Metadata for Integrating Chinese Text and Speech Documents in a Multimedia Retrieval System

Yue-Shi Lee and Hsin-Hsi Chen

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan, R.O.C.
Email: {leeys, hh_chen}@csie.ntu.edu.tw

Abstract

Multimedia documents place new requirements on the conventional text retrieval systems. This paper presents a multimedia retrieval system that employs the content-based strategy to retrieve both text and speech documents. Its input can be a sequence of spoken words which are digitized waveforms or a sequence of characters, and its output is a list of ranked text and/or speech documents. In this system, a new metadata especially designed for both text and speech documents is proposed. The metadata is automatically generated with special consideration of the characteristics of Chinese. The presented approach is very easy to implement and the preliminary tests give very encouraging results.

1 Introduction

As concepts and technologies for single media types are maturing, problems concerning the composition of multimedia documents from single-media components are becoming more acute. However, retrieving complex multimedia objects from multimedia database is more difficult than text retrieval. This paper concentrates the speech retrieval since it may play an important role in a multimedia retrieval system.

For audio and speech documents retrieval, several investigations were proposed. Wilcox and Bush (1991) developed a wordspotting system for voice indexing. Rose, et al. (1991) presented a speech message information retrieval system that classifies speech messages according to predefined topic or message classes. The system accepts speech messages as input and produces an estimated message class as output. Glavitsch, et al. (1994) proposed an information retrieval system that can search for text and speech documents. In their system, three special types of subword units are selected as indexing features: VCV-, CV- and VC-features. V stands for a maximum sequence of vowels and C a maximum sequence of consonants within the same word. The above approaches, however, are specially designed for English. Obviously, they are not appropriate for the non-alphabetic languages like Chinese.

This paper will present a multimedia retrieval system that employs the content-based strategy to retrieve both text and speech documents. A new metadata is proposed with special consideration of the characteristics of Chinese.

This paper is organized as follows. Section 2 gives an overview of our retrieval system. Section 3 presents a set of indexing features for both text and speech documents. Section 4 shows how to identify these indexing features. Section 5 introduces the ways to determine the weight of each indexing feature and estimate the relevant degrees between queries and documents. Finally, Section 6 evaluates our retrieval system.

2 System Description

The proposed system is depicted in Figure 1. The processes of metadata generation for query and multimedia database are shown in Figures. 2 and 3.

Our retrieval system accepts vague (quasi-natural language) queries and adopts the best-match search to find those documents that are relevant to the query. The best-match retrieval paradigm is important for multimedia databases because current speech and image recognition systems are not able to recognize words and images perfectly. The input of the system can be a sequence of spoken words which are digitized waveforms or a sequence of characters. The output of the system is a list of ranked text and/or speech documents that meet the user's requirement. When a user submits a query via keyboard or microphone, the Syllable Recognition Module recognizes the speech input and the Text-to-Syllable Module transforms the text input into the syllabic form, respectively. Because our system allows to search for text and speech documents simultaneously, the phonetic units are selected as the indexing features.

