Speech Retrieval of Mandarin Broadcast News via Mobile Devices

Berlin Chen, Yi-Ting Chen, Chih-Hao Chang, Hung-Bin Chen

Graduate Institute of Computer Science & Information Engineering,
National Taiwan Normal University, Taipei, Taiwan, Republic of China
berlin@csie.ntnu.edu.tw

Abstract

This paper presents a system for speech retrieval of Mandarin broadcast news. First, several data-driven and unsupervised approaches are integrated into the broadcast news transcription system to improve the speech recognition accuracy and efficiency. Then, a multi-scale indexing paradigm for broadcast news retrieval is proposed to make use of the special structural properties of the Chinese language as well as to alleviate the problems caused by the speech recognition errors. Finally, we use the PDA as the platform and Mandarin broadcast news stories collected in Taiwan as the document collection to establish a speech-based multimedia information retrieval prototype system. Very encouraging results are obtained.

1. Introduction

Speech is the primary and the most convenient means of communication between people [1]. Due to the successful development of much smaller electronic devices and the popularity of wireless communication and networking, it is widely believed that speech will play a more active role and will serve as the major human-machine interface for the interaction between people and different kinds of smart devices in the near future. On the other hand, huge quantities of multimedia information, such as broadcast radio and television programs, voice mails, digital archives and so on, are continuously growing and filling our computers, networks and lives. It is obvious that speech is one of the most important sources of information for the great volumes of multimedia, and therefore research on multimedia content understanding and organization using speech is now becoming more and more emphasized. For example, substantial efforts and very encouraging results on broadcast news transcription, retrieval and summarization have been reported in the last few years [2]-[4].

However, in order to obtain better recognition performance, most of the transcription systems required not only large amounts of manually transcribed speech materials for acoustic training in the data preparation phase, but also much time expenditure and memory overhead in the recognition phase. On the other hand, because the speech recognition transcripts often contain errors and lack for organization and structure, it is still not very easy to retrieve and browse the desired contents from the retrieved multimedia documents. With these observations in mind, in this paper we present a system for speech retrieval of Mandarin broadcast news. Unlike [5], that focused on the delivery of multimedia information, and [3], that used the spoken query to retrieve the Mandarin newswire texts, we focus here on automatic transcription, retrieval and browsing of speech information, and both the query and the documents to be retrieved are in spoken form. In order to improve the speech recognition accuracy and efficiency, several data-driven and unsupervised approaches are integrated into the broadcast new transcription system. Moreover, a multi-scale indexing paradigm is proposed to make use of the special structural properties of the Chinese language as well as to alleviate the problems caused by the speech recognition errors. All the above approaches have been successfully integrated into our speech recognition and retrieval systems, while a prototype system for speech access to Mandarin broadcast news via the PDA has also been established.

2. Speech Recognition System

The major constituent parts of the broadcast news system as well as the speech and language data used in this paper will be described in this section.

2.1 Front-End Processing

The front-end processing is conducted with two data-driven feature extraction approaches: the LDA-based (Linear Discriminant Analysis) and the HLDA-based (Heteroscedastic Linear Discriminant Analysis) approaches, while the HLDA-based approach further seeks to remove the equal class variance constraint assumed by the LDA-based approach. The states of each HMM were taken as the unit for class assignment. Both LDA and HLDA analyses are performed by using the outputs of Mel-filter banks obtained from each frame and its neighboring frames. The dimension of the resultant vectors was set to 39. More specially, for the HLDA-based approach, the minimum classification error (MCE) criterion, instead of the maximum likelihood (ML) criterion, was exploited to find the optimal MCE-HLDA transformation matrix.

2.2 Broadcast News Corpus and Acoustic Training

The speech data set consists of about 176 hours of radio/TV broadcast news, which were collected from several radio and TV stations located at Taipei during November 1998 to October 2004. All the speech materials were manually segmented into separate stories, and each of them is a news abstract pronounced by one anchor speaker. Only 7.7 hours of speech data is equipped with corresponding orthographic transcripts, in which about 4.0 hours of data collected during 1998 to 1999 is used to bootstrap the acoustic training and the other 3.7 hours of data (506 stories) collected in September 2002 is for testing. An amount of 104.3 hours of the rest untranscribed speech data is reserved for unsupervised acoustic training.

