ROBUST SPOKEN DOCUMENT RETRIEVAL BASED ON MULTILINGUAL SUBPHONETIC SEGMENT RECOGNITION

Shi-wook Lee, Kazuyo Tanaka
National Institute of Advanced Industrial Science and Technology, JAPAN
AIST Central # 2, Umezono 1-1-1, Tsukuba, 305-8568, Japan

Yoshiaki Itoh
Iwate Prefectural University, JAPAN
Sugo, 152-52, Takizawa, IWATE, 020-0193, Japan

Keywords: Spoken document retrieval, Multi-lingual, Open-vocabulary Speech Recognition, Subphonetic unit.

Abstract: This paper describes the development and application of a subphonetic segment recognition system for spoken document retrieval. Following from the development of an open-vocabulary spoken document retrieval system, where the retrieval process is accomplished in the symbolic domain by measuring the distance between the parts of subphonetic segment results from pattern recognition in the acoustic domain, the system proposed here performs matching based on subphonetic segment as more basic unit than the semantic unit. As such, the system is not constrained by vocabulary or grammar, and can be readily extended to multilingual tasks. This paper presents the proposed spoken document retrieval system including the proposed subphonetic segment recognition scheme, and evaluates the performance and feasibility of the system through experimental application to multilingual retrieval tasks.

1 INTRODUCTION

Recently, information retrieval techniques have been widely adopted for text databases to identify documents that are likely to be relevant to text queries. The aim of spoken document retrieval (SDR) is to provide similar functionality for databases of spoken documents. Such spoken documents, stored in the form of audio signals, may be collected from many different sources, such as news broadcasts on radio and television, voice/video e-mail, and multimedia material on the Web. Furthermore, as the volume of such accessible multimedia databases continues to grow, the demand for user-friendly methods to access, process and retrieve the data has become increasingly important. Therefore, it has become indispensable to be able to retrieve such documents in response to speech queries as well as text queries.

Overall, subword-based retrieval is not as effective as word-based retrieval, but is helpful when the word-based speech recognition output is prone to error or undesirable, as may occur in out-of-vocabulary (OOV) problems and multilingual tasks. With current technology, there is a practical limit on the size of the vocabulary. When new speech queries that are not in the pronunciation dictionary (lexicon) are input, the system undesirably replaces the query with the most probable word. If the system is more sophisticated, it regards the input query as unknown, and fails to offer a result, without providing further details. The system can never output a string of subwords that is not listed in the pronunciation dictionary. Large-vocabulary continuous speech recognition (LVCSR) systems cannot deal with articulation at the subword level unless subword units are used as the fundamental units. The video mail retrieval project is addressing this requirement by developing systems to retrieve stored video material using the spoken audio sound track(Jones, 1996).

In the task of cross-language retrieval, non-native pronunciation characteristics (i.e., foreign accents) in foreign language speech lead to extremely poor performance in SDR. For example, English uttered by a Japanese speaker will retain many Japanese speech characteristics (Japanese-English). Therefore, it is desirable that the system can deal with those speech queries with foreign accents. As the prevalence of multimedia material on the Web continues to grow, the demand for multilingual SDR systems is rapidly strengthening. The development of multilingual SDR systems will therefore be of significant benefit for multilingual task, and will parallel the development of retrieval in these systems.

To accomplish cross-media tasks, for example text

Lee S., Tanaka K. and Itoh Y. (2004).
ROBUST SPOKEN DOCUMENT RETRIEVAL BASED ON MULTILINGUAL SUBPHONETIC SEGMENT RECOGNITION.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 134-139
DOI: 10.5220/0002636201340139

queries of spoken documents and speech queries of text documents, the choice of suitable subword units for multimedia retrieval is important. The advantage of subword units is that the transcript is readable by humans and can be used to translate text queries into subword sequences so as to be acceptable in SDR.