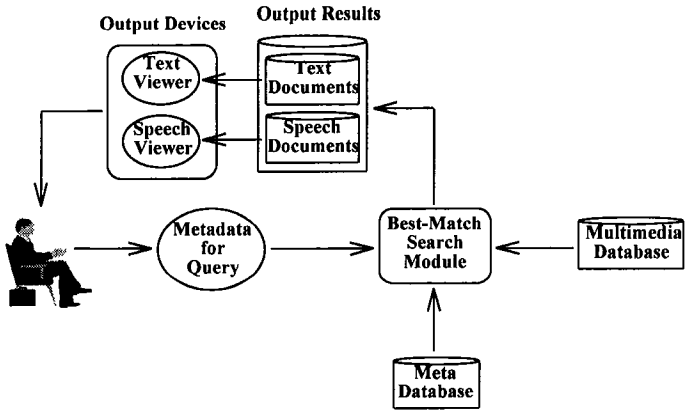

Figure 1 Overview of the Retrieval System

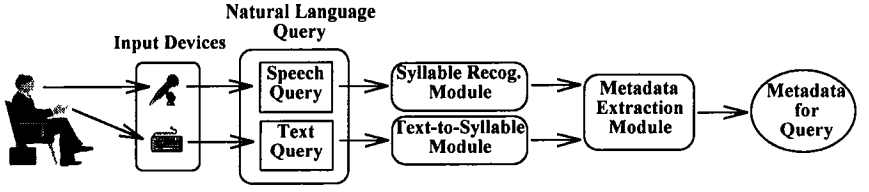

Figure 2 The Processes of Metadata Generation for Query

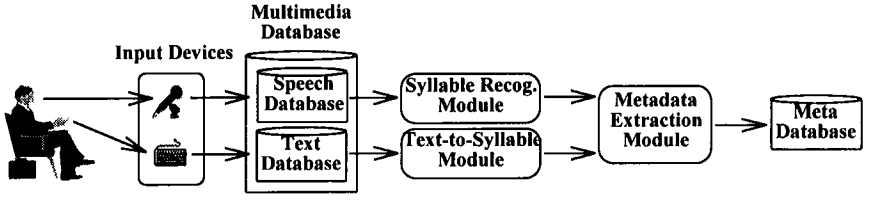

Figure 3 The Processes of Metadata Generation for Multimedia Database

The Metadata Extraction Module then generates the metadata according to these syllable inputs. Similarly, the meta database is generated from multimedia database beforehand by the same processes. The metadata for query, meta database and multimedia database are sent to the Best-Match Search Module for retrieving the relevant documents. Through the text and the speech viewers, the user can browse the retrieved documents.

3 Indexing Features for Both Text and Speech Documents

There are several possible candidates for indexing features, e.g., consonants, vowels, phonemes, syllables, characters, words and phrases. To select an appropriate indexing vocabulary for retrieving both text and speech documents, the following two issues must be considered (Glavitsch, et al. 1994):

1. For the retrieval efficiency, the indexing vocabulary may not be too large.
2. For speech-document retrieval, the indexing features must be phonetic units that are easily recognized by speech recognition systems.

Words or phrases are too large to serve as indexing features. In general, the vocabulary in Chinese has more than 100,000 features. In contrast to words, there are only about 10,000 characters in Chinese. Thus, character-based approaches are wildly used in Chinese text retrieval systems (Chien 1995). However, some characters or words frequently used in query are difficult to be recognized by speech recognition systems. Proper nouns like personal names, organization names and transliterated names are typical examples. From Chinese phonology, the basic phonetic units including 4 tones, 22 consonants and 38 vowels are used for the construction of

1,300 Mandarin syllables. And there are 414 phonologically allowed tone-independent syllables (phonemes). Obviously, consonants and vowels are too small to serve as indexing features. Lin, *et al.* (1995) presented a syllable-based voice retrieval system for retrieving Chinese textual databases. In their system, bi-syllables (i.e., syllable-pairs) are selected as indexing features. Total 1,690,000 indexing features are required. Obviously, such an indexing vocabulary violates our first issue. For the retrieval efficiency and the consideration of speech-document retrieval, the desired indexing features seem to lie between vowels and characters.

Table 1 summarizes the number of the indexing features at different linguistic levels. In this paper, syllables (1,300), phonemes (414 features), bi-phonemes (171,396) features) and a hybrid indexing features are considered as the indexing features in our experiments. The hybrid indexing vocabulary consists of three parts (total 2,342 features), i.e., phonemes (414 features), bi-consonants (484 features) and bi-vowels (1,444 features).

Table 2 shows an example and the corresponding phonetic units. Note that the symbol '@' denotes 'none'. Table 3 demonstrates the different indexing schemes from the example.