Unlike the previous approaches [6]-[7] which aligned the closed-captions with the automatic transcripts and kept only portions that agreed for acoustic training, in this paper, we

developed a verification-based method for automatic acoustic training data acquisition by making use of the confidence measure, consisting of word-level posterior probability as well as subword-level acoustic verification score, to locate the most probably correct words [8]. Gender-independent INITIAL-FINAL models were used.

2.3 Lexicon and N-gram Language Modeling

The recognition lexicon initially consists of 67K words. A set of about 5K compound words was automatically derived using the forward and backward bigram statistics and was then added to the lexicon to form a new lexicon of 72K words. The background language models used in this paper consist of trigram and bigram models, which were estimated using a text corpus consisting of 170 million Chinese characters collected from Central News Agency (CNA) in 2000 and 2001 (the Chinese Gigaword Corpus released by LDC). The *n*-gram language models were trained with Katz backoff smoothing.

2.4 Speech Recognition

The speech recognizer was implemented with a left-to-right frame-synchronous Viterbi tree search as well as a lexical prefix tree organization of the lexicon. At each speech frame, a beam pruning technique, which considered the decoding scores of path hypotheses together with their corresponding unigram language model look-ahead and syllable-level acoustic look-ahead scores, was used to select the most promising path hypotheses. Moreover, if the word hypotheses ending at each speech frame had scores higher than a predefined threshold, their associated decoding information, such as the word start and end frames, the identities of current and predecessor words, and the acoustic score, will be kept in order to build a word graph for further language model rescoring. In this study, the word bigram language model was used in the tree search procedure while the trigram language model was used in the word graph rescoring procedure. Our transcription system is a lightweight system, which can be run in real time on an ordinary Pentium IV PC.

3. Information Retrieval System

The information retrieval system is implemented in a client-server architecture, in which the broadcast news indexing and retrieval are performed at the server side and the query is posed at the client side. The considerations of the structural properties of the Chinese language, the indexing mechanism and the information retrieval model used in this paper are explained as follows.

3.1 Considerations of the Structural Properties of the Chinese Language

In Mandarin Chinese, there is an unknown number of words, though only some (e.g., 80 thousands, depending on the domains) are commonly used. Each word is composed of one or more characters, and each character is pronounced as a monosyllable and is a morpheme with its own meaning. As a result, new words are easily generated by combining a few characters. For example, the combination of the characters "電 (electricity)" and "腦 (brain)" yields the word "電腦 (computer)" while the combination of "火 (fire)" and "山 (mountain)" yields the word "火山 (volcano)". Mandarin Chinese is phonologically compact; an inventory of about 400 base syllables provides full phonological coverage of Mandarin audio, if the differences in tones are disregarded. There is a many-to-many mapping between characters and syllables. Consequently, a foreign word can be translated into different Chinese words based on its pronunciation. For example, Kosovo may be translated into "科索沃/ke1-suo3-wo4/", "科索佛/ke1-suo3-fo2/", "柯索佛/ke1-suo3-fo2/", etc. Different translations usually have some syllables in common, or may have exactly the same syllables. The characteristics of the Chinese language lead to some special considerations when performing Mandarin Chinese speech recognition, e.g., syllable recognition is believed to be a key problem. Recognition performance evaluation is usually based on syllable accuracy and character accuracy, rather than word accuracy.

The characteristics of the Chinese language also lead to some special considerations for the spoken document retrieval task. Word-level indexing features possess more semantic information than subword-level features; thus, word-based retrieval enhances the precision. On the other hand, subword-level indexing features are more robust against the Chinese word tokenization ambiguity, open vocabulary problem, and speech recognition errors; thus, subword-based retrieval enhances the recall. Accordingly, there is good reason to fuse the information obtained from indexing features of different levels.