The present authors have been developing an SDR system in which retrieval is conducted by calculating the distance between the parts of a subphonetic segment(SPS) sequence extracted from underlying speech recognition. As the system is based on matching SPS sequences directly, the system is not constrained in terms of vocabulary or grammar, and is robust with respect to recognition error(Tanaka, 2001)(Lee, 2002). Most existing SDR systems are based on matching text, and speech recognition systems usually employ the integration of likelihood values of acoustic phoneme sequences given from a top-down hypotheses. Thus, it should be possible to merge both acoustic and symbolic processing simultaneously. In this work, the feasibility of subphonetic units for retrieval in an SDR system is investigated. The effect of varying the distance measure is also examined in an attempt to improve the performance of the shift continuous dynamic programming (Shift-CDP) matching based on SPS sequences. Finally, SDR experiments are conducted to evaluate the performance of the proposed system in both monolingual and multilingual tasks.

2 SPOKEN DOCUMENT RETRIEVAL SYSTEM

A spoken document database containing a significantly high proportion of OOV words is assumed, such as names and places. Such words will be susceptible to poor retrieval performance due to misrecognition. Speech retrieval is similar to text retrieval, except for a number of difficulties in actual application such as accurate detection of word boundaries, recognition errors, and acoustic mismatching. For this reason, existing SDR systems perform retrieval using a text-based database linked to multimedia material in the speech-based database. The SDR system proposed here aims to retrieve speech keyphrases directly from the object multimedia database. In the system, if the object multimedia database has parts similar to those included in the input queries, the relevant data can be retrieved using only the accumulated distance between arbitrary durations of SPS sequences. Such a scheme is suitable for an open-vocabulary system. This function can be performed by applying Shift-CDP for optimal matching between SPS sequences. This is an essential difference from the conventional speech processing methods. In the proposed system, the input utterance is first encoded in terms of acoustic features. Then, the SPS extracted by a recognizer is transferred to Shift-CDP(Tanaka, 2001). Figure 1 shows the overall block diagram of the proposed SDR system.

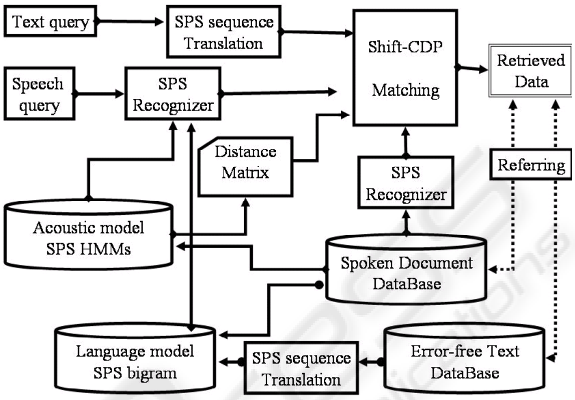

Figure 1: Block diagram of proposed SDR system based on subphonetic segments

3 SUBWORD UNITS

In order to allow user-friendly queries of a multimedia database, speech signals are converted into words, phonemes, or other subword units, using a speech recognition system. This work focuses on a SPS-based approach, where spoken documents are recognized as SPS sequences and the retrieval process is carried out based on matching the dynamic programming scores of these transcriptions. Although word-based approaches have consistently outperformed phoneme approaches(Voorhees, 1998), there are several compelling reasons for using SPS, as mentioned above.

The present authors have been developing an architecture for speech processing systems based on the universal phonetic code (UPC)(Tanaka, 2001). All of the speech data in the systems are once encoded into UPC sequences, and then the speech processing systems, such as recognition, retrieval, and digestion, are constructed in the UPC domain. The international phonetic alphabet (IPA) or extended speech assessment methods phonetic alphabe (XSAMPA) is the candidate set for the UPC set. Here SAMPA is a machine-readable phonetic alphabet. The SPS is derived from XSAMPA and is refined under the consideration of acoustic-articulatory effects. For example, the XSAMPA (i.e., IPA) contains partly extra-detailed categorization to be modeled in an engineering sense. Therefore, only primary IPA symbols are adopted, and minor phonetic variations are represented by statistical distributions in the acoustic domain. A simple example of an SPS converted from XSAMPA sequences consisting of stationary and non-stationary segments in the speech stream is given below. The advantage of training SPS models is that pronunciation variation is trained directly into the acoustic model, and does not need to be modeled separately in the dictionary.

