Expanded Vector Space Model based on Word Space in Cross Media Retrieval of News Speech Data

S.Takao, J.Ogata and Y.Ariki

Department of Electronics and Informatics
Ryukoku University, Seta, Otsu-shi, Shiga, 520-2194, Japan
tail@arikilab.elec.ryukoku.ac.jp

ABSTRACT

News On Demand System using speech technology usually employs automatic speech transcriptions to retrieve the news data. In the retrieval, users specify a few keywords or sentences as a query and the related news data can be retrieved using the speech transcription. However when users can’t give a query clearly, a video shot of news program which users are watching will become a good query to retrieve the related news data. As one of such kinds of news data retrieval, we propose here to employ video captions as a query and to retrieve the related news data using speech transcription. We call this kind of retrieval as cross media retrieval due to its media cross over. Conventionally available method in cross media retrieval is standard cosine measure in vector space model. In this conventional method, there is a problem of impossibility of semantic level retrieval. To solve this problem, we propose here an expanded vector space model based on a word space. Experimental results found that the expanded vector space model based on the word space has superiority to the conventional vector space model.

1. INTRODUCTION

Recently, TV news programs are broadcast from all over the world owing to the broadcast digitization. In this situation, TV viewers want to select and watch the most interesting news. In order to satisfy this requirement, many studies [1]-[8] have been done to construct news database with automatic topic segmentation and retrieval function for broadcast news speech.

Broadcast TV news consists of speech, video caption and video images. The speech mainly conveys the news contents and the video caption summarizes them. The video images send the evidence or the atmosphere to the broadcast digitization. In this situation, TV viewers want to select and watch the most interesting news. In order to satisfy this requirement, many studies [1]-[8] have been done to construct news database with automatic topic segmentation and retrieval function for broadcast news speech.

Fig.1 shows an outline of a cross media retrieval system. All the spoken documents in news speech database are already transcribed. Users can retrieve related spoken documents by giving the video caption as a query, when he can’t give the query in language clearly.

In cross media retrieval, there are two problems. The first is error words caused by speech recognition. The second is error words caused by video caption recognition. These error words decrease the retrieval performance. In particular, if important words are deleted, the retrieval performance deteriorates. For example, if word “Mori” is deleted from the video caption “Mr. Mori assumes premiership” almost of the spoken documents including word “Mori” will not be retrieved, because the similarity between speech transcription and video caption is computed based only on the overlapping words in the conventional vector space model.

To solve this problem, the similarity must be computed between different words instead of word overlapping. From this viewpoint, in this paper, we propose an expanded vector space model which can compute the similarity between different words. The outline of the cross media retrieval system based on the conventional vector space model and expanded vector space model are respectively shown as follows.

Conventional vector space model:

1. Speech transcription is carried out for Japanese broadcast news documents.
2. OCR is carried out for video caption.
3. Weighting term is computed for each word in the transcribed
Japanese broadcast news documents. Then the important words
are extracted.
4. Spoken documents related to the video caption given as a query
are retrieved based on the standard cosine measure in the vector
space model.

Expanded vector space model:
1. Speech transcription is carried out for Japanese broadcast news
documents.
2. OCR is carried out for video caption.
3. Weighting term is computed for each word in the transcribed
Japanese broadcast news documents. Then the important words
are extracted.
4. Spoken documents related to the video caption given as a query
are retrieved based on the word space method in the expanded vector
space model.

3. SPEECH TRANSCRIPTION

3.1. EXPERIMENTAL CONDITION

We carried out speech transcription for the NHK spoken news program-
s containing 42 documents, using a language model and an acoustical
model. The language model is the word bigram constructed from R-
WC text database which was produced by morphologically analyzing the
MAINICHI Japanese newspaper of 45 months from 1991 to 1994. The
number of the words in the dictionary is 20,000. The word bigram was
back-off smoothed after cutting off at 1 word.

