SPEECH RECOGNITION BASED ON THE DISTANCE CALCULATION BETWEEN INTERMEDIATE PHONETIC CODE SEQUENCES IN SYMBOLIC DOMAIN

Kazuyo TANAKA and Hiroaki KOJIMA
Electrotechnical Laboratory, 1-1-4 Umezono, Tsukuba, 305, JAPAN
ktanaka@etl.go.jp, and hkojima@etl.go.jp

ABSTRACT

This paper proposes a speech recognition method alternative to the conventional sample-based statistical methods which are characterized by the necessity of large amounts of training speech data. To resolve this type of heavy processing, the proposed method employs an intermediate phonetic code system and the calculation of distance between phonetic code sequences in symbolic domain. It realizes high efficiency when compared with direct processing of acoustic correlates, although some deterioration will be expected in recognition scores. We first describe the distance calculation method and present specific procedures for obtaining the intermediate code sequence from input utterances and for spotting words using the calculation of distance in the symbolic domain. Preliminary experiments were examined on isolated word recognition and phrase spotting in continuous speech. Word recognition results indicate that the recognition scores obtained by the proposed method are comparable compared with ordinary phone-HMM-based speech recognition.

1. INTRODUCTION

In the past decade, speech technology has achieved remarkable progress in large vocabulary speech recognition. Statistical techniques such as Hidden Markov Models (HMMs) and large-scale speech databases are two major contributions for this progress. The implication is, however, that in such a framework, speech recognition performance inevitably depends upon the speech samples and language data concerning their respective acoustic and language environments. In other words, implementing a recognition system needs speech and language corpora that are collected in environments similar to those used in the system. Thus, we were always faced with heavy data collection projects when implementing those systems. To overcome this situation, we propose a framework in which speech signals are once represented by an *intermediate phonetic code (IPC)* system and this IPC sequence is then treated as a primary input for the upper processing, without using likelihood values for the IPC sequence. The upper processing employs distance calculation in the symbolic domain, which utilizes distance matrices predetermined for the IPCs and individual phonetic code systems.

The distance calculation in the symbolic domain is much more efficient compared with that in the acoustic domain. This framework therefore also makes it possible to decrease the amount of processing for conventional statistical recognition procedures. We have already applied this framework to the estimation of the degree of recognition difficulty for a given word vocabulary and predicting word candidates which follow the first candidate in large vocabulary recognition systems [Tanaka et al, 1997b].

In this paper we first describe the basic framework of our method. Next we present a specific procedure to convert speech utterance to the IPC sequence that consists of a subphonetic code set, and also present a method for calculating the distance between phonetic symbol sequences, which is improved from our previous papers [Tanaka et al, 1997b]. To confirm feasibility, two recognition experiments are examined. One is an experiment on speaker-independent word recognition, where the results of the proposed method are compared with those of ordinary phone-HMM-based speech recognition. The other is a preliminary experiment on phrase spotting in continuous speech utterances.

2. BASIC FRAMEWORK

The basic framework for the proposed processing scheme is shown in ***Fig.1***. As indicated in this figure, the IPC sequence converted from an input speech is transformed into a symbol sequence of corresponding phonemic system by calculating the distance between individual symbol sequences. Therefore, if the distance matrices between the IPC symbols and individual

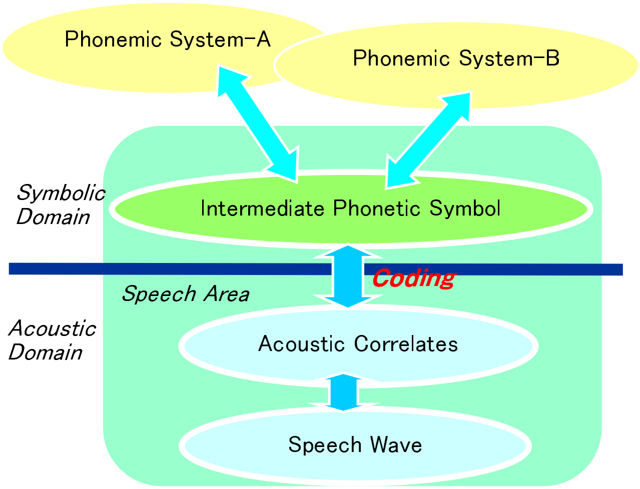

Fig. 1 *Framework for the separation of symbolic domain processing from acoustic domain.*