3.2 Indexing Mechanism

A retrieval index was generated over the entire collection of indexed broadcast news documents. Both the word-based and syllable(subword)-based indexing approaches were used here to represent the broadcast news documents. Every recognized word sequence was also automatically converted into its equivalent syllable-level sequence. For the word-based indexing approach, single words are taken as the index terms, while for the syllable-based approach, both the single syllables and overlapping syllable pairs are the index terms.

3.3 Information Retrieval Model

The vector space model (VSM) widely used in many text information retrieval systems was used here for simplicity, though our previous experiments on Mandarin spoken document retrieval have demonstrated the HMM-based retrieval models have superior retrieval performance over VSM [9]. In VSM, a document D can be represented by a set of feature vectors $\vec{d}_j$, each consisting of information for one type of indexing terms, such as a single word, a single syllable or a syllable pair. Each component x_{jt} of the feature vector $\vec{d}_j$ for a document D is associated with the weighted statistics of a specific indexing term t:

$$x_{jt} = \left[1 + \ln\left(\sum_{i=1}^{n_t} c_t(i)\right)\right] \cdot \ln\left(N / N_t\right), \quad (1)$$

where $c_t(i)$, ranging from 0 to 1, is the confidence measure evaluated for the *i*-th occurrence of the specific indexing term t within the document D, and n_t is the total frequency counts for the occurrences of the specific indexing term t in the document. Therefore the value of $\left[1 + \ln\left(\sum_{i=1}^{n_t} c_t(i)\right)\right]$ denotes the term frequency of the specific indexing term t but evaluated in terms of the confidence measure, and the logarithmic operation is to compress its distribution.

$\ln(N/N_t)$ is the Inverse Document Frequency (IDF), where N_t is the number of documents that include the term t and N is the total number of documents in the collection. A query Q is also represented by a set of feature vectors $\vec{q}_j$ constructed in the same way. The Cosine measure is used to estimate the query-document relevance for each type of indexing terms:

$$R_j(\vec{q}_j, \vec{d}_j) = \left(\vec{q}_j \bullet \vec{d}_j\right) \Big/ \left(\left\|\vec{q}_j\right\| \cdot \left\|\vec{d}_j\right\|\right). \quad (2)$$

The overall relevance is then the weighted sum of the relevance scores of all types of indexing terms:

$$R(Q, D) = \sum_j w_j \cdot R_j(\vec{q}_j, \vec{d}_j), \quad (3)$$

where w_j are empirically tunable weights.

4. Experimental Results

4.1 Broadcast News Transcription

In this subsection, a series of experiments was conducted to assess recognition performance as a function of the improved approaches presented in this paper. As the results shown in the second row of Table 1, the baseline system (using the LDA features) initially achieves a character error rate of 20.89% and a syllable error rate of 14.97%. Because the word error rate is not a good performance measure for the Chinese language, we just list the corresponding results here for reference. Rows 3 to 5 of Table 1 are respectively the recognition results obtained when the MCE-HLDA feature extraction, automatic compound word extraction and unsupervised acoustic training are further integrated ("+") into the broadcast news system. As can be seen, the character error rate can be significantly reduced to 14.85% while the syllable error rate to 9.39%. The last row of Table 1 shows the results when online unsupervised MLLR adaptation was included. Unlike the conventional approach using the top 1 recognized word sequence for adaptation (whose results are shown in the parentheses), the adaptation approach presented in this paper performed on the word graph (WG) instead and exploited the word-level posterior probability as well as subword-level acoustic verification score to weight the accumulated statistics. It can be found that such an approach could provide slightly superior performance over the conventional one and the system finally yielded a character error rate of 14.29% and a syllable error rate of 8.91%.

