Phone Vector DHMM to Decode a Phone Recognizer's Output

Bong-Wan Kim[1], *Dae-Lim Choi*[1], *Yongnam Um*[1], *Yong-Ju Lee*[2]

[1]Speech Information Technology & Industry Promotion Center, Wonkwang Univ., Korea
[2]Division of Electrical Electronic and Information Engineering, Wonkwang Univ., Korea
{bwkim,dlchoi,umyongnam}@sitec.or.kr, yjlee@wonkwang.ac.kr

ABSTRACT

In this paper we introduce a Phone Vector Discrete HMM (PVDHMM) that decodes a phone recognizer's output. The proposed PVDHMM treats a phone recognizer as a vector quantizer whose codebook size is equal to the size of its phone set. To examine the proposed method we perform two experiments. First, the output of a phone recognizer is recognized by the PVDHMM, and its results are compared with those of a continuous speech recognizer (CSR). Second, to investigate its potential application in the field of open-vocabulary spoken document retrieval, a retrieval experiment through word spotting is carried out on the output of a phone recognizer, and its results are compared with those of retrieval through the phone-based vector space model.

Index Terms: acoustic modeling, PVDHMM, phone sequence recognition.

1. INTRODUCTION

Recently, stimulated by the increase of multimedia data, a number of studies on the spoken document retrieval (SDR) have been done in order to develop how to search in the stored data for speech data which are related to the query. The SDR is to retrieve multimedia information contents by means of speech recognition and information retrieval techniques.

There are several approaches to SDR. One of them is the conventional text retrieval that uses keyword transcriptions of spoken documents that are obtained through keyword spotting technique. Another one is the text retrieval that uses word transcriptions that are obtained through large vocabulary continuous speech recognition (LVCSR) system. These two approaches have problems in that keywords or vocabulary should be known in advance for indexing spoken documents. However, out-of-vocabulary (OOV) words may be deleted or replaced during speech recognition. Furthermore, more significant problems arise when query words are OOV. Thus, researchers have studied on how to use phone sequences generated through speech recognizers to avoid the problem of OOV in word-based approaches [1,2,3]. Researchers have also tried to use Probabilistic String Matching (PSM) [1], phone n-grams [2], or phone confusion probability [3] to overcome the problem of errors due to low recognition rate since recognition rates of phone recognizers are low in general.

If certain ways are found in which the results of the phone recognizer including its errors can be decoded, one could use them to complement a word-based approach. With this goal in mind, we introduce a Phone Vector Discrete HMM (PVDHMM) to decode the results of the phone recognizer. The proposed PVDHMM treats the phone recognizer as a vector quantizer whose codebook size is equal to the number of phones, and it uses discrete feature vectors generated from the results of the phone recognizer. In this paper we use Korean as the language for experiments.

This paper is organized as follows. In Section 2, we describe the text corpus and speech corpus that are used for the experiments in this paper. In Section 3, we describe the CSR that is used for comparing the performances of PVDHMM and the phone recognizer used for training PVDHMM. In Section 4, we describe PVDHMM and its training and testing procedures. In Section 5, we describe the results of the experiments. Finally, in Section 6 we come to the conclusion.

2. TEXT AND SPEECH CORPORA

2.1 Text corpora (KSR-2002-TN) [4]

For the language model we have used this text corpus which was created by the Speech/Language Technology Research Department of the Electronics and Telecommunications Research Institute (ETRI). The corpus is the collection of 20M words or 1.4M sentences from the texts of Korean daily newspapers from January 1, 2000 to December 31, 2000.

2.2 Speech corpora [5]

2.1.1 Speech database for Korean dictation (Dict01)

This is a speech corpus created and distributed by the Speech Information Technology and Industry Promotion Center (SiTEC) in Korea. It is the collection of speech of 400 speakers (about 105 sentences per speaker). 20,833 sentences composed of highly frequent 10K words selected from a text corpus of 40M words were used for prompts. The speech data was recorded through the Andrea ANC 750 microphone. In this paper, the speech data of 360 speakers (66.95 hours) from this corpus is used for training acoustic models, and the speech data of 40 speakers (7.87 hours) is used for testing.

2.1.2 Speech database for Korean address (Addr01)

This is a speech corpus created and distributed by the SiTEC to develop the location based service system such as the navigation system. Prompts contain 2,110 phrases such as addresses, names of apartments, and buildings. The speech data of 300 speakers (about 140 phrases per speaker) was recorded through the Labtec Axis-301 microphone. This is used for testing the performance of the proposed model through word spotting on the phone sequences in the retrieval task. Data which contains 4 sets of 2,110 phrases read by 60 speakers (5.98 hours) is used for testing.

