimated) by using those learned vowels.

Vowels and voiced consonants are recognized by Bayes' discriminant functions, which are obtained from the renewed standard patterns of each phoneme. Semi-vowels are recognized by applying rewriting rules to a recognized noisy phoneme string.

To reduce the influence of missegmentation on the system performance, the segmentation process is designed so that a voiced part may be divided into a segment finer than a phoneme. This division will be recovered in word identification process, for example, a vowel can be allowed to match with up to three segments and a consonant with up to two segments.

To the segment thus processed are given the first candidate phoneme, the second one, the degree of confidence (reliability) of the first candidate, and the segment duration. Also the string of segments would be corrected by phonological rules.

This phoneme string is translated into a word. It is based on matching technique of a recognized phoneme sequence against a phoneme sequence given by an entry in a word dictionary. This matching uses a phoneme similarity matrix and Dynamic Programming.

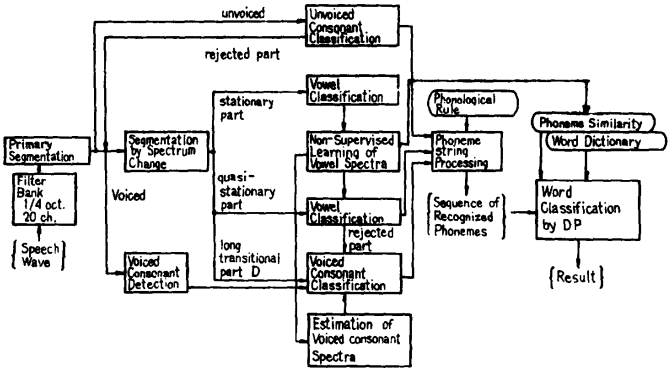
unvoiced
Unvoiced Consonant Classification
rejected part
Vowel Classification
stationary part
Phonological Rule
Phoneme Similarity
Word Dictionary
Primary Segmentation
Segmentation by Spectrum Change
Non-Supervised Learning of Vowel Spectra
Phoneme string Processing
Filter Bank 1/4 oct. 20 ch.
Voiced
quasi-stationary part
Vowel Classification
Sequence of Recognized Phonemes
Word Classification by DP
rejected part
Speech Wave
Voiced Consonant Detection
long transitional part D
Voiced Consonant Classification
Result
Estimation of Voiced consonant Spectra

Fig. 1 Block diagram of spoken words recognition.

3. Partially Non-Supervised Learning of Vowel Spectra

In non-supervised learning, the system must be able to treat training data very carefully in order to avoid mislearning. From our preliminary experiments, we found that vowels can be recognized very accurately in the stationary parts extracted from the secondary segmentation. Therefore, the spectrum of a stationary part where vowels are recognized with high reliability is a suitable sample for (partially) non-supervised learning. Let us consider the way to learn the average spectrum m_i^j of the vowel i for the speaker j. Given N samples, x_1, x_2, ... x_N , uttered by the speaker j and recognized as the vowel j , we can use the method for learning the mean vector[5].

The value of m^{j}_{i} at the (N-1)th stage of learning is denoted by $m^{j}_{i N-1}$. This is renewed by the N-th sample X_N as follows:

$$m^{j}_{i,N} = \frac{1}{\alpha+N}\,(\alpha\cdot m_{i} + N\cdot X_{1,N}) = \frac{1}{\alpha+N}\{(\alpha+N-1)m^{j}_{i N-1} + X_{N}\}$$

$$X_{1,N} = \frac{1}{N}\sum_{k=1}^{N} X_{k}$$

$$m^{j}_{i,0} = m_{i}$$

where m_i is a mean vector for all speakers, and α is a constant value representing the degree of ambiguity of m^{j}_{i}. We substituted 10 for α in this experiment.

The system selects training samples automatically, and recognize vowels and syllabic nasal.

4. Word Classification from Phoneme String

4-1 Similarity between two Phonemes

Phoneme recognition is performed using statistics of the spectrum (means and covariance matrices in 20-dimensional vectors) for phonemes. Thus, if the Phoneme Recognizer makes mistakes in phoneme recognition, we should consider that such errors are caused by the fact that statistics calculated from an uttered phoneme are very similar to those of a misrecognized phoneme. The errors are generally divided into three kinds: a) substitution, b) insertion, c) omission.

Word matching is fundamentally defined as the process which makes a one-to-one correspondence between each phoneme of a recognized phoneme string and each phoneme of an entry (a word) in the word dictionary. To evaluate a degree of matching between two phonemes, we introduce a concept of similarity between two phonemes.

