SPEECH DATA RETRIEVAL SYSTEM CONSTRUCTED ON A UNIVERSAL PHONETIC CODE DOMAIN

Kazuyo TANAKA, Yoshiaki ITOH**, Hiroaki KOJIMA*, and Nahoko FUJIMURA**
*National Institute of Advanced Industrial Science and Technology(AIST)
**Iwate Prefectural University
{kaz.tanaka; h.kojima}@aist.go.jp, y-itoh@iwate-pu.ac.jp

ABSTRACT

In this paper we propose a novel speech processing framework, where all of the speech data are encoded into universal phonetic code (UPC) sequences and speech processing systems, such as speech recognition, retrieval, digesting, etc., are constructed on this UPC domain. As the first step, we introduce an IPA-based sub-phonetic segment (SPS) set to deal with multilingual speech and develop a procedure to estimate acoustic models of the SPS from IPA-like phone models. The key point of the framework is to employ environment adaptation into the SPS encoding stage. This makes it possible to normalize acoustic variations and extract the language factor contained in speech signals as encoded SPS sequences. We confirm these characteristics by constructing speech retrieval system on the SPS domain. The system can retrieve key phrases, given by speech, from different environment speech data in vocabulary-free condition. We show several preliminary experimental results on this system, using Japanese and English sentence speech sets.

1. INTRODUCTION

We are developing a framework and architecture for constructing universal-phonetic-code-based speech processing systems. In this framework, we encode all of the speech data, contained in or entered into information systems, into *universal phonetic code (UPC)* data and then construct the speech processing systems, such as recognition, synthesis, retrieval, indexing, digesting, etc, on this code domain, as illustrated in **Fig.1**. We present, in the paper, a procedure of introducing a candidate set of the UPCs and show several advantages of this framework by constructing a speech data retrieval system on the UPC set.

Introducing the UPC set is motivated by the necessity of handling various dialects including non-native speaker's pronunciations, where individual language-dependent phonological systems might be ambiguous. The IPA (International Phonetic Alphabet) or XSAMPA [1], which is the ASCII code of the IPA, will be a candidate set for this purpose. We, however, propose a more fine segment, called *sub-phonetic segment (SPS)* for this purpose, which is derived from XSAMPA on the basis of the acoustic-articulatory considerations.

A speech data retrieval system is constructed on the SPS code domain, as an application system. Speech retrieval systems usually function either as retrieving key words, given by spoken words, from text-based DB, or retrieving key words, given by text, from speech-based DB [2]. In addition, key words are usually limited in vocabulary size because of speech recognition performance. If both key words and objective DB are real speech, it is difficult to function well because the acoustic characteristics of those two are, in usual cases, considerably different. The proposed retrieval system architecture resolves these difficulties by introducing phrase spotting in SPS domain computation and environment-dependent SPS coding. The system basically uses phonetic coding, does not use word recognition, so that it is a vocabulary free system. The system architecture is in part based on a speech recognition framework using phonetic symbol domain computation, previously proposed [3,4].

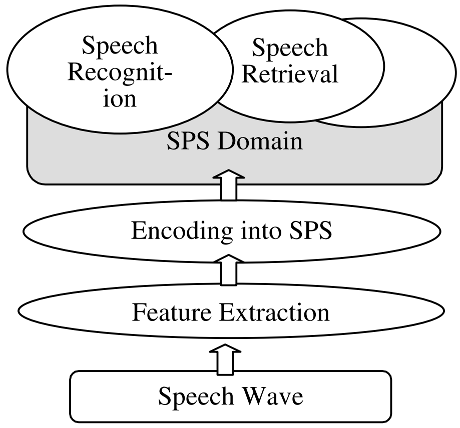

Fig.1 Proposed framework for speech processing.

2. SUBPHONETIC SEGMENT (SPS)

2.1 From XSAMPA to SPS Labels

The SPS labels are obtained from the XSAMPA segment sequences using conversion rules. The rules are created by considering the acoustic-articulatory characteristics. Before deriving the SPS set, we slightly modify XSAMPA symbols to be convenient to deal with in speech processing systems. For example, the XSAMPA (i.e., IPA) contains partly extra-detailed categorization to be modeled in engineering sense. Therefore, we basically adopt only primary IPA symbols and represent minor phonetic variations by statistical distributions in acoustic domain.