3. PHONE AND WORD RECOGNIZERS

3.1 Phone set

The phone set consists of 46 phones (including silence and short pause). The plosives in the Korean syllable final are modeled differently from those in the syllable initial to distinguish them.

3.2 Acoustic models

Two versions of acoustic model - context independent (CI) and context dependent (CD) models - were created. The speech data of 360 speakers from Dict01 was used for training. The CI model has 128 mixtures per state, and the CD model is the cross-word triphone which has 16 mixtures per state. HTK [6] was used for training and testing. The parameter used for acoustic models was "MFCC_E_D_N_Z" and the number of coefficients was 25.

3.3 Language models

We created a phone 2-gram language model for phone recognizer and a 2-gram language model for highly frequent 10K words for CSR by using KSR-2002-TN. We used the CMU-Cambridge Statistical Language Modeling toolkit [7] for language modeling.

3.4 Phone recognizers and a CSR

We created two versions of the phone recognizer: CIPHONE using the CI acoustic model and CDPHONE using the CD acoustic model. Performance of each phone recognizer for the Dict01 test set is shown in Table 1 below. In order to compare the performances of PVDHMM in the CSR task, we created a CSR system – CSR10K – that used the CD model for the acoustic model and highly frequent 10K words from the corpus KSR-2002-TN for the vocabulary.

Table 1: Performance of each phone recognizer for the Dict01 test set when its beam width is 100.

Model	Phone Corr. (%)	Phone Acc. (%)	Realtime factor
CIPHONE	72.55	69.46	0.34
CDPHONE	83.06	80.10	1.01

4. PHONE VECTOR DHMM

4.1 PVDHMM

PVDHMM is a DHMM that treats a phone recognizer as a vector quantizer whose codebook size is equal to the size of its phone set. Discrete feature vectors are created for training and testing the PVDHMM by using phone symbol sequences which are produced by the phone recognizer. Two problems should be solved in order to convert phone symbol sequences into discrete feature vectors. First, phone symbol sequences should be converted into VQ indices. We solve this problem simply by obtaining the index of each phone symbol from the sorted phone list. Second, what is more crucial, the problem should be solved that the length of generated discrete feature vectors becomes shorter than the number of the states of the PVDHMM. We solve this problem by obtaining the information on the time alignment of each phone at the stage of training and repeating VQ indices as many times as the number of frames. In case of the lack of information on the time alignment in the phone sequences, we solve the problem by using the mean frame length information on each phone obtained in the training procedure. Figure 1 below shows the process by which discrete feature vectors are generated from the phone sequence "A B C" in case of the lack of time information, on the assumption that the index of 'A' is 1 on the sorted phone list and the mean length of the phone is 4 frames.

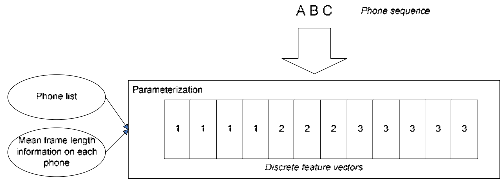

Figure 1: An example of conversion from a phone sequence to discrete feature vectors in case of the lack of information on phone length.

4.2 Training and testing procedures

The following procedures are followed to train the PVDHMM.

A. Reference phone labels are created from speech data for training. Reference labels are obtained from the forced alignment by a speech recognizer or by manual labeling.

B. Phone transcriptions with time information for training data are obtained by using a phone recognizer.

C. Discrete feature vectors are created from the phone transcriptions. At this point, the mean frame length of each phone is calculated to parameterize phone sequences without time information.

D. The PVDHMM is trained with the created discrete feature vectors. The procedure by which the PVDHMM is trained is the same as for a conventional DHMM.

The procedures for training the PVDHMM are diagrammed in Figure 2.

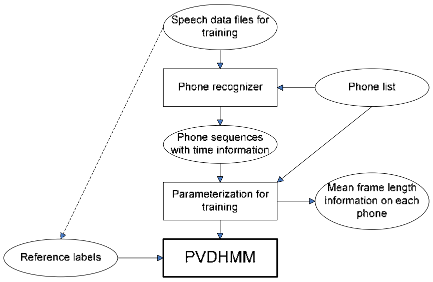
Speech data files for training
Phone recognizer
Phone list
Phone sequences with time information
Parameterization for training
Mean frame length information on each phone
Reference labels
PVDHMM

Figure 2: Training procedure for PVDHMM.