The Bhattacharyya distance is closely related to the confusion matrix constructed from the results of phoneme recognition based on Bayes' rule. We calculate the similarity $S(i,j)$ for a pair of phonemes (i and j) by the linear transformation of the distance. If i belongs to the set (a,i,u,e,o,N) and j to the set (m,n,$\tilde{g}$,b,d,g,r,z), $S(i,j)$ is decreased by 5. On the other hand, if i belongs to the second set or equals to /y/ or /w/ and j is in the first set, S(i,j) is increased by 5. Such modification of S(i,j) is to compensate for the unsymmetric performance of the phoneme recognition. If either i or j is unvoiced, S(i,j) is derived from the confusion matrix. The resulting similarity matrix is shown in Table 1. The column "in" denotes phonemes in the lexicon, and the column "out", recognized phonemes. In this table, the pseudo phoneme /*/ is treated as an unvoiced plosive, except that it is not associated with the silence group.

Table 1 Phoneme similarity matrix

In \ Out	*a*	*i*	*u*	*e*	*o*	*m*	*n*	*η*	*p*	*t*	*k*	·	—	/	···
a	100	36	51	69	75	51	55	58	5	5	5	5	5	5	···
i	36	100	83	73	56	72	74	85	50	50	50	5	5	5	···
u	51	83	100	74	80	81	83	89	5	5	5	5	5	5	···
e	69	73	74	100	69	59	69	74	5	5	5	5	5	5	···
o	75	56	80	69	100	64	65	75	5	5	5	5	5	5	···
m	61	82	91	69	74	100	92	86	5	5	5	5	5	5	···
n	65	84	93	79	75	92	100	88	5	5	5	5	5	5	···
η	68	95	99	84	85	86	88	100	65	85	85	30	5	5	···
*	5	5	5	5	5	5	5	5	100	100	100	5	5	5	···
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮ ⋮ ⋮

4-2 Word Dictionary

All words which are used in recognition are preregistrated in the word dictionary. In order to reduce matching time and computer storage, the dictionary describes each word as a string of phonemic symbols in only one way. Some phonemes in a word are often influenced by phoneme environments, while other phonemes are sometimes devocalized. Because of this influence, such phonemes are often omitted or misrecognized as other phonemes. By introducing the sub-phoneme 'k' in addition to the main-phoneme 'I', we denote these situations in the dictionary by I/k(c). This notation means that the phoneme 'I' can be replaced by the phoneme 'k', where c (0≤c≤1.0) means the weight of the sub-phoneme 'k'. Both phonemes are equally treated if it is 1.0, and the sub-phoneme is neglected if it is 0. By this description, we can represent an optional phoneme, i.e., one that is omissible or addible.

Table 2 shows a part of the word dictionary. The special symbols (+ and -) indicate changes of restrictions for DP matching. The maximum and minimum durations indicate the range of duration time required for uttering a word.

Theserepresentations for given words are automatically constituted by the constructing rules of the word dictionary.

An output of the Phoneme Recognizer is a sequence of segments, each consisting of the first and second candidates of phonemes, the degree of confidence of the frtst candidate, and the segment duration. The first three constituents of the j-th segment in a sequence will be denoted by J, l and $p(0 \leq p \leq 1.0)$, respectively. An element by which a word in the word dictionary is described is either a main-phoneme or a sub-phoneme plus a weighting factor. The constituents of the i-th element of a lexical entry will be denoted by I, k and c, respectively. In order to match a portion of a segment string against a word, we must first define the similarity between a segment and an element in the entry. This similarity is defined by the following equation:

$$S(I,k,c;J,l,p) = \max \begin{cases} S(I,J) \\ c \times S(k,J) \\ p \times S(I,J) + (1-p) \times S(I,l) \\ p \times c \times S(k,J) + (1-p) \times c \times S(k,l) \end{cases}$$

We simply denote $S(I,k,c;J,l,p)$ as $S_0(i,j)$.

Table 2 Example of entries in the Word Dictionary

Word	Symbol	Phoneme Representation	Maximum Duration	Minimum Duration
ichi	1	i ·/c(1.0) c i/c (1.0)	350ms	100ms
ni	2	n i	300	100
san	3	s a N	550	200
yon	4	y/g (0.9) o N	450	150
go	5	g o	300	100
roku	6	r/p(0.85) o ·/k(0.95) k u/a(1.0)	450	100
nana	7	n a/N(0.85) n/a(0.85) a/N(0.85)	550	200
hachi	8	h a/N(0.85) ·/c(1.0) c i/c(1.0)	500	150
kyu	9	i/c(0.95) k/c(0.95) y/u(0.95) u	500	200
rei	0	r/p(0.85) e i/e(0.95)	400	100

4-3 Word Matching by DP

We make the following restrictions with respect to the matching between an input string and an element string in the Word Dictionary. These could be regarded as reasonable restrictions,· judging from the performance of the Phoneme Recognizer.

(1) Except for elements marked with the symbol (-), a vowel and the syllabic nasal in the word dictionary can be associated with three or less segments in a recognized phoneme string.