The SPS sequences converted from XSAMPA sequences consist of stationary and non-stationary (transitional) segments in the speech stream, for example,

[eIkl]: #e, ee, eI, II, Ik, kcl, kk, kl, ll, l#
[koobe]: #kcl, kk, ko, ooo, ob, bcl, bb, be, ee, e#

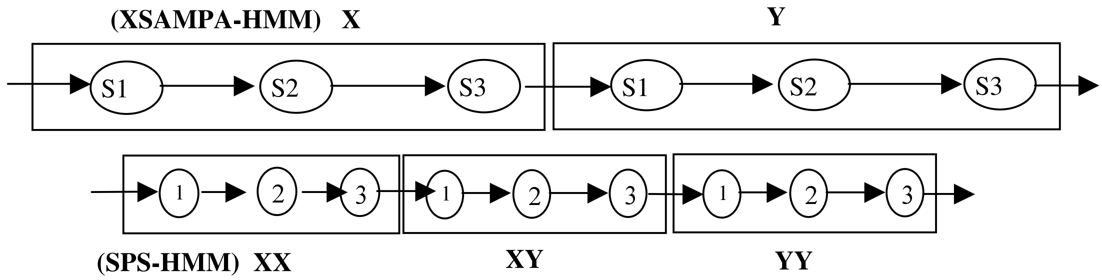

Fig.2 Schematic diagram of relationship between XSAMPA-HMMs (upper) and SPS-HMMs (lower).

The SPS set is extended from those originally proposed for recognizing Japanese utterances[5]. The advantages of SPS-type segments was confirmed by the experiments shown in the Japanese speech recognition[6]. Similar segment category was also proposed and used in an Italian speech recognition system[7]. The SPS set can deal with multilingual speech in the same level as the XSAMPA (or IPA) can. However, the SPS set contains a large number of segmental labels, so that a definite procedure is required to estimate segment models when we apply it to individual speech data sets.

2.2 Procedure for Estimating SPS-HMMs

Acoustic models of the XSAMPA segments are represented simply by LR-HMM with three states and three loops. Each state has a single continuous distribution. We can estimate XSAMPA-HMMs using speech samples with XSAMPA labels by an ordinary HMM training method[8]. Acoustic parameters are the same as those used in paper[4]. We represent acoustic models of the SPSs by the same form of XSAMPA-HMM and initial values of the SPS-HMMs are derived from the XSAMPA-HMMs, given in the following procedure.

(1) Calculating the initial values of SPS-HMMs using values of adjacent XSAMPA segments.
(Typical relationship of location between the XSAMPA segment and the SPS is shown in **Fig.2**.)
(2) Labeling all training speech samples using the above SPS-HMMs.
(3) Reestimating SPS-HMMs using the speech samples labeled in stage (2). Go to (2), until the HMM values are convergent.

The training procedure of (1), (2) and (3), that is, an adaptation technique to a new speech sample set can be described in more detail, but it will be given in another contribution.

4. ENCODING SPEECH SIGNALS INTO SPS SEQUENCES

As describe in section 1., speech signals are once encoded into SPS sequences by SPS unit recognition using SPS-HMMs, where SPS-HMMs are adapted to the corresponding speech signal data environments, such as recording environment, speakers' mother tongue, male/female/child, etc., but not necessary to adapt to individual speakers. We also use SPS label pair grammar in optimizing the SPS sequence.