4.2 Broadcast News Retrieval

There were totally about 21,000 broadcast news documents used here to evaluate the information retrieval performance. Both the word-based and syllable-based indexing approaches described previously in Section 3 were evaluated. On the other hand, a set of 20 simple queries, in both spoken and written forms, and their corresponding relevant news documents were manually created to support the retrieval experiments. Four speakers (two males and two females) were instructed to produce the 20 queries, respectively, over an Acer n20 PDA using the original microphone and in a recording environment with slight background noise. To recognize the spoken queries, another read speech database consisting of about 8.5 hours of speech produced by other 39 male and 38 female speakers over the same type of PDAs was used for training the speaker-independent HMMs for automatic recognition of the spoken queries. The recognition results are presented in Table 2. It can be found that the character (CER) and syllable (SER) error rates for the spoken queries are 27.61% and 19.47%, respectively. The results are not as good as that of broadcast news transcription reported earlier, it is probably because that most of the test queries contain one to several out-of-vocabulary (OOV) words, such as personal names and new organization or event names, which apparently occur much more frequently in the queries than in the broadcast news documents and may degrade the speech recognition performance severely.

The final retrieval results are evaluated in terms of the mean average precision (*m*AP) at different document cutoff values k. The retrieval results are shown in Table 3. Columns 3, 4 and 5 respectively show the results using the word-level indexing features (W), syllable-level indexing features (S) and both of them (S+W), which are evaluated at different document cutoff values and with either text queries (TQ) or spoken queries (SQ). As can be seen, the syllable-level indexing features are better than the word-level features for either text queries or spoken queries, while using both of them gives significant improvements than using any of them alone. Moreover, the retrieval results for the spoken queries are much worse than that of the text queries, but the combination of word-level and syllable-level features do help to reduce the performance gap between the spoken and the text queries.

5. Prototype System

5.1 System Description

We implemented a prototype system that allows the user to search for Mandarin broadcast news via the PDA using a spoken natural language query. The framework of the system is shown in Figure 1. There is a small client program on the PDA, as illustrated in Figure 2, which transmits the speech waveform or acoustic feature data of the spoken query to the information retrieval server. The information retrieval server then passes the speech waveform or acoustic feature data to the large vocabulary continuous speech recognition (LVCSR) server, which works in the similar way as the broadcast news transcription system shown earlier in Section 2. The recognition result is then passed back to the information retrieval server to act as the query to generate a ranked list of relevant documents. When the retrieval results are sent back to the PDA, the user can first browse the summaries of the retrieved documents and then click to read the speech transcripts of the relevant broadcast news documents or play the corresponding audio files from the audio streaming server. The summaries were automatically generated with a dynamic programming procedure by using the word-level linguistic score, significance score, as well as confidence measure [4]. On the other hand, the huge collection of broadcast news documents, as described previously in Section 4, is offline recognized by the broadcast news transcription system and the resultant transcripts are then utilized by the multi-scale indexer to generate the word-level and syllable-level indexing terms. The final retrieval indices, including the vocabularies and document occurrences of indexing terms of different types (word- and syllable-level indexing terms), are stored as inverted files for efficient searching and comparison.

	WER (%)	CER (%)	SER (%)
Baseline	28.26	20.89	14.97
+ MCE-HLDA	27.42	20.15	14.23
+Compound Words	27.84	19.00	13.34
+Unsupervised Acoustic Training	23.06	14.85	9.39
+MLLR (WG)	22.29 (22.36)	14.29 (14.32)	8.91 (8.97)

Table 1: The recognition results for the testing broadcast news stories, expressed in terms of word error rate (WER), character error rate (CER) and syllable error rate (SER).

WER (%)	CER (%)	SER (%)
49.16	27.61	19.47

Table 2: The recognition results for the set of 20 queries spoken by 2 male speakers and 2 female speakers, expressed in terms of word error rate (WER), character error rate (CER) and syllable error rate.

		W	S	W + S
Document Cutoff 10	TQ	0.7130	0.7588	0.8038
	SQ	0.5799	0.5901	0.6237
Document Cutoff 30	TQ	0.6006	0.6416	0.6692
	SQ	0.4827	0.4986	0.5232
Document Cutoff 50	TQ	0.5277	0.5610	0.5840
	SQ	0.4217	0.4388	0.4586

Table 3: The retrieval results achieved by using word-level features (W), syllable-level features (S) and both of them (W+S), evaluated in terms of the mean average precision at different document cutoff values.