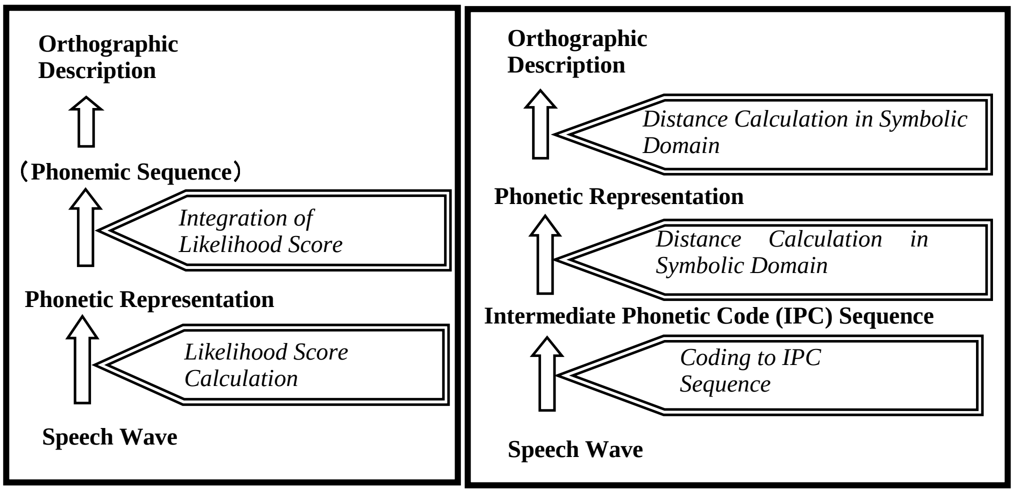
Orthographic Description
（Phonemic Sequence）
Integration of Likelihood Score
Phonetic Representation
Likelihood Score Calculation
Speech Wave
Orthographic Description
Distance Calculation in Symbolic Domain
Phonetic Representation
Distance Calculation in Symbolic Domain
Intermediate Phonetic Code (IPC) Sequence
Coding to IPC Sequence
Speech Wave

Fig.2 *Illustration for the distinction between the ordinary phone-based recognition procedure (the left box) and the proposed one (the right box).*

phonetic symbols are determined in advance by using relatively small speech data sets, it will be possible to flexibly change phonemic systems to be hypothesized. ***Fig.2*** illustrates the distinction between this method and the conventional speech recognition method. As indicated, the proposed method separates the acoustic domain data from the symbolic domain processing except for the IPC sequence, whereas the conventional method uses the integration of likelihood scores for the input phonetic symbol sequence.

Currently, we use subphonetic units, called demiphonemes as the IPC set. The demiphoneme were originally proposed for precise representation of Japanese speech [Tanaka et al, 1986], so that it depends upon language. (However, we intend to adopt a more common code system in the future.) The framework will be effective in its application to phrase spotting tasks, where the system can deal simultaneously with multi-categorical hypotheses for the input speech.

