TOWARDS RETRIEVAL OF VIDEO ARCHIVES BASED ON THE SPEECH CONTENT

Mei-fang HUANG, Kuan-ting CHEN, and Hsin-min WANG

Institute of Information Science,
Academia Sinica, Taipei
{mfhuang, kenneth, whm@iis.sinica.edu.tw}

ABSTRACT

Huge collections of video and audio recordings which have captured events of the last century remain an untapped resource of historical value. Accordingly, there are many digital library projects worldwide studying how multimedia digital libraries can be established and used. In this paper, we will report on some interesting findings from our recent work towards retrieval of video archives for Taiwan's humanity and social activities based on the speech content. We are currently focusing on the recordings about the aboriginals in Taiwan. Based on the acoustic models trained by broadcast news speech and language models trained by newswire texts, the recognition accuracy, which is 15.92% for syllables and 8.18% for characters, is disappointedly low. After applying the model adaptation techniques using some domain-specific training speech and text corpora, we are able to improve the accuracies to 30.04% and 22.08%, respectively. Though the accuracies are definitely not satisfactory, we found that it is still feasible to build a speech retrieval system for the target video archives.

1. INTRODUCTION

With the flourishing growth of mass media and the movie industry during the last century, how to store and utilize data from this amazing period has become the most challenging task facing researchers in the multimedia area in the 21st century. Extensive research has been done for both image and speech retrieval in various types of multimedia and brought out many digital museum projects worldwide. The best known one is The Informedia Digital Video Library Project conducted at CMU [1].

The "Digital Museum of Taiwan's Social and Humanities Video Archive" sponsored by the National Science Council in Taiwan aims to establish a digital museum based on the presentation of archival footage in video format, and to develop this collection into a world-class video database. Our current work is to help developing a video retrieval system using speech recognition and spoken document retrieval technologies and, then, to combine this with the video shot change detection, watermarking, metadata and interactive website devised by the members of other sub-projects. The raw video content is contributed by the chief administrator of this project, Professor Daw-Ming Lee of Taipei National University of the Arts, who has spent more than 10 years in recording the humane activities in Taiwan and has produced thousands of video archives. Undoubtedly, the establishment of this digital museum will preserve lots of valuable material on Taiwanese culture and society. In the first year, we are focusing on the recordings about the aboriginals in Taiwan.

Our past one-year centralized study on this project has led to some interesting findings. In this paper, we will introduce the discovered problems from the work towards speech retrieval of documentary films and share the experience of starting a new task using the existing general domain setup. Our team has many years of experience in dealing with Chinese broadcast news speech recognition and retrieval [2]. However, the content in this Digital Museum (DM) project is different from the broadcast news in the subject matter, speech style, recording environment, etc. Moreover, the programming format also varies from video to video. It's difficult and complex to do such a challenge work in a short time, but we are glad to say that after a preliminary analysis and test, we have brought up new ideas of retrieving speech segments in video archive and have overcome the first examination of the low recognition accuracy by adaptation methods for speech recognition. A reasonable recognition accuracy is critical to the success of developing a speech retrieval system, therefore to raise the recognition accuracy is the first task to deal with. Successively, non-speech filtering and automatic segmentation will be our second stage to achieve. We also construct a retrieval system to verify and realize speech retrieval technology.

The rest of this paper is organized as follows: The target database to address is introduced in Section 2. The baseline speech recognition evaluation is discussed in Section 3. Further analysis of the speech content and the model adaptation experiments are presented in Sections 4 and 5. Finally, the preliminary retrieval system is described in Section 6 and the concluding remarks are made in Section 7.

2. DATA COLLECTION AND ANALYSIS

The target database to address was originally in betacam video cassettes, each with a 30 minutes recording. For the sake of reservation and utility, the video was digitized to MPEG2 while the audio was digitized to 224kbps, 44.1kHz, 16bit, stereo format using MPEG1 layer 2 standard (ISO/IEC 11172-3) The compressed audio was decoded to waveforms for recognition and indexing purposes.

First of all, we spot-checked the content in some videos and found they widely ranged from interviews, chats, conversations to outdoor activities such as singing, dancing, ceremony and many other events. Furthermore, the speech was usually mixed with Mandarin Chinese, aboriginal languages and other dialects. Also, we found that the quality of some recordings was strikingly poor because of the noisy recording environment and the speakers' non-standard accents. It is foreseeable that the

characteristics of the speech in the videos we want to use in this project are much more complicated than any database we have handled before.

In order to assess the recognition accuracy for the new task, we have to create a test set. First we manually transcribed 13 speech segments from the video pool. This test set (termed DMtest in the rest of this paper) is made up of 10 speakers and the total length is about 25 minutes.