Table 1 Total Indexing Features at Different Linguistic Levels

Indexing Feature	# of Indexing Feature
Word	100,000
Bi-Word	10,000,000,000
Character	13,000
Bi-Character	169,000,000
Syllable	1,300
Bi-Syllable	1,690,000
Phoneme	414
Bi-Phoneme	171,396
Vowel	38
Bi-Vowel	1,444
Consonant	22
Bi-Consonant	484

Table 2 An Simple Example

English	I	am	a		Chinese		
Chinese	我	是	一	個	中	國	人
Syllables	wo3	shi4	yi2	ge5	zhong1	guo2	ren2
Phonemes	wo	shi	yi	ge	zhong	guo	ren
Consonants	@	shi	@	g	zhi	g	ri
Vowels	wo	@	yi	ê	weng	wo	en

Table 3 Indexing Feature Extraction

Indexing Features	Extracted Indexing Features						
Syllable	wo3	shi4	yi2	ge5	zhong1	guo2	ren2
Bi-Phoneme	wo - shi	shi - yi	yi - ge	ge - zhong	zhong - guo	guo - ren	
Phoneme	wo	shi	yi	ge	zhong	guo	ren
Hybrid	wo	shi	yi	ge	zhong	guo	ren
	@ - shi	shi - @	@ - g	g - zhi	zhi - g	g - ri	
	wo - @	@ - yi	yi - ê	ê - weng	weng - wo	wo - en	

4 Syllable Recognition and Text-to-Syllable Modules

Last section proposes phonetic units and their combinations as indexing features, so that a Syllable Recognition Module and a Text-to-Syllable Module are indispensable in our multimedia retrieval system. The details are described in the following.

4.1 Syllable Recognition Module

Since Mandarin speech is a mono-syllabic and tonal language, it is suitable for hierarchical recognition (Wang, et al. 1991). The tone, the vowel and the consonant of each syllable are separately adopted and hierarchically recognized in our system. For 1,300 Mandarin syllables, a hierarchical neural network employs pitch vector at the first level, vowel vector at the second level, and the consonant vector at the third level to recognize tone, vowel and consonant, respectively. The system produces a set of syllable candidates when a character is pronounced by a user. Table 4 lists the experimental results for our Syllable Recognition Module. The performances of top n, i.e., the correct results appear within the top n candidates, are shown.

Table 4 Experimental Results for Syllable Recognition Module

	Top 1	Top 2	Top 3	Top 9
Syllable	55.72%	68.07%	72.88%	76.75%
Phoneme	71.87%	83.07%	89.05%	94.86%
Consonant	74.56%	91.14%	95.76%	99.36%
Vowel	87.03%	96.32%	98.47%	99.93%

Because phoneme are used in our final model, we further use a Markov model to identify the correct phoneme from the top n candidates. The model is defined as follows. Given a sequence of m phoneme-sets and these sets form a phoneme lattice. The model attempts to find a path ρ from the lattice such that the phoneme sequence $<p_{\rho(1)}, p_{\rho(2)}, \ldots, p_{\rho(m)}>$ denotes the correct result. Here $p_{\rho(i)}$ is one of the phoneme candidates in i-th phoneme set. The path ρ is defined by the following formula.

$$\rho \cong \underset{\rho}{\operatorname{argmax}} \ \mathrm{Prob}(p_{\rho(1)}) * \prod_{i=1}^{m-1} \mathrm{Prob}(p_{\rho(i+1)} | p_{\rho(i)})$$

$\mathrm{Prob}(p_{\rho(i)})$ denotes the probability of phoneme $p_{\rho(i)}$ and $\mathrm{Prob}(p_{\rho(i+1)}|p_{\rho(i)})$ denotes the conditional probability of phoneme $p_{\rho(i+1)}$ given phoneme $p_{\rho(i)}$. Table 5 lists the experimental results of phoneme recognition.

Table 5 Experimental Results for Phoneme Recognition

	1	2	3	9
Accuracy	71.87%	74.96%	72.54%	52.71%

It shows that the performance increased by 3.09% when each phoneme set has 2 candidates. But, if the phoneme is identified from the top 9 candidates, the performance becomes 52.71%. The Markov process in this work obtains a little improvements.

4.2 Text-to-Syllable Module

As mentioned above, Mandarin speech is a mono-syllabic language. Each character is pronounced by a syllable. However, more than 1,000 characters, so-called Poyinzi (homographs), have multiple pronunciations in Chinese. Hence, the Chinese text-to-syllable or the character-to-syllable conversion is also not trivial.