- Speech: *She had your dark* $\cdots$
- XSAMPA-Phoneme: *# S i h E dcl dZ @ r dcl d A kcl k* $\cdots$
- SPS: *# #S SS Si ii ih hh hE EE EdZ dcl dZdZ dz@ @@ @r rr rd dcl dd dA AA Ak kcl kk* $\cdots$
- Core SPS: *# SS ii hh EE dZdZ @@ rr dd AA kk*$\cdots$

where # denotes a pause or silence interval. A total of 429 SPSs are extracted from the 43 phonemes for Japanese (including 3 silence types), and 1352 SPSs are extracted from the 42 phonemes of English (including 3 silence types). Theoretically, 1610 SPSs can be extracted from the 42 English phonemes,however, some concatenations of phonemes do not exist in real language. The remarkably fewer Japanese SPSs is due to the fact that most Japanese syllables consist of 1 consonant and 1 vowel (C+V). Therefore, concatenations of consonants are very rare in Japanese, resulting in a lesser degree of acoustic-articulation than in English. Acoustic models of English and Japanese are simply represented by a left-to-right hidden Markov model (HMM) with 3 states, each with a single mixture diagonal distribution for simplicity.

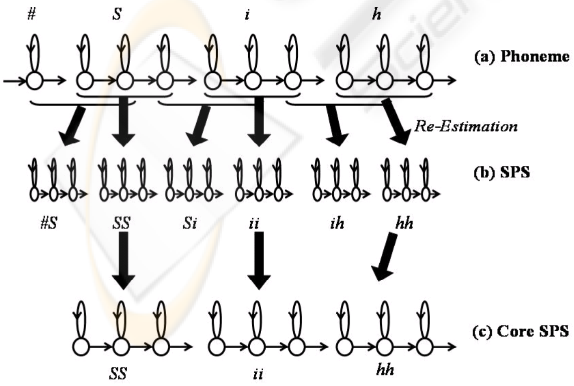

Figure 2: Acoustic models of subword units: phoneme, SPS, core SPS

4 SHIFT CONTINUOUS DYNAMIC PROGRAMMING

When subword sequences are recognized directly, with higher error rates than for words, selection of a good matching approach becomes much more important. Shift-CDP is an algorithm that identifies similar parts between a reference pattern R_N and the input pattern sequence I_T synchronously. The pre-fixed part of the reference pattern, called the unit reference pattern (URP), is shifted from the start point of the reference pattern to the end by a certain number of frames. The matching results for each URP in the reference pattern are then compared and integrated. Shift-CDP is an improved version of the reference interval-free CDP (RIFCDP), and performs matching between arbitrary parts of the database and arbitray parts of the query input(Itoh, 2001).

$$R_N = \{R_0, \cdots, R_\tau, \cdots, R_{\tau+r}, \cdots, R_{N-1}\} \quad (1)$$
$$I_T = \{I_0, \cdots, I_t, \cdots, I_{t+i}, \cdots, I_{T-1}\} \quad (2)$$

The first URP is taken from R_0 in the reference pattern R_N. The next URP is then composed of the same number of N_{URP} frames from the $(N_{shift}+1)_{th}$ frame. In the same way, the k_{th} URP is composed of N_{URP} frames from the $k \times (N_{shift}+1)_{th}$ frame. Thus, the number of URPs becomes $[N/N_{shift}]+1$, where [] indicates any integer that does not exceed the enclosed value. Shift-CDP is then performed for all URPs in the reference R_N. It is not necessary to normalize each cumulative distance at the end frame of a URP because all URPs are of the same length. Actually, Shift-CDP is a very simple and flat algorithm that performs CDP for each URP and integrates the results.(Itoh, 2001)