The following is an example of the encoding result for a real utterance of a sentence:

<u>[i] Original sentence from TIMIT-DB</u>
She had your dark suit in greasy wash water all year. (sa1)
<u>[ii] XSAMPA description converted from TIMIT-DB labeling, slightly modified by our rules</u>
h# S i h E dcl dZ @ r dcl d A kcl k s u q N gcl g r i z i w OO S PAU w OO dcl d @ q OO l j I @ r h#
<u>[iii] SPS sequence converted from the above by rules</u>
h# #S SS Si ii ih hh hE EE EdZ dcl ddZ dZ@ @@ @r rr rd dcl dd dA AA Ak kcl kk ks ss su uu uq qq qN NN Ng gcl gg gr rr ri ii iz zz zi ii iw ww wO OOO OS SS S# PAU #w ww wO OOO Od dcl dd d@ @@ @q qq qO OOO Ol ll lj jj jI II I@ @@ @r rr r# h#
<u>[iv] SPS sequence obtained by SPS-HMM decoding</u>
h# #S SS Si ii ih hh hE EE EdZ dcl ddZ dZi ii id dcl dd dA AA Ak kcl kk ks ss si ii iI II IN NN Ng gcl gg gr rr ri ii iz zz zs ss sI II Iw ww wO OOO OS SS Sw ww wO OOO Od dcl dd dA AA A@ @@ @q qq qO OOO Ol ll lj jj jI II Ir rr r# h#

We can see that the SPS sequence obtained by SPS-HMM decoding shows a sequence similar to that of original labeling. Currently the adaptation is just HMM retraining using additional speech samples, with compensating the mean vectors and variances of SPS-HMMs when the number of SPS label samples are small.

5. SPEECH RETRIEVAL SYSTEM

5.1 System Architecture

The speech retrieval system proposed here can retrieve key phrases, given by speech, from object speech DB by matching in the SPS symbol domain. It characterizes a vocabulary and grammar free system, that is, if the object speech DB have some sections similar to those included in user's key phrase, they can be extracted (see **Fig. 3**), using

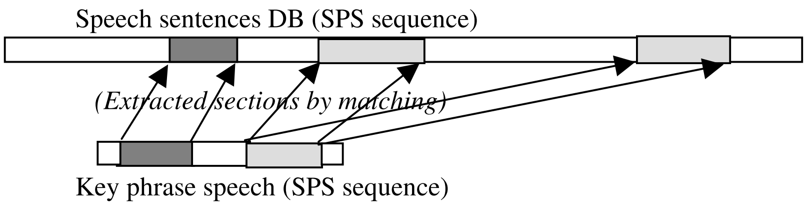

Fig.3 Key phrase spotting used in the speech retrieval system. The key phrase has no explicit boundary in itself.

only the accumulated distance between arbitrary durations of SPS sequences, which is calculated by *Shift Continuous Dynamic Programming* (*Shift-CDP*) [9]. Also it has such performance as effectively working even when the quality or environmental condition of those two speech data is considerably different each other, because the environment-dependent SPS coding plays the role of the normalization.

The system configuration is shown in **Fig. 4**. The procedure is given in the following, where *(A), (B),.., (H)* indicate the block labels in the figure.

(1) Estimate base-form (or standard) SPS-HMMs from a basic training speech sample set, and using these SPS-HMMs, prepare the distance matrix *(H)* for all SPS pairs.

(2) Adapt the base-form SPS-HMMs to the environment of speech DB *(A)* to create SPS-HMM *(E)*. Then encode speech DB *(A)* using the SPS-HMMs to obtain SPS sequence DB *(B)*. The obtained SPS sequence is like that shown in [iv] of section 4, for example.

(3)As the same way, adapt the base-form SPS-HMMs to the environment of the key phrase speech *(C)* to create SPS-HMMs *(F)*, then encode key phrase using the SPS-HMM to obtain SPS sequence *(D)*.

(4)Match SPS sequence *(D)* of the key phrase with the SPS sequence DB *(B)* successively using the Shift-CDP and detect adequate part of sentences from the object speech DB.

In stage (1), it is necessary to define a distance between two HMMs. The distance can be calculated by Kullback-Leibler divergence, but simply approximated as follows: Let us denote the centroid of each state by

$c_{ij}(k)$: centroid vector of the jth distribution in state i of category k

$v_{ij}(k)$: diagonal variance vector corresponding to $c_{ij}(k)$

then we define the distance between HMM- k and -l as

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

where the norms between centroid vectors are normalized by $v_{ij}(k)$.