3. SPEECH RECOGNITION EVALUATION

Our first step is to conduct speech recognition on the test set based on the baseline setup for large vocabulary continuous speech recognition available at our laboratory.

3.1 Signal Processing

In our speech recognizer, spectral analysis is applied to a 20 msec frame of speech waveform every 10 msec. For each speech frame, 12 mel-frequency cepstral coefficients (MFCC) and the logarithmic energy are extracted, and these coefficients along with their first and second time derivatives are combined to form a 39-dimensional feature vector. In addition, to compensate for channel noise, utterance-based cepstral mean subtraction (CMS) is applied to the training and testing speech.

3.2 Acoustic Model

The acoustic units used in our speech recognizer are intra-syllable right-context-dependent INITIAL/FINAL, including 112 context-dependent INITIAL's and 38 context-independent FINAL's. Each INITIAL or FINAL is represented by a continuous density HMM with 1 to 4 states. The Gaussian mixture number per state ranges from 4 to 64, depending on the amount of corresponding training data. In addition, the silence model is a 1-state CDHMM with 64 Gaussian mixtures trained by using the non-speech segments. The baseline acoustic models were trained by using a database with 16 hours of broadcast news speech collected on air from several radio stations located at Taipei and finally a total of 11004 mixtures were obtained.

3.3 Language Model

The baseline syllable-based and word-based *N*-gram language models were trained by using a newswire text corpus consisting of 65 million Chinese characters collected from Central News Agency (CNA) in 1999. Word segmentation and phonetic labeling were performed for the training text corpus based on a 61521-word lexicon for training the *N*-gram language models.

3.4 Baseline Speech Recognition Experiments

The performance of the baseline acoustic and language models was evaluated by using the recognition results of 23 pieces of radio news which came from the same experimental condition with the training data. The syllable accuracy and character accuracy was 63.68% and 69.61%, respectively. We used the same setup with the previous baseline experiment to evaluate the DMtest data. Unfortunately, the recognition accuracy was disappointingly low, dropping from 63.68% and 69.61% to 15.92% and 8.18% for syllable accuracy and character accuracy respectively. The poor accuracy could result from poor recording quality, compression during digitization, domain mismatch, etc. Further analysis of the speech and further evaluation of the speech recognition are definitely necessary.

4. FURTHER ANALYSIS OF THE SPEECH

We carefully examined the content in the video and attempted to find the major cause of the recognition accuracy declination. One thing we have not dealt with was the channel selection problem. The original audio in the database was recorded in stereo but the recognition was only conducted on the left channel in the baseline experiments. Another putative reason was the compression effect. According to our experience on the Internet-accessible compressed Mandarin speech, the compression effect would be a major factor of low recognition accuracy [3].

4.1 Channel Effect

We reconsidered the channel selection by examining the right channel data and the 50%-50% mixed channel data of the DMtest set. The recognition results are summarized in Table 1.

File Name	Length (min:sec)	Mixed Channel	Left Channel	Right Channel
09-2059_1	0:52.93	* 26.42	25.00	24.53
09-2059_2	2:16.31	17.82	*18.39	17.82
03-2221_1	1:08.16	10.76	*12.79	12.50
03-2221_2	1:57.23	*12.66	11.39	10.13
08-2181	2:49.27	*8.26	5.92	N/A
12-2475	1:24.16	30.55	*31.99	N/A
14-2541	3:13.62	10.68	11.86	*12.37
13-2421	2:33.54	11.09	*16.34	9.34
17-2164	2:02.94	19.77	*22.46	19.39
2769	0:30.09	11.34	*15.46	11.34
2882_1	1:33.69	14.01	*22.25	17.03
2882_2	2:05.18	12.20	*14.26	10.88
05-2259	2:00.47	13.22	11.23	*14.10
Overall Result		14.55	15.92	14.20

Table 1. Speech recognition results with the different channel data. (* Remarks the highest speech recognition accuracy.)

It's hard to find a rule of which channel would outperform the other one from the speech recognition results of both the right and left channels. More interesting was that sometimes the mixed channel would have a better recognition result. After discussion with the owner of the films, Professor Lee, we understood that the stereo channels were produced from two different recorders; one was embedded on the betacam machine while the other was worn by or directed to the interviewee. Either left or right channel could be the major channel used by the interviewee while recording, and the equipment setup was according to the convenience and the restriction on the recording environment. Moreover, the major channel could change in the same video, if it contained several different recording segments. If we could dynamically select the correct channels, the syllable accuracy of the DMtest data would rise to 16.21%. We also attempted to judge the major channels by human hearing, but the ones

identified as major channel were not always exactly those with higher recognition accuracy. Moreover, owing to the constraint on visual effect, the microphone was kept off from the camera and the constraint thus resulted in the lower speech fidelity.