To test PVDHMM, phone sequences are converted into feature vectors by using the mean frame length information on each phone which is obtained during the process of training, and created feature vectors are recognized by the trained PVDHMM.

5. EXPERIMENTAL RESULTS

5.1 CSR task

To verity the effect of the performance of phone recognizer, we trained two versions of PVDHMM acoustic model through the training procedure described above: the one is trained on phone sequences generated from the CIPHONE phone recognizer which uses monophone acoustic model having 128 mixtures per state, and the other is trained on phone sequences generated from the CDPHONE phone recognizer which uses crossword triphone acoustic model having 16 mixtures per state. Performances of both phone recognizers (CIPHONE, CDPHONE) for Dict01 test set were shown in Table 1. To generate phone sequences to train PVDHMMs, phone recognition experiments were performed on the Dict01 training set which was used to train each phone recognizer itself, and PVDHMMs were trained with the 1-best phone recognition results. Two versions of PVDHMM are both CI models with 45 phone set excluding short pauses from recognition results of each phone recognizer.

We created two versions of phone sequence recognizer (PSR): PSR10K_CIPHONE which uses PVDHMM acoustic model trained on phone sequences generated from phone recognizer CIPHONE, and PSR10K_CDPHONE which uses PVDHMM acoustic model trained on phone sequences generated from phone recognizer CDPHONE. For the language model of both PSRs, we used the same language model as for the CSR10K -- 2-gram language model for highly frequent 10K words. Both were tested with phone sequences without time information which each phone recognizer produced from Dict01 test set. Performances of PSR10K_CIPHONE and PSR10K_CDPHONE as compared with CSR10K are shown in Figure 3.

We can see that the performance of the phone recognizer used for training the PVDHMM can affect that of the PSR, seeing that the performance of the PSR10K_CDPHONE is better than that of the PSR10K_CIPHONE. Although PSR10K_CDPHONE uses CI acoustic models, it does not show much difference in word accuracy from the CSR10K, and when the beam width is 75, its word accuracy is better (7.07%) than CSR10K. We can also see that the speed of PSR10K_CDPHONE at the beam width of 75 is 2.3 times higher than the speed of CSR10K at the beam width of 100, whereas word accuracy shows 2.72% decrease. As the beam width is widened, the CSR10K shows significant improvement in word accuracy, whereas the PSRs show just a little improvement in word accuracy and the speed of recognition decreases more rapidly than CSR10K. This suggests that some techniques should be devised to improve the speed and performance of PSR.

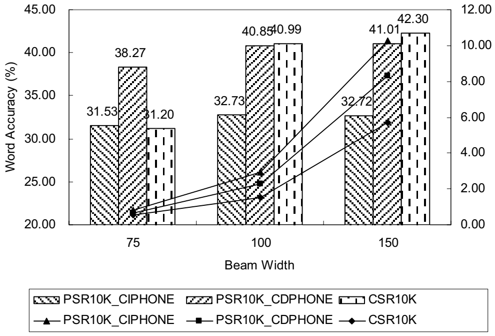

Figure 3: Performances of phone sequence recognizers as compared with CSR10K.

5.2 Retrieval task using PVDHMM word spotter

To evaluate the potential of SDR using PVDHMM, retrieval through word spotting is performed by PVDHMM on the output of a phone recognizer, and its results are compared with those of retrieval through the conventional phone-based vector space model (VSM) [8], which is usually used for SDR. For retrieval experiment on the Addr01 test date set, we selected ten queries which were composed of 4 ~ 8 phones.

VSM creates a space in which both documents and queries are presented by vectors. Given a query *Q* and a document *D*, two *T*-dimensional vectors *q* and *d* are generated, where *T* is the total number of possible indexing terms. Each component of *q* and *d* represents a term frequency. The inner product of *q* and *d* is then used to estimate a measure of similarity between query *Q* and the document *D*. To combine phone N-grams of different lengths, we used the relevance score in [3] as follows:

$$S_{1,2,3}(q,d) = \frac{1}{6}\sum_{N=1}^{3} N \cdot S_N(q,d) \qquad (1)$$

Where S_N represents the relevance score obtained with the set of *N*-gram indexing terms.

After recognition was performed by the CIPHONE and CDPHONE on the speech data (a total of 8,247 sentences) of 60 speakers from Addr01, retrieval performances (VSM_CIPHONE and VSM_CDPHONE) were measured by obtaining the relevance scores for queries. On the Addr01 data set, the phone accuracy of the CIPHONE is

59.62% and that of the CDPHONE is 64.06%. To examine the upperbound performance of the phone-based VSM, retrieval performance (VSM_NOERROR) was measured with the error-free phone transcriptions of prompting items.