5.2 Preliminary Experiments

We conducted the following experiments to confirm the feasibility of the proposed method. The basic configuration of the experiments was to extract such sentences as those including key phrases in their parts, from a sentence speech DB. The key phrases were set to one or two parts of the sentences, as described in each experimental condition. Therefore, if we can extract the sentence that include the specified key phrase, then it is “hit”; if not, then “missing”; and “false alarm” is to mis-extract such sentence that includes no key phrase . In the following experiments, the distance matrix *(H)* was calculated using base-form SPS-HMM set. The results depends on the boundary condition, so that the following shows results in adequate conditions.

[Experiment-Ia] The speech sentence DB was ATR-A set which includes 50 sentences[10],uttered by five Japanese male speakers, and the key phrases were those uttered by five female speakers, corresponding to a beginning part of 0 to 2 second in each utterance sample, that is, these phrases are just limited by time, not by vocabulary (see **Fig. 5**) . The number of test queries is 250, complete(100%) hitting gives 1250 (=250x5) samples, and the maximum number of possible false alarms (FAs) is 61250 (=62500-1250).

(Result-Ia) hit: 1174(93.9%), missing: 76 (6.1%), and false alarm: 41.

[Experiment-Ib] The condition was the same as the experiment-Ia, except that the key phrases were chosen from parts of five male speakers’ utterances different from the above. Therefore, both key phrase and DB speech are uttered by males.

(Result-Ib) hit: 1117 (89.5%), missing: 133 (10.5%), and false alarm: 54.

From the above two results, we can see that the performance of *(Result-2)* is almost the same as *(Result-1)*, therefore the acoustic difference between males and females are normalized by the adaptation scheme shown in Fig.4.

[Experiment-IIa] The experimental condition was similar as the Experiment-Ia, but the test data was TIMIT English

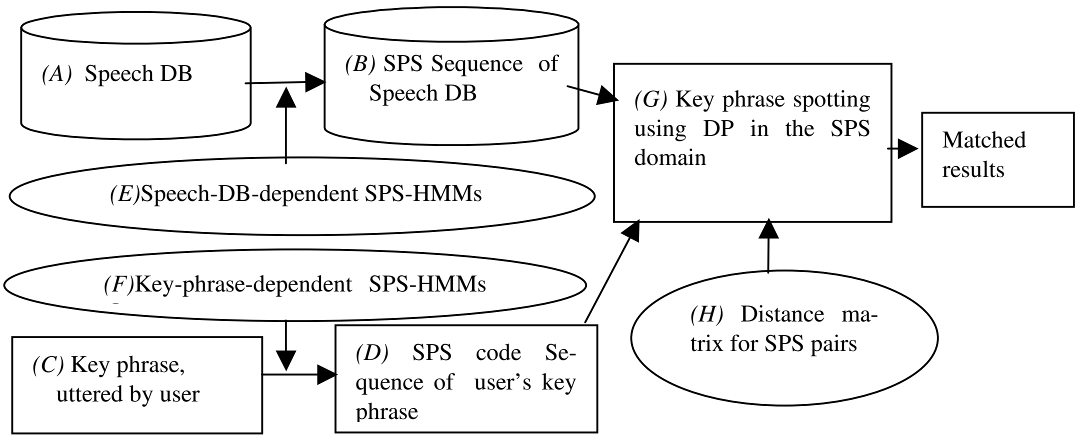

Fig. 4 Block diagram of the proposed speech retrieval system which retrieve user’s key phrase speech from a speech DB.