4.2 Compression Effect

To see whether compression is a major factor, we let the developing set of the broadcast news speech undergo the same compression/decoding process (MPEG1-layer2). The recognition results are summarized in Table 2. A very interesting observation is that, when the compression ratio was not heavy, the MPEG compression/decoding process did not degrade the recognition accuracies but slightly improved them. The reason could be due to the sub-band filtering and psychoacoustic modeling techniques, which function as noise reduction or speech enhancement, embedded in the MPEG audio compression algorithm [4]. However, with a higher compression ratio, the recognition accuracies did drop though slightly.

	Syllable Acc %	Character Acc %
Original Wave	63.68	69.61
MPEG Bit rate 112 (Ratio 1: 2.2)	63.72	70.39
MPEG Bit rate 40 (Ratio 1: 6.4)	61.82	66.82

Table 2. Speech recognition results of the developing set with or without the MPEG1- layer2 compression/decoding process.

The compression effect might be the reason for the accuracy declination, but having the broadcast news training speech undergo the compression/decoding process might not be of much help in itself. It was clear from the above experiment that the compression would not have a significant effect in speech with higher quality. However, the compression would damage the DMtest speech, which contained noise and background sounds and was with a low signal to noise ratio (SNR). We tested the DMtest speech and found the SNR ranged from 3.6dB to 16dB. Another probable reason might be the different speaker characteristics. The average speaking rate of the DMtest data ranged from 2.61 to 5.48 syllables per second. Furthermore, the speech type was mixed with spontaneous and planned forms and was quite different from the training speech. We believed that if we could eliminate the mismatch between the training data and the DMtest data, we would be able to get a better recognition result. Therefore, we applied the adaptation techniques to see if the recognition accuracy would be raised or not.

5. ADAPTATION OF SPEECH RECOGNITON

The adaptation was applied to both the acoustic model and the language model to enhance the speech recognition accuracy.

5.1 Acoustic Model Adaptation

In order to make the acoustic models better matched with the DM data, we selected several speech sections from the video database and manually transcribed 3.25 hours speech for both acoustic model re-training and MAP [5] adaptation. After 6 iterations of re-training and MAP adaptation, the syllable accuracy was significantly improved from 15.92% to 27.17% and 26.05%, respectively. This result showed that using a small amount of domain-specific training speech data to mix with the original training speech data is a good way to improve the acoustic model performance. Though the amount of domain-specific training speech (3.25 hours) is much less than that of broadcast news training speech (16 hours), it's interesting to find that training a whole new set of acoustic models outperformed adapting the acoustic models via MAP though the difference is not very significant.

5.2 Language Model Adaptation

After applying the acoustic model adaptation techniques, we then focused on the language model adaptation work. We found that many specific words for the target domain don't appear in the CNA99 corpus and are not included in the lexicon either. We have thus collected an 8MB text corpus from websites about the Formosan aboriginals. After removing some functional terms from the HTML files, the final plain text corpus (termed WEB in the rest of this paper) consists of 3.2 million Chinese characters (6.4MB). The perplexities of the test set under the syllable and word bigrams trained from the CNA99 corpus are 92.8 and 1383.7, respectively, but are significantly reduced to 59.1 and 491.2 under the language models trained solely from the domain-specific corpus (WEB), based on the baseline 61521-word lexicon. The perplexities of the test set under the language models obtained by interpolating the language models trained from the CNA99 corpus and from the domain-specific corpus (WEB) using different weights are summarized in Table 3. These results showed that the language models trained solely from the domain-specific corpus have the lowest perplexity, and the interpolated language models based on different weights always have lower perplexities than the language models trained from the CNA99 corpus. From the above results, we inferred that the target data was well centralized in the domain-specific corpus we collected from the Internet.

Weight (CNA:WEB)	1:0	0.75:0.25	0.5:0.5	0.25:0.75	0:1
Syllable Bigram	92.8	73.0	64.8	60.0	59.1
Word Bigram	1383.7	736.3	603.0	542.9	491.2

Table 3. The perplexity of the test set under the interpolated language models.

Weight (CNA:WEB)		1:0	0.75:0.25	0.5:0.5	0.25:0.75	0:1
Old Lexicon 61521 words	Unigram	3536.9	2336.1	1930.4	1699.4	1368.3
	Bigram	1383.7	736.3	603.0	542.9	491.2
New Lexicon 65195 words	Unigram	3619.4	2400.7	1976.0	1735.8	1393.5
	Bigram	1439.5	802.2	655.9	589.9	532.9

Table 4. The perplexity of the test set under the word-based language models.