Speaker independent cross-word triphone HMMs were constructed. They
were trained using 21,782 sentences spoken by 137 Japanese males. These
speech data is taken from the database of acoustical society of Japan. The
acoustic parameters are 39 MFCCs with 12 Mel cepstrum, log energy and
their first and second order derivatives. Cepstrum mean normalization was
applied to each sentence to remove the difference of input circumstances.

In the transcription experiment, we used HTK (HMM Toolkit) as the
decoder which can perform Viterbi decoding with beam search using
above mentioned language model and acoustic model.

Table 1: Acoustic Analysis(AA) and HMM

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling frequency</td>
<td>12kHz</td>
</tr>
<tr>
<td>High-pass filter</td>
<td>(1 - 0.972^{-1})</td>
</tr>
<tr>
<td>Feature parameter</td>
<td>MFCC,Fow,Δ,ΔΔ(39th)</td>
</tr>
<tr>
<td>Frame shift</td>
<td>20ms</td>
</tr>
<tr>
<td>Window type</td>
<td>Hamming window</td>
</tr>
<tr>
<td>Learning method</td>
<td>Concatenated training</td>
</tr>
<tr>
<td>Type</td>
<td>Left to right continuous HMM</td>
</tr>
<tr>
<td>Number of states</td>
<td>3 states with 3 loops</td>
</tr>
<tr>
<td>Number of mixtures</td>
<td>8</td>
</tr>
</tbody>
</table>

3.2. TRANSCRIPTION RESULT

The transcription was carried out for the NHK spoken news programs con-
taining 42 documents broadcast in 1998. We used three streams of news
programs as shown in the most left column in Table 2. The total number
of time duration and sentences were 1.3 hours and 347 respectively. The
transcription result is shown in Table 2. In the table, the “Corr” indicates
the correctness defined by Eq.(1). On the other hand, the “Acc” indicates
the accuracy defined by Eq.(2). The reason why the transcription result is
a little lower is explained as follows. The language model was constructed
from the MAINICHI Japanese newspaper published from 1991 to 1994. On
the other hand, the test data was NHK spoken news broadcast in 1998.
This time difference mainly seems to cause the lower transcription result.
Further more, most of the topics included in the MAINICHI Japanese
newspaper are economy and politics, and other various topics are equally
distributed in the test data. This uneveness of topic distribution also seem-
s to cause the lowerness. This transcription result was used for spoken
document retrieval.

\[
\text{Percent Correct} = \frac{N - D - S}{N} \times 100\% \quad (1)
\]

\[
\text{Percent Accuracy} = \frac{N - D - S - I}{N} \times 100\% \quad (2)
\]

\[
N : \text{Total number of labels}
\]

\[
S : \text{Number of substitution errors}
\]

\[
D : \text{Number of deletion errors}
\]

\[
I : \text{Number of insertion errors}
\]

Table 2: Transcription result(%)

<table>
<thead>
<tr>
<th></th>
<th>Corr</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>19980820-12:00NHK</td>
<td>77.81</td>
<td>77.42</td>
</tr>
<tr>
<td>19980824-12:00NHK</td>
<td>75.8% (47/62)</td>
<td>75.34</td>
</tr>
<tr>
<td>19980825-12:00NHK</td>
<td>75.8% (47/62)</td>
<td>75.8% (47/62)</td>
</tr>
<tr>
<td>Total</td>
<td>75.34</td>
<td>72.44</td>
</tr>
</tbody>
</table>

4. CHARACTER RECOGNITION

In this section, the method to recognize video captions is described. Fig.2
shows the recognition flow. It consists of the following procedures.

1. The time sections (frames) where video captions appear are de-
tected at first.
2. Then character regions are extracted.
3. Finally the character regions are binarized and recognized.

4.1. DETECTION OF VC APPEARING SECTIONS

We have already proposed a technique to detect VC (video caption) appearing
sections in [10]. But the detection accuracy of VC appearing sections
was low. To solve this problem, we employed a feed-back method which
rejects the VC appearing sections where characters are not extracted. The
accuracy of VC appearing sections was improved (75.8-34.8=41.0%) by
using this method. Experimental results are shown in Table 3. In the
table, “Recall” indicates the ratio of the correctly extracted sections to
the ratio of total sections. On the other hand, the “Precision” indicates
the ratio of the correctly extracted sections to the number of total extracted sections.