For retrieval using the proposed model, we composed the network as in Figure 4 below, and we performed query word spotting experiment on the recognition results of the CIPHONE and CDPHONE, using two PVDHMMs trained as described in subsection 5.1 (PVSPOT_CIPHONE and PVSPOT_CDPHONE). Retrieval performance was measured by using the normalized acoustic score for the query word as the relevance score. Normalized acoustic score is the acoustic score divided by the duration of the segment. Mean realtime factor of word spotting tests for each query is 0.01.

We evaluate the retrieval performance by means of precision, recall, and mean average precision (mAP) as used in TREC [9]. A perfect retrieval system would result in a mAP measure of 1.

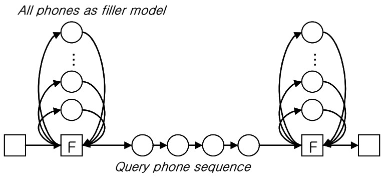

Figure 4: Query word spotting network using PVDHMM.

Recall-Precision Graph of the experimental results is shown in Figure 5. The mAP measure of VSM_NOERROR is 0.95, whereas those of VSM_CIPHONE and VSM_CDPHONE are 0.35 and 0.53 respectively. We can see that deterioration in performance due to the recognition errors of the phone recognizer is not trivial. The mAP measures of PVSPOT_CIPHONE and PVSPOT_CDPHONE are 0.67 and 0.71, showing 62% relative improvement on average in performance, which indicates that they can be used for complementing the low performance of the phone recognizer.

6. CONCLUSION

We have introduced the PVDHMM for decoding the phone sequences which are the output of the phone recognizer. Using the proposed model, we performed CSR experiment and retrieval experiment through word spotting, and we compared their results with those of other conventional models. By using the proposed model, indexing and retrieval functions for speech data can be performed on a well known HMM framework. Further studies on a variety of other methods may be required in order to improve the performance of the PVDHMM. Though we used only the 1-best results of the phone recognizer in the experiments, the further experiment that uses the N-best results may be necessary for improvement as well as the experiment to improve the speed by reducing the vector length appropriately in generating discrete feature vectors from phone sequences. A hierarchical approach for SDR may be tried which uses word transcriptions that are obtained through LVCSR system if query words are in-vocabulary, and which uses PVDHMM and phone sequences converted from the word transcriptions if query words are OOV words.

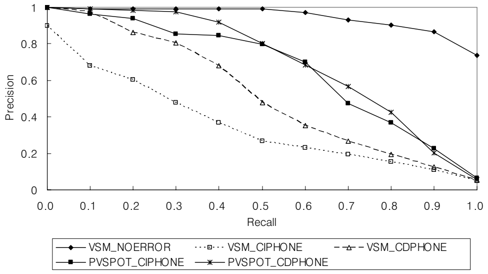

Figure 5: Retrieval performances of PVDHMM spotters, compared with phone-based VSM models.

7. ACKNOWLEDGEMENTS

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (the Center for Healthcare Technology Development, Chonbuk National University, Jeonju 561-756, Korea).

8. REFERENCES

[1] Martin Wechsler, *Spoken Document Retrieval Based on Phoneme Recognition*, PhD Dissertation, Swiss Federal Institute of Technology (ETH), Zurich, 1998

[2] Kenny Ng, *Subword-based Approaches for Spoken Document Retrieval*, PhD Dissertation, Massachusetts Institute of Technology (MIT), Cambridge, MA, 2000

[3] Nicolas Moreau, Hyoung-Gook Kim, Thomas Sikora, "Phone-based Spoken Document Retrieval in Conformance with the MPEG-7 Standard," 25th International AES Conference (Metadata for Audio), London, UK, 2004.

[4] Speech/Language Technology Research Department in ETRI, http://voice.etri.re.kr

[5] SiTEC (Speech Information Technology and Industry Promotion Center), http://www.sitec.or.kr

[6] HTK (Hidden Markov Model Toolkit), http://htk.eng.cam.ac.uk

[7] CMU-Cambridge Statistical Language Modeling toolkit, http://mi.eng.cam.ac.uk/~prc14/toolkit.html

[8] Gerald Salton, Michael J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983

[9] TREC, Common Evaluation Measures, NIST, 10th Text Retrieval Conference (TREC 2001), Gaithersburg, MD, 2001