(2) A consonantcan be associated with two or less segments.

(3) Three or more successive elements cannot be associated with one segment except when there is an element marked with the symbol (+).

(4) When the total duration time of three successive segments is beyond 250ms, a vowel element (except for an elongated vowel) does not have to be associated with the three segments.

(5) When the duration time of one segment is not beyond 100ms, an elongated vowel element cannot be associated with only this segment.

(6) If a word matching is performed outside the range of the duration time specified by the lexical entry, the matching score is decreased.

Now, we consider how to calculate the likelihood for a given word. Let $L(i,j)$ be the highest cumulative similarity (or score), when considering the i-th element of the lexical entry for this word and the j-th segment of a recognized phoneme string. In other words, $L(i,j)$ is determined by evaluating all possible paths from the point (1,1) to the point (i,j) on the lattice plane. When the i-th element is a vowel or the syllabic nasal, $L(i,j)$ is calculated successively by the following equation:

$$L(i,j) = \max \{L_1(i,j),\ L_2(i,j),\ L_3(i,j),\ L_4(i,j),\ L_5(i,j),\ L_6(i,j)\},$$

where

$$L_1(i,j) = L^*(i-1,j) + S_0(i,j)$$

$$L_2(i,j) = L\ (i-1,j-1) + S_0(i,j)$$

$$L_3(i,j) = L^*(i-1,j-1) + \{S_0(i,j-1) + S_0(i,j)\}/2$$

$$L_4(i,j) = L\ (i-1,j-2) + \{S_0(i,j-1) + S_0(i,j)\}/2$$

$$L_5(i,j) = L^*(i-1,j-2) + \{S_0(i,j-2) + S_0(i,j-1) + S_0(i,j)\}/3$$

$$L_6(i,j) = L(i-1,j-3) + \{S_0(i,j-2) + S_0(i,j-1) + S_0(i,j)\}/3$$

If the i-th element is marked with the special symbol (+), $L^*(i,j) = L(i,j)$; otherwise $L^*(i,j) = \max\{L_2(i,j), L_3(i,j), L_4(i,j), L_5(i,j), L_6(i,j)\}$. This selection of $L^*(i,j)$ corresponds to the restriction (3) in matching. The boundary (or initial) conditions are the following:

$$L(1,j) = 0 \qquad : j \geq 4$$

$$L(1,1) = S_0(1,1)$$

$$L(1,2) = \{S_0(1,1) + S_0(1,2)\}/2$$

$$L(1,3) = \{S_0(1,1) + S_0(1,2) + S_0(1,3)\}/3$$

When the i-th element is a consonant, silence or an element with the special symbol (-), $L(i,j)$ is calculated by the following equation:

$L(i,j) = \max\{L_1(i,j), L_2(i,j), L_3(i,j), L_4(i,j)\}$

$L(1,3) = 0$

If the numbers of elements in a lexical entry and input string are i_0 and j_0, respectively, the likelihood of this word is calculated as $L(i_0, j_0)/i_0$. Input utternace is recognized as a word which has the highest likelihood out of all words. Fig.2 shows the graphic representation of word matching.

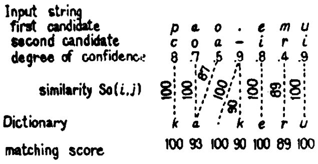

Fig.2 Graphic representation of word matching

5. Experimental Results of Spoken Digits Recognition

The experiments were composed of the following three types:

Experiment 1. The common reference patterns (similarity matrix, reference spectra) for the ten male speakers, whose different speech materials were used in the system design, were employed.

Experiment 2. The personally tuned reference patterns were used for each of the same ten speakers.

Experiment 3. The digits spoken by 10 new speakers were recognized by using the common reference patterns used in *Experiment 1.*

For each experiment, each of the ten speakers uttered every digit five times. The recognition rates of (1), (2), and (3) were 96.4%, 97.4%, and 97.6%, respectively.

References

(1) V.M.Velichko and N.G.Zagoruyko: Automatic Recognition of 200 Words, Int.J. Man-Machine Studies, Vol.2,p.223 (1970)

(2) H.Sakoe and S.Chiba: A dynamic Programming Approach to Continuous Speech Recognition, Seventh International Conference on Acoustics, Budapest,p.65 (1971)

(3) M.Kohda, S,Hashimoto and S.Saito: Spoken Digit Mechanical Recognition System, Jour. of IECEJ, Vol.55-D, p.186,(1972) (in Japanese)

(4) S.Nakagawa: A Machine Understanding System for Spoken Japanese Sentences, Doctral Thesis, Kyoto University, Oct. 1976

(5) N.J.Nilsson: Learning Machines, MacGraw-Hill Co. New York (1965)