Table 3: Detection of VC (video caption) appearing sections

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>19980820-12:00NHK</td>
<td>70.4% (47/67)</td>
<td>7.8% (47/62)</td>
</tr>
</tbody>
</table>
### 4.2. EXTRACTION OF CHARACTERS REGIONS

We carried out the method described in [10] for extraction of character regions from the VC appearing sections. The total number of character regions was 96. The character region extraction method is based on local line density. The experimental result is shown in Table 4.

**Table 4: Characters Region Extraction**

<table>
<thead>
<tr>
<th>Video Caption</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>19980820-12:00NHK</td>
<td>81.3% (78/96)</td>
<td>56.5% (78/137)</td>
</tr>
</tbody>
</table>

### 4.3. BINARIZATION and RECOGNITION

We carried out the method described in [10] for binarization and recognition of video captions. The binarization method is based on a three-level floating thresholding method. The binarized characters were read by commercially available OCR. The result is shown in Table 5.

TV viewers usually pay attention to video captions showing the outline of TV news documents, because video captions almost shows the outline of TV news documents. For this viewpoint, we selected video captions showing the outline of TV news documents from the extracted character regions shown in Table 4. Experimental results are shown in Table 5. As a result, their number of character regions showing the outline is 14. Their recognition results were used as queries.

### 5. CROSS MEDIA RETRIEVAL METHOD

#### 5.1. STANDARD COSINE MEASURE

In this section, we describe a method to retrieve related spoken documents by using video caption recognition results. At first, linguistic morphological analysis is carried out for video caption recognition results. Then a vector is constructed for one video caption using the extracted nouns. Next, each spoken document is converted into a vector using extracted important words. We employed mutual information incorporating TF-IDF to extract the important words from the spoken documents. Mutual information incorporating TF-IDF is described in section 5.3. The similarity degree is computed between the video caption vector \( X_k \) and the spoken document vector \( X_l \) as follows.

\[
\cos \theta = \frac{\mathbf{X}_k \cdot \mathbf{X}_l}{\| \mathbf{X}_k \| \| \mathbf{X}_l \|} = \frac{\sum_{i=1}^{n} x_{k,i} x_{l,i}}{\sqrt{\sum_{i=1}^{n} x_{k,i}^2} \sqrt{\sum_{i=1}^{n} x_{l,i}^2}}
\]

Here the vector \( X_k \) of video caption is defined as:

\[
X_k = (x_{1,k}; x_{2,k}; \ldots; x_{N,k})^T
\]

\[
x_{n,k} : \text{Normalized frequency of word appearing in only video caption } t_k
\]

\[
x_{n,k} : \text{Normalized frequency of word appearing in both video caption } t_k \text{ and spoken document } t_l
\]

The vector \( X_l \) of spoken document is defined as:

\[
X_l = (x_{1,l}; x_{2,l}; \ldots; x_{N,l})^T
\]

\[
x_{n,l} : \text{Normalized frequency of word appearing in only spoken document } t_l
\]

\[
x_{n,k} : \text{Normalized frequency of word appearing in both video caption } t_k \text{ and spoken document } t_l
\]

If \( \cos \theta \) nearly equals 1, the similarity between the video caption vector \( X_k \) and the spoken document vector \( X_l \) is regarded as high. Consequently, spoken documents whose similarity to the video caption vector is higher than some thresholds are retrieved as related spoken documents to video caption vector \( X_k \). But a large difference in number of words between the video caption vector \( X_k \) and the spoken document vector \( X_l \) decreases the retrieval performance, because Eq.(3) shows only word overlapping between \( X_k \) and \( X_l \). To solve this problem, we employed the similarity between different words as described in the next section.
5.2. WORD SPACE METHOD