5 DISTANCE MEASURE

In the Shift-CDP algorithm, the DP matching score is calculated using a pre-measured SPS distance matrix. Therefore, the system is directly influenced by the distance measure, and selecting a proper measure is important for the performance. Distance measures have been widely applied in a number of speech technologies. For speech coding, distance measures are used in the design scheme for vector quantization algorithms and as objective measures of speech quality. In speech and speaker recognition, the spectral difference between two speech patterns is measured to compare patterns and make similarity decisions. Motivated by these speech recognition techniques, some unit-selection algorithms for speech synthesis and optimal-joining algorithms now use the distance measure between feature vectors. Here, the distance measures D_{AB} between two multivariate Gaussian distributions, $N(\mu_A, \Sigma_A)$ and $N(\mu_B, \Sigma_B)$, are considered. The Bhattacharyya distance D_{BHAT}, which is covered in many texts on statistical pattern recognition(Fukunaga, 1990), is a separability measure between two Gaussian distributions:

$$D_{BHAT} = \frac{1}{N}\sum_{n=1}^{N}\frac{(\mu_{An}-\mu_{Bn})^2}{8}\left[\frac{\Sigma_{An}+\Sigma_{Bn}}{2}\right]^{-1} + \frac{1}{2}ln\frac{|\frac{\Sigma_{An}+\Sigma_{Bn}}{2}|}{\sqrt{|\Sigma_{An}||\Sigma_{Bn}|}} \qquad (3)$$

where N is the number of HMM states and $N = 3$ states is used throughout this work. The first term of Eq. (3) provides the class separability from the difference between class means, while the second term gives the class separability from the difference between class covariance matrices. Here, considering the insufficiency of training data, the distance measure derived directly from the difference between class mean is adopted, that is, the first term of Eq. (3), as formulated below. This distance is very close to the weighted Mahalanobis distance.

$$D_{AB} = \frac{1}{N}\sum_{n=1}^{N}\left((\mu_{An}-\mu_{Bn})^2\frac{\Sigma_{An}+\Sigma_{Bn}}{2}\right)^{-1} \qquad (4)$$

Figure 3 shows the phonetic distance matrix between English and Japanese phonemes measured by Eq. (4). The radius of the spots is linearly proportional to the distance between the English and Japanese phoneme HMMs. Thus, a larger spot area indicates a longer distance.

6 EXPERIMENTAL EVALUATION

6.1 Multilingual Corpus and Models

For research into the underlying speech recognition and information retrieval technologies based on subword units, it was necessary to prepare a sufficient corpus of spoken documents. The observation vector consists of 12th-order mel-cepstra, their delta, power, and its derivative. Thus, a 26-dimensional feature vector is extracted from each 5 ms analysis frame. HMMs for English and Japanese are estimated separately on language-dependent native-speaker speech data as in typical monolingual speech recognition. The phoneme models used here were 42 monophones (including 3 silence types) for English, and 43 monophones (including 3 silence types) for Japanese. The English phoneme models were first estimated from TIMIT phonetically labeled data. The acoustic models for the 1352 English SPSs were estimated using Wall Street Journal data (WSJ0). In the case of Japanese acoustic models, Japanese newspaper articles sentences (JNAS) were used to obtain phonemes and the 429 SPSs. For all training utterances, phoneme and SPS sequences were generated from the text transcription and a dictionary. The Carnegie Mellon University (CMU) pronunciation dictionary (120,000 words) was used for English, and the IPA pronunciation dictionary (60,000 words)(Kawahara, 1998) was used for Japanese to translate all English and Japanese spoken documents into phoneme/SPS sequences for forced-alignment in training acoustic models and subword n-gram language models. English acoustic phoneme models were primarily built using the TIMIT 61 label set. In order to use the CMU pronunciation dictionary (39 phonemes) to estimate English SPS models, the English phoneme models were collapsed down to the 39 labels. To increase underlying subword recognition accuracy, phoneme/SPS bigram language models were also estimated from the same corpus used to train the acoustic models. The training material used for acoustic and language models is summarized in Table 1.