3. CONVERSION FROM SPEECH TO THE IPC SEQUENCE

3.1 Subphonetic Units as an IPC set

As mentioned above, the IPC (intermediate phonetic code) set is now a set of demiphonemes which belongs to a category of subphonetic segmental units. It is defined for a given phoneme sequence from an acoustic-phonetic sense[Tanaka et al, 1986]. For Example:

Phoneme sequence example: */yokohama/*(city name)
Demiphoneme sequence:
<y-yy-yo-oo-ok-qk-kk-ko-oo-oh-hh-ha-aa-am-mm-ma-aa-a>

Here note that demiphonemes consist of steady and transitional segments and transitional segments, such as *yo, ok*, etc., depend on their phonemic contexts. According to our previous work [Tanaka et al, 1990], about 320 segment labels cover the variations of the common Japanese language transcriptions, and roughly 600 segment labels cover acoustic-phonetic variations.

3.2 Conversion Procedure

Each demiphoneme is represented by a left-to-right HMM with three states and three loops, and every state is represented two mixture continuous density distributions. The input speech is converted to a demiphoneme sequence in a kind of phonetic unit recognition, using only Japanese syllabic and acoustic-phonetic constraints which were previously established as a rule from labeled speech database [Tanaka et al, 1990]. The constraints are compiled to limit the labels that can follow the current label. For Example, demiphoneme labels (*aa, ap, at, …*) are possible to follow label *ka*. This is expressed by:

ka-followers: aa, ap, at, ak, am, an, ab, ad, ag,….

An example of the demiphoneme sequence converted from an input speech sample is shown in ***Fig.3***.

Input: Japanese speech utterance in phonemic description :
/iqshuukaNbakari nyuuyookuoshuzaishita/

Output: Demiphoneme sequence obtained by the phonetic unit recognition:
#i, ii, ish, shq, ssh, shu, uk, qk, kk, ka, aN, Nb, ba, aa, ak, qk, kk, ka, ar, ri, ii, i#, sil2, sil1, sil1, #n, nn, ny, yu, uy, yy, yo, oo, ok, qk, kk, ku, uo, oo, osh, shi, iz, zzz, za, aa, ai, ish, ssh, shit, qt, tt, ta, aa, a#, silE.

Fig. 3 *An example of the phonetic unit recognition results. This sample indicates an utterance of a sentence, where silence intervals were assumed between every phrase periods and automatically identified. (This is indicated as a part of the sequence, "i#, sil2, sil1, sil1, #n".)*

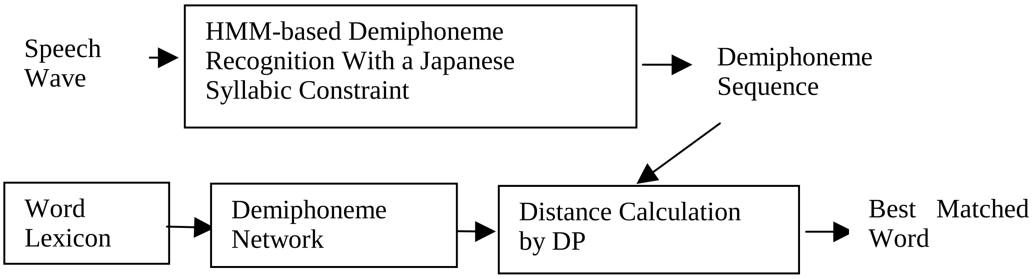
Speech Wave
HMM-based Demiphoneme Recognition With a Japanese Syllabic Constraint
Demiphoneme Sequence
Word Lexicon
Demiphoneme Network
Distance Calculation by DP
Best Matched Word

Fig. 4 *Block diagram illustrating word recognition system based on the symbolic domain matching.*

4. DISTANCE CALCULATION IN THE SYMBOLIC DOMAIN

Now we consider two words that are expressed by demiphoneme sequences, and let us denote the distance between word *i* and word *j* by *d(i, j)*, then *d(i, j)* is calculated by dynamic programming (DP) under the following allowed path conditions:

$$G(m,n)=\min\begin{Bmatrix} G(m-1,n-2)+3D^2(m,n) \\ G(m-1,n-1)+2D^2(m,n) \\ G(m-2,n-1)+3D^2(m,n) \end{Bmatrix} \quad (1)$$