In general, the problem of Chinese homograph disambiguation can be defined as: how to convert a sequence of characters $C=c_1, c_2, \ldots, c_n$ into a sequence of syllables $S=s_1, s_2, \ldots, s_n$. Here, each s_i represents one of the 1,300 Chinese syllables and each c_i represents one of the 13,000 Chinese characters. Because Chinese words have far fewer homographs than Chinese characters, our first work is to separate a sequence of Chinese characters into a sequence of Chinese words (Lee, et al. 1994; Chen & Lee 1994). Most of the ambiguities in homographs are eliminated in word strings. The remaining ambiguities are resolved by heuristic rules. Such a design achieves 99.69% correct rate (Chen & Lee 1995). It has much higher performance than the Phoneme Recognition Module described in the last section.

5 Retrieval Model

This section describes how the relevant documents are retrieved on the basis of the metadata. Assume the number of indexing features is m. The metadata for each document (i.e., document description vector) and query (i.e., query description vector) are defined as m-dimension vectors shown below:

$$D_i = (d_{i,1}, d_{i,2}, \ldots, d_{i,m-1}, d_{i,m})$$

and

$Q_j = (q_{j,1}, q_{j,2}, \ldots, q_{j,m-1}, q_{j,m})$

The coefficients $d_{i,j}$ and $q_{i,j}$ denote the weights of indexing feature ϕ_j in document d_i and query q_i, respectively. The relevance of a document with respect to the query is estimated in terms of these two vectors:

$$Sim(d_i, q_j) = \frac{D_i^T \cdot Q_j}{\sqrt{D_i^T \cdot D_i} * \sqrt{Q_j^T \cdot Q_j}}$$

This method is the well-known cosine measure widely used in information retrieval models (Salton 1989). The retrieval system presents a list of ranked documents to the user in increasing order of the angle.

Besides, the weight of an indexing feature in the description vector should be high if the feature characterizes the document or query well. A weighting scheme that meets this requirement is shown as follows:

$d_{i,j} = ff(d_i, \phi_j) * idf(\phi_j)$

The feature frequency, $ff(d_i, \phi_j)$, denotes the number of occurrences of the indexing feature ϕ_j in the document d_i. The inverse document frequency $idf(\phi_j)$ of an indexing feature ϕ_j defined below is a measure of the specificity of a feature.

$$idf(\phi_j) = \log\left(\frac{n+1}{df(\phi_j)+1}\right)$$

The symbol n denotes the total number of documents in the collections. The document frequency $df(\phi_j)$ is the number of documents containing ϕ_j.

6 System Evaluation

For testing the performance of the proposed approaches, an experimental environment is set up. The text database contains 5,000 news articles and the speech database contains the speech version of these 5,000 news articles. Total 100 queries are issued by two users. Figure 4 shows four combinations of the queries and the databases. Intuitively, the combination of the text query and the text database contains less noise. On the contrary, the combination of the speech query and the speech database has much noise. The experimental results are listed in Table 6.

Table 6 The Performance of the Retrieval System

Query Type	Document Type	Indexing Feature		
		Phoneme	Bi-Phoneme	Hybrid
Speech	Text	68%	90%	90%
*Speech	Text	88%	96%	96%
Text	Speech	68%	86%	84%
Speech	Speech	88%	96%	96%

In this paper, the hit rates of the top 1 document, i.e., the average precision rates of the top 1 document, are shown.