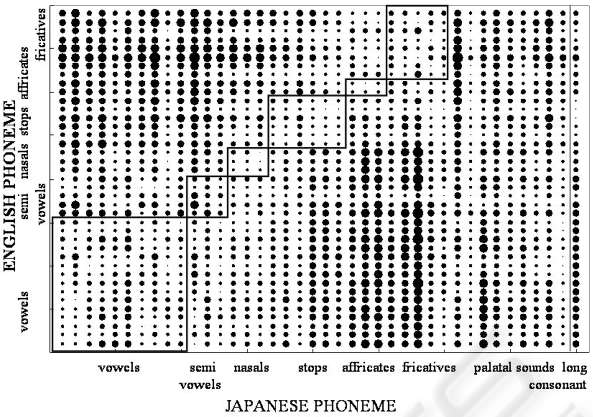

Figure 3: Phonetic distance matrix between English (# 39) and Japanese (# 40) phonemes calculated by Eq.(4). The 3 silence models are not presented here.

6.2 Experimental Results

A set of 10 short keyphrase queries were prepared for SDR experiments. Each query had 9 relevant documents in each language-dependent target database of 2000 sentences. The input queries and target database for experimental evaluation are detailed in Table 2.

The underlying recognition system for decoding

Table 1: Training material used for acoustic and language models

Language	English	Japanese
# Sentences	29150	13786
Length	55.68 h	25.62 h
# Phonemes	42	43
# SPSs	1352	429

Table 2: Analysis of test material for monolingual and multilingual SDR tasks

Speaker	Japanese	Japanese	English
Speech	Japanese	English	English
Ave. length	1.19 s	2.00 s	1.01 s
# Phonemes	12.4	7.9	8.4
# SPSs	26.0	19.1	19.4
# Core SPSs	11.3	8.7	9.0
# Relevant	9	9	9
Target DB	Japanese	English	
# Sentences	2000	2000	
Length	3.29 h	1.87 hr	

subwords was a single-pass beam search decoder based on the JULIUS system(Kawahara, 1998). Table 3 summaries the error rates for each language and subword unit.

Table 3: Subword error rates (%) for each language and subword unit

	English	Japanese
Phonemes	42.97	45.76
SPSs	50.90	46.21
Core SPSs	41.66	35.66

Figures 4 and 5 show the language-dependent monolingual SDR performance for English and Japanese according to subword units. The speech of input queries and target DBs were uttered by native speakers. Both in English and Japanese SDR experiments, the SPS-based SDR outperformed the phoneme-based and core SPS-based schemes remarkably. Longer subword units can capture word or phrase information, while shorter units can only model word fragments. The trade-off is that the shorter units are more robust to error and word variants than the longer units. There were no significant differences between the performance of SDR based on phonemes and core SPSs. As seen in Tables 2 and 3, although the error rates when using core SPSs are lower than for the use of phonemes, the SDR performance is heavily dependent on the number of subwords per unit time. This also demonstrates that the amount of information per unit time is an essential consideration in subword-based SDR. The performance of SDR for English is worse than that for Japanese, due to the extremely large number of variant pronunciations in English, and the larger amount of information per unit time in Japanese. The former cause can be counteracted to some extent by preparing various pronunciations in a dictionary, however, cross-word coarticulation cannot be predicted.

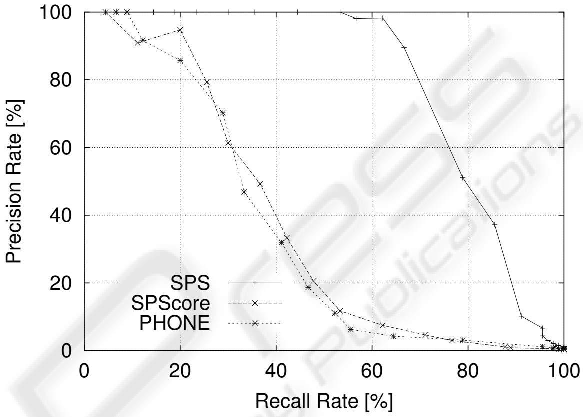

Figure 4: Performance of Japanese SDR according to subword units, phoneme(PHONE), subphonetic segment(SPS), and core SPS(SPScore)