$m=1,2,\dots,M,\ \ n=1,2,\dots,N$

where *G(m,n)* and *D(m,n)* indicate an accumulated distance value, and distance between the *m*th demiphoneme segment of word *i* and the *n*th segment of word *j*, respectively. Then *G(M,N)* is normalized as

$$\{d(i,j)\}^2=G(M,N)/(M+N-1) \qquad (2)$$

The distance between demiphonmes is calculated using the values of individual HMMs, as follows. Let us denote the centroid vectors of each state by

$c_{ij}(k)$: centroid vector of the *j*th distribution in state *i* of category *k*.

Then we define the distance between demiphonemes *k* and *l* as

$$D(k,l)=\sum_{i=1}^{3}\min_{j,j'}[\{c_{ij}(k)-c_{ij'}(l)\}^2]$$

This formulation means that the distance is estimated by using the centroid vector of the nearest distribution in each state of the demiphoneme HMMs. Of course there are several alternatives to estimate the distance. They basically depend upon criteria for estimating phone models and at the same time upon the applications. For example, very fast search for spoken document retrieval, a simple matching is considered [Foote et al, 1997].

5. EXPERIMETAL RESULTS ON ISOLATED WORD RECOGNITION

5.1 Word Recognition via Demiphoneme Sequences

Speaker-independent word recognition experiments were carried out in order to confirm the feasibility of the proposed method. The block diagram of the proposed recognition procedure is shown in ***Fig.4***, which indicates that an input speech utterance is first converted to a demiphoneme sequence and then its word name is identified by matching the demiphoneme sequence with demiphoneme networks generated from word lexicon by rules. This matching is done by the DP method described in section 4.

5.2 Recognition Results

To compare the performance of the proposed method with that of the ordinary recognition method, we have used the same speech sample set for both recognition methods. The test sample set consisted of 492 words uttered by four different male speakers. The ordinary recognition system was implemented by a phoneme-HMM-based recognition, where phoneme HMMs were represented by three states, three loops and two mixture continuous density distributions [Tanaka et al, 1997a]. Other conditions, such as acoustic features and HMM training, were set up into almost the same conditions.

The recognition results by the proposed method was an 87% correct rate when using the most basic demiphoneme sequence generation rule and 91% when using a more general rule. On the other hand, the base line recognition score by the conventional phoneme-HMM-based method was 89% for the same sample set. Therefore the recognition rate was almost comparable order in both methods.

6. PRELIMINARY EXPERIMENT ON PHRASE SPOTTING

6.1 Procedure

The proposed framework is proper for phrase spotting in speech stream. Here we have examined extracting hypothesized phrases from continuous speech utterances based on this framework, that

(a) Phonemic sequence of the input speech sample: /arayuru geNjitsuo subete …./
(b) Word to be spotted: /geNjitsu/
(c) Subphonetic code chunk of /genjitsu/: gg-ge-ee-eN-NN-Nj-jj-jjj-(ji-ii-its | jit)-qts-tts-(tsu-uu | tsux)
(d) Subphonetic code sequence result converted from the input (actual example):
#qk-kk-ka-ar-ra-ay-yy-yu-ur-re-ek-qk-kk-ke-eN-NN-Nj-jjj-ji-its-qts-tts-tsu-uo-o#-sil2-sil1-#s-ss-…
(e) The network of (c) is matched with the sequence of (d) using the continuous DP.

Fig.5 *Procedure for the phrase spotting.*

is, by the distance calculation between a demiphoneme sequence converted from the input speech and demiphoneme networks representing hypothesized phrases (i.e., the references). The distance is calculated by a frame synchronous dynamic programming, called continuous DP [Oka, 1978]. This is basically formulated similar to eq.(1), except that the accumulated distance is normalized by the length of the reference symbol sequence. Therefore, when assuming that *m* indicates the reference direction, the following formulation is used:

$$G(m,n) = \min\left\{\begin{matrix} G(m-1,n-2)+D^2(m,n) \\ G(m-1,n-1)+D^2(m,n) \\ G(m-2,n-1)+2D^2(m,n) \end{matrix}\right\} \qquad (4)$$

We have used a simple algorithm to extract key phrases to be spotted. The place of satisfying all the following conditions is extracted:

a) The distance value takes the minimum in duration that satisfies the distance value less than *delta1*.
b) The distance takes the value less than *delta2 (< delta1).*
c) The extracted places are distant more than five symbols each other.