Because many specific words in the domain-specific corpus are not included in the baseline lexicon, we augmented the lexicon with 3674 new words extracted from the domain-specific corpus using the keyword extraction tool developed by Chen and Ma [6]. We then re-tokenized the CNA99 corpus and the domain-specific corpus, and re-trained the language models. The perplexities of the test set under the word-based language models obtained by combining the word-based language models trained from the CNA99 corpus and from the domain-specific corpus are summarized in Table 4. Using the augmented lexicon slightly raised the perplexity because of the larger number of word

entries it contained. At this stage, we did not remove any words from the old lexicon before augmenting it. We believe that if we can further manipulate the lexicon, the perplexity can be reduced.

In addition to the perplexity, we also evaluated the domain-specific language models and the augmented lexicon using the recognition accuracy. The results are summarized in Table 5. Although the augmented lexicon increased the test data perplexity as shown in Table 4, it improved the character accuracy slightly from 14.65% to 15.07%. The reason could be due to the characteristics of the *N*-gram language model framework. Some of the domain-specific terms in the test data might exist in the CNA99 corpus. They would be tokenized as multi-character words using the augmented lexicon while tokenized as a string of mono-character words using the old lexicon. It is clear that these terms are more likely to be correctly recognized using the augmented lexicon. When we applied the language models trained from the domain-specific corpus, the character accuracy was further improved to 22.08%. It's interesting that the character accuracy obtained based on the old lexicon and the domain-specific corpus was slightly better than that obtained based on the augmented lexicon and the domain-specific corpus. The reason could be that even though a domain-specific term is an OOV for recognition, there is still a chance that our recognizer produces sequentially mono-character words. In summary, when we applied the new acoustic models, new lexicon, and new language models trained from the domain-specific corpus in the speech recognizer, the syllable accuracy was improved from 15.92% to 30.04% while the character accuracy was improved from 8.18% to 22.08%.

Language Model Corpus		CNA99 130MB	CNA+20*WEB (~Weight 0.5:0.5)	WEB 6.4MB
Syllable Accuracy %		27.17	29.53	30.04
Character Accuracy %	Old Lexicon	14.65	19.90	22.11
	New Lexicon	15.07	18.68	22.08

Table 5. Speech recognition results with or without the augmented lexicon and the domain-specific language models.

6. THE RETRIEVAL SYSTEM

The goal of this work is to establish a video retrieval system based on the speech content. For demonstration and evaluation purposes, we have randomly selected 8 betacam cassettes for building the prototype retrieval system. The speech wave was automatically chopped into 28-32 seconds segments based on energy detection. A total of 386 segments were obtained. Every speech segment was regarded as a document and has its own set of indexing vectors, each consists of overlapping syllable *N*-gram (*N*=1, 2, or 3) or overlapping syllable pair separated by *n* syllables (n=1, 2, and 3) [2].

We have implemented a web-based retrieval system, which can accept text queries to retrieve spoken documents. A text query is first converted into a syllable string and, then, a set of indexing vectors can be derived via exactly the same way as we process the documents. Cosine measure is used to estimate the distance between a query and a document. The retrieval system has been integrated into the search interface of the Digital Museum of Taiwan's Social and Humanities Video Archive. As shown in Figure 1, in addition to the speech retrieval function, the

Figure 1. The search interface of the Digital Museum of Taiwan's Social and Humanities Video Archive.

interface also provides a full-text search function and a metadata search function.

We have randomly selected 40 domain-specific terms as queries to evaluate the retrieval performance. 77.5% of the queries were able to get the relevant speech segment when only one segment was returned, while 87.5% were able to get at least one relevant segment within 3 retrieved segments. By taking advantage of this situation, we can apply relevance feedback techniques to enhance retrieval performance.

7. CONCLUSION AND FUTURE WORK

We have presented some preliminary work towards retrieval of video archives using the speech content. Substantial work to further improve recognition accuracy as well as to investigate further problems is definitely necessary. Our ongoing studies mainly include unsupervised speaker/environment adaptation and automatic classification, including music speech detection, language identification, speaker identification and clustering, etc. Hopefully, these techniques can further improve recognition accuracy and retrieval performance.

8. REFERENCES

[1] CMU Informedia Digital Video Library project http://www.informedia.cs.cmu.edu/

[2] B. Chen, H. M. Wang, and L. S. Lee, "Retrieval of broadcast news speech in Mandarin Chinese collected in Taiwan using syllable-level statistical characteristics," *ICASSP2000*.

[3] W. P. Hsieh, B. Chen, K. T. Chen and H. M. Wang, "Recognition of Internet-accessible compressed speech for audio indexing," *ISCSLP2000*.

[4] D. Pan, "A tutorial on MPEG/audio compression", *IEEE Multimedia*, pp. 60-74, 1995.

[5] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Acoustics Speech and Signal Processing*, vol. 2, pp. 291-298, April 1994.

[6] K. J. Chen and W. Y. Ma, "Unknown word extraction for Chinese documents," *COLING2002*.