

Application of Speech Recognition with Closed Caption for Content-Based Video Segmentation

Jongmok Son, Jinwoong Kim**, Kyungok Kang**, Keunsung Bae**

*School of Electronic and Electrical Engineering, Kyungpook National University, Taegu, Korea

**Broadcasting Technology Dept., Electronics and Telecommunications Research Institute, Taejeon, Korea

ABSTRACT

An important aspect of video indexing is the ability to segment video into meaningful segments, i.e., content-based video segmentation. Since the audio signal in the sound track is synchronized with image sequences in the video program, a speech signal in the sound track can be used to segment video into meaningful segments. In this paper, we propose a new approach to content-based video segmentation. This approach uses closed caption to construct a recognition network for speech recognition. Accurate time information for video segmentation is then obtained from the speech recognition process. For the video segmentation experiment for TV news programs, we made 56 video summaries successfully from 57 TV news stories. It demonstrates that the proposed scheme is very promising for content-based video segmentation.

1. INTRODUCTION

Advances in computer and communication techniques have produced a flood of information. Comparing to the text and audio information, the amount of video information in particular has led to an unprecedented high volume of data. As more and more video databases are digitized and stored in accessible archival files, the need grows for effective ways to search and retrieve useful information from them. To achieve this goal, more efficient and appropriate ways for making video summary and video indexing should be developed [1][2][3]. An important aspect of making video summary and video indexing is the ability to segment video into meaningful segments, in other words, the ability of content-based video segmentation.

Since the audio signal in the sound track is synchronized with image sequences in the video program, a speech signal in the sound track can be used to segment video into

meaningful segments. Thus some researchers tried to make use of speech recognition and closed captioning data for video segmentation[3], but they could not get satisfactory results due to poor recognition performance. They did not use closed caption directly in the recognition process, but they used it for time alignment with inaccurate speech recognition results. Closed caption is spoken portion of television show or video program that appears as text at bottom of screen when processed by a decoder installed in set.

To avoid difficulties in the recognition problem, in this paper, we propose a new approach to video segmentation using speech recognition with closed caption. We use closed caption directly in the recognition process by constructing a network with transcripts of closed caption. Since the closed caption contains the contents of video program, it can be used effectively for the purpose of content-based video segmentation. However, the time that closed caption appears does not coincide with that of video. So it is necessary to find the time of video that corresponds to the contents of closed caption. This job can be done using closed caption-based speech recognition technique because the speech signal in the sound track is synchronized with video. The proposed scheme has been tested with real TV news program, and experimental results are presented and discussed.

2. CONTENT-BASED VIDEO SEGMENTATION

2.1 Video Segmentation with Speech Recognition

The proposed system of content-based video segmentation using speech recognition with closed caption is shown in Figure 1. From the script of closed caption, a keyword or key-sentence that contains the video information to be segmented is searched first. Since most of closed caption appears several seconds later than the speech signal, it is needed to find the time interval of the speech/video signal that corresponds to the contents of closed caption. Speech recognition with closed caption is used to find this time

This work was funded by the Broadcasting Technology Dept., Electronics and Telecommunications Research Institute in Korea

information from the speech signal, which corresponds to the video frames to be segmented.

Video segmentation algorithm with speech recognition is summarized as follows.

- Step 1: Input the keyword or key-sentence that includes contents of video to be segmented.
- Step 2: Search the keyword or key-sentence from the script of closed caption, and find the beginning and end frame numbers of video in which the caption of keyword or key-sentence appears.
- Step 3: Determine the search region in the sound track for speech recognition based on the beginning and end video frame numbers. Caption usually appears 2~7 seconds later than the audio signal in the TV news.
- Step 4: Construct the recognition network by adding more words before and after keyword or key-sentence and concatenating them.
- Step 5: Detect the keyword or key-sentence in the sound track using speech recognition with the network, and determine the corresponding time information of the audio signal.
- Step 6: Compute the beginning and end video frame numbers from the time information of the audio signal.

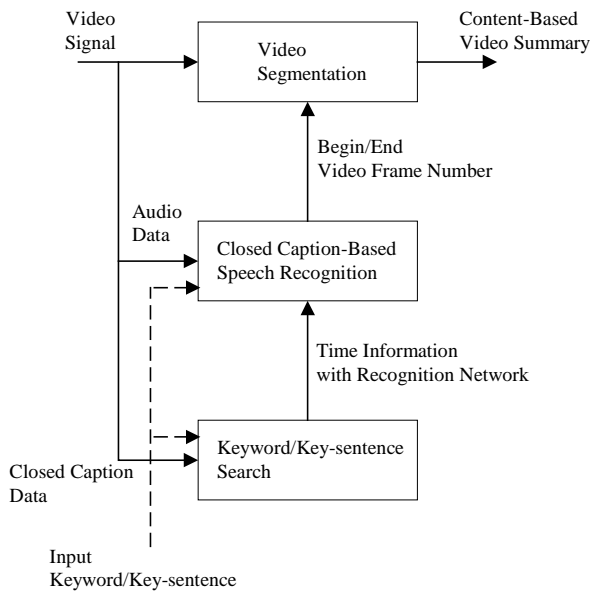


Figure 1. Overview of the proposed content-based video segmentation system

- Step 7: Segment the video using the beginning and end frame numbers obtained from the step 6. Then video summary that includes contents of the keyword or key-sentence as well as audio signal is obtained.