Fig. 5 illustrates the procedure for the phrase spotting from a continuous speech utterance.

6.2 Experimental Results

A set of 50 Japanese sentences, named ATR A-set, was adopted for the test data. The key phrases were conveniently determined as selecting all adjacent word pairs contained in the test sentence set. The number of such different phrases was 261 and lengths of the phrase samples range about 10 to 25 phonemes.

The ASJ Continuous Speech Corpus [Kobayashi et al, 1992] was used for the test utterance set. The total number of the test utterances was 500 by ten different male speakers. ***Table 1*** shows the recognition results depending upon threshold *delta2*, where *delta1* was kept in constant (*delta1* =1.2). The denominator of each percent value is the number of (hit + missing), that is, 261x10 samples.

Table 1*: Experimental Results for the Phrase Spotting in a Sentence Set.*

Threshold *delta2*	0.6	0.8	1.0
% correct (hit)	84.6	89.6	91.8
% insertion error	0.6	2.1	13.5
% missing error	15.3	10.3	8.1

7. CONCLUDING REMARKS

We have proposed a new framework for speech recognition, which will be an alternative for the conventional statistical speech recognition methods. The method will be particularly effective for phrase spotting and multilingual speech recognition because of its being less dependent on the acoustic environment [Tanaka 1998]. Further investigations are of course needed for deriving a proper IPC set and confirming the feasibility of several recognition applications.

In the last, we wish to thank Nahoko Fujimura for her technical assistance and also the members of the Speech Processing group in ETL for their usual discussion.

REFERENCES

[Foote et al, 1997] J.T. Foote, S.J. Young, G.J.F. Jones, "Unconstrained keyword spotting using phone lattice with application to spoken document retrieval," Computer Speech and Language 11, pp.207-224 (1997).

[Kobayashi et al, 1992] T. Kobayashi, S. Itahashi, S. Hayamizu, T. Takezawa, "ASJ continuous speech corpus for research (in Japanese)," J. Acoust. Soc. Jpn, 48, 2, pp.888-893 (1992).

[Oka, 1978] R. Oka, "Continuous words recognition by use of continuous dynamic programming for pattern matching (in Japanese)," ASJ Tech. Report S78-20 (1978).

[Tanaka et al, 1986] K. Tanaka, S. Hayamizu, K. Ohta, “A demiphoneme network representation of speech and automatic labeling techniques for speech data base construction”, Proc. of ICASSP-86, pp.309-312 (Apr. 1986).

[Tanaka et al, 1990] K. Tanaka, S. Hayamizu, K.Ohta, "Sorting and clustering of acoustic-phonetic variations based on a fine-labeled speech database with applications for automatic word recognition (in Japanese)", IEICE Trans. J73-D-II, pp.1619-1629 (1990).

[Tanaka et al, 1997a] K. Tanaka, H. Kojima, “A method of extracting time-varying acoustic features for speech recognition”, Proc. of ICASSP97, pp.1391-1394 (Apr. 1997).

[Tanaka et al, 1997b] K.Tanaka, H. Kojima, “A between-word distance calculationin a symbol domain and its applications to speech recognition”, Proc. of ICONIP-97, pp.1107-1111 (Nov. 1997).

[Tanaka, 1998] Tanaka, K., ”Next major application systems and key techniques in speech recogniton technology”, Proc. of ICASSP'98, pp.1057-1060 (May 1998).