5.2 PDA Programming and Multimedia Streaming

The PDA client system was programmed to run on the WinCE 4.0 Pocket PC2003 Operating System, and the development environment is the Embedded Visual C++ 4.0, which is a freeware released by the Microsoft Corporation. The client system is connected with the server side via the wireless network and following the IEEE 802.11b/g protocol. On the other hand, a streaming server runs on the Windows Server 2000 Operating System and resides at the server side, while the client system programmed with the Active Template Library (ATL) hosts the Windows CE Multimedia Player to play the streaming files. The Media Encoder, which is also a freeware released by the Microsoft Corporation, was used beforehand to process the huge collections of broadcast news recordings and to encode them into the WMA 9.0 format.

6. Conclusions

This paper presents the initial results of a research project toward automatic recognition, retrieval and organization of speech information. Several data-driven and unsupervised approaches to Mandarin broadcast news speech recognition and retrieval were integrated into the prototype system, which allows the user to search for Mandarin broadcast news via the PDA using a spoken natural language query. Very encouraging experimental results were demonstrated.

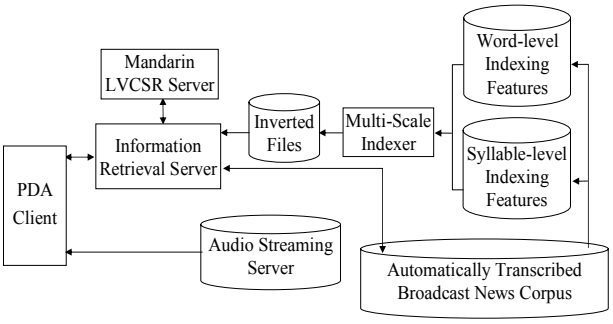

Figure 1: The framework for speech retrieval to Mandarin broadcast news.

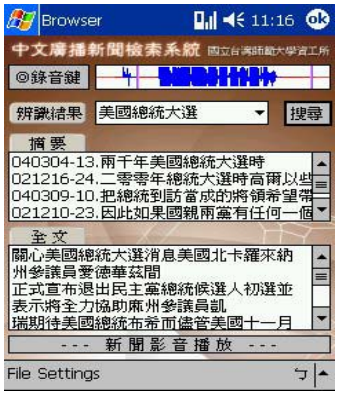

Figure2: The PDA client system.

7. References

[1] B.H. Juang and S. Furui, "Automatic Recognition and Understanding of Spoken Language—A First Step Toward Natural Human–Machine Communication," *Proceedings of the IEEE*, vol. 88, NO. 8, August 2000.

[2] B. Chen et al., "Discriminating Capabilities of Syllable-Based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information in Mandarin Chinese", *IEEE Trans. on Speech and Audio Processing*, Vol. 10, No. 5, 2002.

[3] E. Chang et al., "A System for Spoken Query Information Retrieval on Mobile Devices," *IEEE Trans. on Speech and Audio Processing*, Vol. 10, No. 8, 2002

[4] S. Furui et al., "Speech-to-Text and Speech-to-Speech Summarization of Spontaneous Speech," *IEEE Trans. on Speech and Audio Processing*, Vol. 12, No. 4, July 2004.

[5] D. Gibbon and L. Begeja, "Multimedia Processing for Enhanced Information Delivery on Mobile Devices," in *Proc. MobEA 2004*.

[6] L. Lamel, et al., "Lightly Supervised and Unsupervised Acoustic Model Training," *Computer Speech and Language*, Vol. 16, 2002.

[7] L. Nguyen and B. Xiang, "Light Supervision in Acoustic Model Training," in *Proc. ICASSP 2004*.

[8] B. Chen et al., "Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription," in *Proc. ICASSP 2004*.

[9] B. Chen et al., "A Discriminative HMM/N-Gram-Based Retrieval Approach for Mandarin Spoken Documents," *ACM Trans. on Asian Language Information Processing*, Vol. 3, No. 2, 2004.