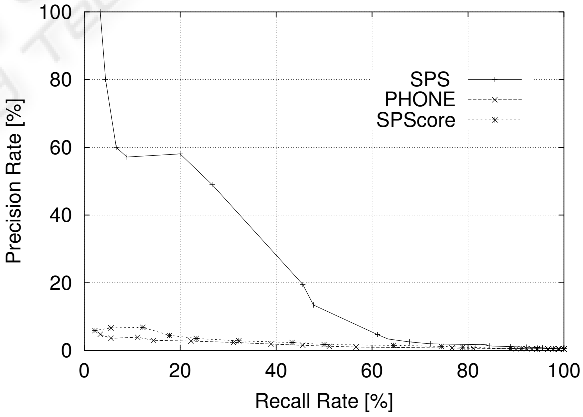

Figure 5: Performance of English SDR according to subword units

The two HMM sets, Japanese and English, are used together in English SDR tasks for Japanese-English, labeled as $EJ.model$ in Figure 6. The acoustic $EJmodel$ contains 1778 SPS HMMs from the 1352 English HMMs, and 429 Japanese HMMs. The 3 silence models are chosen from the English models. Despite the simple method for combining the two acoustic models, the performance is improved considerably for Japanese-English tasks. This result demonstrates that this approach is a feasible means of handling foreign accents in multilanguage SDR.

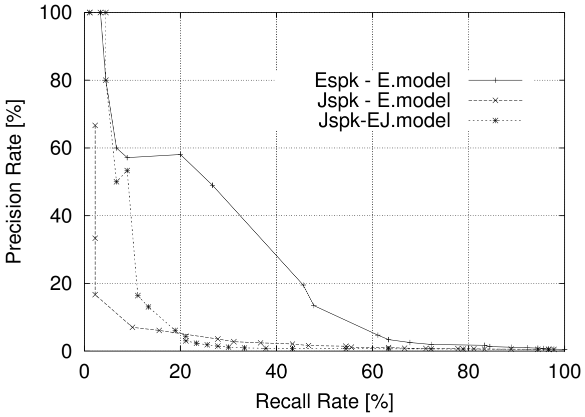

Figure 6: Performance of SPS-based English SDR according to speaker type and acoustic model. Native English speaker (Espk) and Japanese speaker (Jspk), English acoustic model (E.model), and Multi-lingual acoustic model (EJ.model)

7 CONCLUSIONS

This paper presented the development of an open-vocabulary SDR system and the use of SPSs as a new subword unit. Experimental evaluation demonstrated that the SPS-based approach significantly improves the performance of both monolingual and multilingual SDR. Future work will concentrate on extending the system to other languages, as well as coupling the scheme with LVCSR-based systems.

REFERENCES

E. Voorhees and D. Harman (1998). “Overview of the Seventh Text REtrieval Conference” In *Proc. of the 7th Text Retrieval Conference (TREC-7)* pp. 1–24 .

K. Ng (2000). “Subword-based approaches for Spoken Document Retrieval” In *Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA* .

M. A. Siegler, et al. (1997). “Automatic Segmentation, Classification and Clustering of Broadcast News Audio” In *ARPA Speech Recognition Workshop* pp. 97–99.

K. Spärck Jones, G. J. F. Jones, J. T. Foote and S. J. Young (1996). “Experiments in spoken document retrieval” In *Information Processing and Management* 32(4):pp. 399-417.

K. Tanaka, et al. (2001). “Speech data retrieval system constructed on a universal phonetic code domain” In *Proc. of ASRU2001* pp. 1–4.

S. Lee, et al. (2002). “Evaluation of speech data retrieval system using sub-phonetic sequence” In *Proc. of Autumn Meeting of the Acoustical Society of Japan* pp. 159–160.

Y. Itoh and K. Tanaka (2001). “Automatic Labeling and Digesting for Lecture Speech Utilizing Repeated Speech by Shift CDP” In *Proc. of EUROSPEECH-2001* pp. 1805-1808.

K. Fukunaga (1990). “Introduction to Statistical Pattern Recognition” Academic Press

T. Kawahara, et al. (1998). “Sharable software repository for Japanese large vocabulary continuous speech recognition” In *Proc. of ICSLP’98* pp. 3527–3260.

The CMU Pronouncing Dictionary (v. 0.6), In *http://www.speech.cs.cmu.edu/cgi-bin/cmudict*.