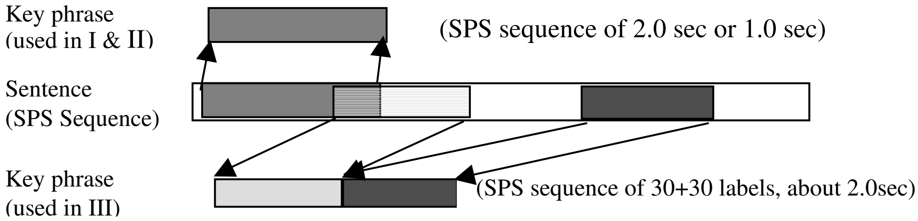

Fig. 5 Illustration for composing key phrases used in the experiments. I, II, III indicate the experiment No.s.

sentence DB released from LDC. It contains 326 speakers (3260 sentences) for training and 112 speakers for test. We used two sentences of the test set as creating the key phrases in the same manner, and chose the training set as the object speech sentence DB. The number of queries is 224, the complete hitting gives 73024 samples, and the max number of possible FAs is 365120.

(Result-IIa) hit: 71699 (98.2%), missing: 1325 (1.8%), false alarm: 811.

[Experiment-IIb] The condition was same as those in [IIa], except that the duration of key phrases was 1.0 sec., shorter than that of [IIa].

(Result-IIb) hit: 69161(94.7%), missing: 3863 (5.3%), false alarm: 18490

[Experiment-III] The used speech data were the same as the experiment-I, except that the number of speakers was 10 males and 10 females. The key phrases were created by connecting two sections of each sentence to be one phrase, as shown in Fig. 5. Each section length corresponded to 30 SPS labels (about 1.0 sec), and matched sections were detected if the accumulated distance of the corresponding section length were under the threshold value. This was a test for effectiveness of the Shift-CDP [9]. In this case, the number of test queries is 100, the complete hitting gives 1000 samples, and the maximum number of the possible false alarm is 49000.

(Result-III) is shown in **Table 1**.

6. CONCLUDING REMARKS

We have proposed a speech processing framework in which all speech data are once encoded into universal sub-phonetic segment (SPS) sequences and speech application systems, such as recognition, retrieval, indexing, are constructed on this SPS domain. The proposed speech retrieval system is still under developing. It will be improved in SPS modeling and phrase spotting algorithm, and applied to more real tasks. We will also construct several other speech application systems, digesting, indexing of speech data, for example, in the near future.

References

[1] http://www.phon.ucl.ac.uk/home/sampa/home

[2] 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.

[3] K. Tanaka, H. Kojima, "Speech recognition based on the distance calculation between intermediate phonetic code sequences in symbolic domain", Proc. of ICSLP 98, pp.361-364, 1998.

[4] K. Tanaka, H. Kojima, "Speech recognition method with a language-independent intermediate phonetic code," Proc. of ICSLP2000, Vo.IV, pp.191-194, 2000.

[5] 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, 1986.

[6] K. Tanaka, H. Kojima, "A between-word distance calculation in a symbolic domain and its applications to speech recognition," International Journal of Information Sciences, Vol. 123, No.1-2, pp.25-41, Elsevier Science, (2000-1).

[7] D. Albesano, R. Gemello, F. Mana, "Hybrid HMM-NN modeling of stationary-transitional units for continuous speech recognition," Proc. ICONIP-97, Vol.2, pp.1112-1115, 1997.

[8] S. Young, *The HTK Book*, Entropic Cambridge Research Lab,1996.

[9] Y. Itoh, K.Tanaka, "Automatic Labeling and Digesting for Lecture Speech Utilizing Repeated Speech by Shift CDP," Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'2001), (to be appeared in Sep. 2001).

[10] T. Kobayashi, S.Itahashi, et al, "ASJ Continuous Speech Corpus for Research,"(in Japanese) J. Acoust. Soc. Japan, Vol.48, No.2, pp.888-893, 1992.

Table 1 Result of experiment-III, where complete hit gives 1000 samples.

Thresh-old	short length*		long length**	
	hit	false alarm	hit	false alarm
0.7	722	6	293	0
0.8	804	15	396	0
0.9	862	31	500	0
1.0	915	78	602	0
1.1	943	183	693	0
1.2	964	440	768	2
1.3	983	1022	845	17

* Sentence detection is done if the number of SPS labels contained in the matched section(s) is more than 25.

** The number is more than 40.