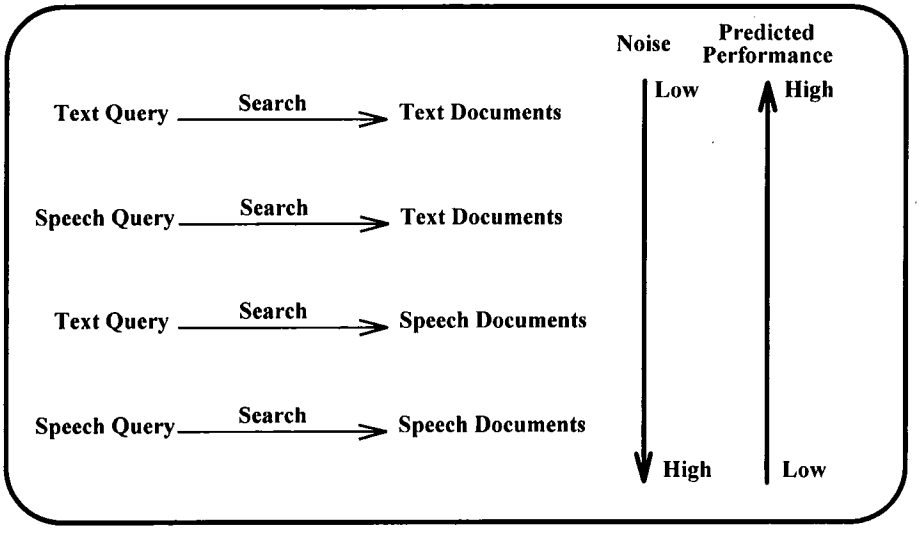

Figure 4 Four Relationships between the Queries and the Databases

The first row demonstrates the results of speech access to text documents. The second row shows the performance of the system when the speech query is correctly recognized by the Syllable Recognition Module. The performance increases in all three indexing schemes. The text access to speech documents has the worse performance shown in the third row. An interesting result is the performance of speech query to speech documents. It is equal to the performance in the second row. Intuitively, this combination is expected to have worse performance than other combinations. The reason is the same syllable recognition module is adopted to process speech query and speech database. The noises that appear in the treatment of the speech queries also exist in the treatment of the speech documents. In this way, the influence of noise may be decreased. In some cases, however, the noise cannot be canceled out, since the background noise may be different at document creation and query time. The experimental results also show both the bi-phonemes and the hybrid indexing features are good schemes. However, the size of the hybrid indexing features (2,342 features) is much smaller than that of bi-phonemes (174,724 features). Thus the former is a better design than the latter from the issue of the retrieval efficiency.

7 Concluding Remarks

With the advent of large digital libraries containing multimedia information, a modern information retrieval system should have capability to manipulate and integrate the heterogeneous multimedia documents. This paper deals with a speech retrieval model in a multimedia retrieval system. Our model is a combination of a conventional text retrieval model (vector space model) and a speech recognition model (hierarchical neural network model). A set of indexing features is also presented. These features are employed not only in speech documents but also in text documents. That is, they can be used to retrieve documents of both types simultaneously. Similarly, the query can be either a speech query or a text query. Hence, our approach is a step towards multimedia retrieval systems where the retrieval model deals coherently with the different media. Although the preliminary tests give very encouraging results, some important issues especially for the ability to handle spontaneous speech must be concerned. Bu it must be investigated further.

References

Chen, H.H., and Lee, J.C. 1994. The Identification of Organization Names in Chinese Texts. *Communication of Chinese and Oriental Languages Information Processing Society* 4(2):131-142.

Chen, H.H., and Lee, Y.S. 1995. An Adaptive Learning Algorithm for Task Adaptation in Chinese Homophone Disambiguation. *Computer Processing of Chinese and Oriental Languages* 9(1):49-58.

Chien, L.F. 1995. Fast and Quasi-Natural Language Search for Gigabytes of Chinese Texts. In Proceedings of ACM SIGIR, 112-120.

Glavitsch, U., et al. 1994. Metadata for Integrating Speech Documents in a Text Retrieval System. *SIGMOD RECORD* 23(4):57-63.

Lee, J.C., et al. 1994. Identification of Personal Names in Chinese Texts. In Proceedings of ROCLING, 203-222.

Lin, S.C., et al. 1995. Unconstrained Speech Retrieval for Chinese Document Databases with Very Large Vocabulary and Unlimited Domains. In Proceedings of EUROSPEECH, 1203-1206.

Rose, R.C., et al. 1991. Techniques for Information Retrieval from Voice Messages. In Proceedings of ICASSP, 317-320.

Salton, G. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley.

Wang, J.F., et al. 1991. A Hierarchical Neural Network Model Based on a C/V Segmentation Algorithm for Isolated Mandarin Speech Recognition. *IEEE Transaction on Signal Processing* 39(9):2141-2146.

Wilcox, L.D. and Bush, M.A. 1991. HMM-Based Wordspotting for Voice Editing and Indexing. In Proceedings of EUROSPEECH, 25-28.