Eq.(3) counts only the number of overlapping words, but it doesn’t take into consideration the similarity between different words. This causes the decreasing of retrieval performance in a case where the number of words included in video captions are small. To solve this problem, we propose word distance in a word space computed as similarity between different words. Though mutual information and co-occurrence are usually used as the similarity measure, they show the similarity between words in only one dimensional space. On the other hand, word distance shows the similarity between words in a three-dimensional space. Consequently, the similarity between words can be correctly computed by using word distance better than the mutual information and co-occurrence. The word distance between word \( w_i \) and \( w_j \) is computed as follows.

\[
WD(w_i, w_j) = \frac{1}{m_n} \sum_m \left( (TF(w_i,t_m) - TF(w_j,t_m))^2 + (IDF(w_i) - IDF(w_j))^2 + (t_m(w_i) - t_m(w_j))^2 \right)^{1/4}
\]

In Eq.(4), \( TF(w_i,t_m) \) shows the term frequency by which word \( w_i \) occurs in VC appearing section \( t_m \), \( IDF(w_i) \) shows inverse document frequency of the word \( w_i \), \( |t_m(w_i) - t_m(w_j)| \) shows mutual information of word \( w_i \) to VC appearing section \( t_m \). The \( m_n \) also shows the number of video caption appearing sections. Then, word distance shows the distance between word \( w_i \) and \( w_j \) in all VC appearing sections in the Mutual-TF-IDF space.

Next, the standard cosine measure is expanded to incorporate the word distance as shown in Eq.(5).

\[
(X_k, X_l) = \sum_i \sum_j x_{ik} x_{lj} \times \frac{1}{WD(w_i, w_j)}
\]

5.3. MUTUAL INFORMATION INCORPORATING TF-IDF

The conventional mutual information shows high value even when occurrence of word \( w_i \) is low, because mutual information is computed based on probability, not on word frequency. Therefore mutual information ignores occurrences of word \( w_i \). This is a weak point of the conventional mutual information. On the other hand, TF considers occurrences of word \( w_i \) and compensates for a weak point of the conventional mutual information. IDF shows degree how word \( w_i \) depends on document \( t_k \) by counting the number of documents which include word \( w_i \). Therefore it can be said that IDF shows the co-occurrence of word \( w_i \) and document \( t_k \) in different viewpoint from mutual information. This is the reason we propose here the combination of the mutual information and TF-IDF, as shown in Eq.(6) to extract important words better than the conventional methods. We call this method as mutual information incorporating TF-IDF.

\[
i(t_k, w_i) \propto (TF-IDF)
\]

\[
= (i(t_k) - i(t_k, w_i)) \times TF(w_i, t_k) \times IDF(w_i)
\]

\[
= (\log \frac{P(t_k|w_i)}{P(t_k)}) \times TF(w_i, t_k) \times IDF(w_i)
\]

6. CROSS MEDIA RETRIEVAL RESULTS

In evaluation experiments, we compared our proposed method with a conventional method with standard cosine measure described in section 5.1. Corpus we used consist of video and speech news data during August of 1998. The number of spoken documents and video captions query are 42 and 14 respectively. We have evaluated the cross media retrieval experiments by recall, precision and Fmeasure. Table 6 shows the result of cross media retrieval experiment. Table 6 shows the effectiveness of the cross media retrieval system we constructed. In the table, the recall is defined as the ratio of the number of the correctly retrieved documents to the number of total documents. On the other hand, the precision is defined as the ratio of the number of the correctly retrieved documents to the total number of retrieved documents. The proposed method improved Fmeasure by 30% compared to the conventional method.

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>Fmeasure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine</td>
<td>33.53</td>
<td>66.66</td>
<td>44.44</td>
</tr>
<tr>
<td>Word Space</td>
<td>91.66</td>
<td>64.70</td>
<td>75.85</td>
</tr>
</tbody>
</table>

7. CONCLUSION

We proposed the cross media retrieval technique which can retrieve news speech by video captions. We proposed the expanded vector space model based on a word space in cross media retrieval. The experimental results show that the word space method has superiority to the standard cosine measure in the vector space model. However, the results are still preparatory, because amount of database is small. In future, we will carry out experiments by using a large amount of database.

8. REFERENCES