2.2 Speech Recognition with Closed Caption

A context dependent semi-continuous HMM is used for each phoneme-like unit HMM[4-7]. The topology for a state transition of a HMM is a 3-state left-to-right Bakis model as shown in Figure 2. The recognition network is constructed by concatenating words that include keyword or key-sentence and about 10 words before and after them, respectively. Figure 3 shows an example of a recognition network for detection of keyword in the closed caption. Each word in a recognition network is represented by cascading phoneme-like unit HMMs. Considering the pause duration between words or sentences in the speech signal, a silence model with skipping is included in the network. In the figure, token represents a set of parameters that contains information of time index of input speech signal, the value of partial likelihood ratio in each state, and the history of a token's route through the recognition network.

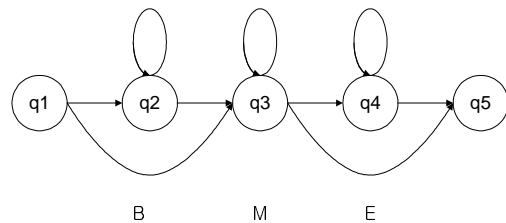


Figure 2. The topology of a state transition model for a phoneme-like unit HMM.

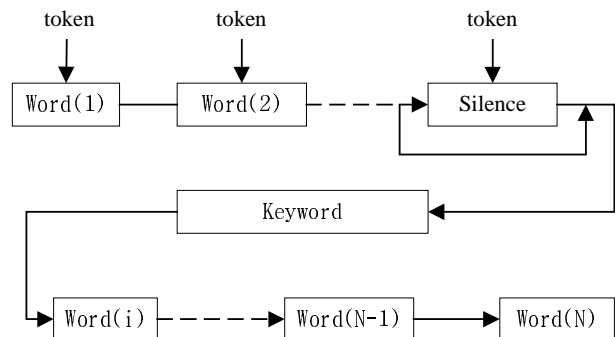


Figure 3. An example of the recognition network.

An alternative formulation of the Viterbi algorithm, i.e., a token passing algorithm is used to detect the keyword in the recognition network. The detection procedure of the keyword from the speech signal using speech recognition in the search region of the sound track is summarized as follows.

- Step 1: Insert initialized tokens into the every word of a recognition network that can be matched with the beginning of the search region of the audio signal. Generally, token is assigned to every word locating before the keyword in the network.
- Step 2: Compute the partial likelihood ratio and pass a copy of token in each state of HMM to the connected next state in the network at each time index t of the input signal, and then update the token information.
- Step 3: Examine the tokens in every state and remove them if estimated partial likelihood ratio is lower than a predefined threshold.
- Step 4: Repeat step 2 and 3 until the time index t reaches the end of the search region of the audio signal.
- Step 5: Select the token having the highest LR value, where LR is defined as the sum of 90% of total likelihood ratio and 10% of partial likelihood ratio in the keyword region. The state having the highest LR is matched with the end of the audio signal in the search region.
- Step 6: Detect the keyword region in the audio signal using route information obtained from the selected token.

3. EXPERIMENTS AND DISCUSSION

Experiments of video segmentation have been carried out with the proposed scheme. The database consists of approximately 60-minute long MPEG1 formatted TV news with closed caption. In our experiment, closed caption and audio signal were separated manually from the MPEG1 formatted data. The script of the closed caption was given as a text file, and the audio signal in the sound track was given as a PCM wave file with sampling rate of 16 kHz and 16 bits/sample quantization.

As feature parameters for speech recognition 12th order MFCC(Mel-Frequency Cepstral Coefficient) and signal energy were used. The energy and MFCC features were augmented by delta energy and 12th order delta MFCC calculated over 7 frames. Analysis condition of the audio signal for speech recognition is shown in Table 1. Each phoneme-liked unit HMM in the speech recognizer was

Table 1. Analysis condition of audio data

Preemphasis factor	0.95
Analysis window	Hamming
Analysis window size	20 ms (320 samples)
Frame rate	10 ms (160 samples)
Feature parameters	Energy Delta Energy 12th order MFCC 12th order delta MFCC

trained with 445DB and 611DB made by Electronics and Telecommunications Research Institute in Korea.

Figure 4 shows examples of keyword and key-sentence detection in the speech signal from a male anchor's voice in the TV news program. The rectangular box denotes extracted speech segment. Figure 4(a) displays the extracted speech segment corresponding to the keyword in the caption, that is pronounced as */bukbanghangaeseon/*. Figure 4(b) is the result for key-sentence in the caption. The boxes in the figure denote the beginning and end words of the sentence, respectively. By listening the extracted speech segments, it was confirmed that they coincided well with the keyword and key-sentence in the caption. Similar result for a female anchor's voice with keyword pronounced as */kimjeongil/* is shown in Figure 5.

Using the scheme we proposed, we did the content-based video segmentation to make TV news summaries. From the 60-minute long TV news program, 57 different stories were selected and key sentences were obtained from the closed caption. Then the beginning and end of video frame numbers for segmentation were computed from the time information of the speech signal that was obtained from speech recognition results. When we consider the segmented video having inserted or deleted speech portion at the beginning or end of each key sentence as a failed one, we made 56 video summaries successfully from 57 TV news stories. With the viewpoint of word detection task, we achieved 97.34% of detection rate. Among the 114 spoken words there were 3 errors. One error was due to the spoken word not given in the closed caption, and others were caused by the reporter's voice with severe noise in the TV news. Reporter's voice usually contains large amount of ambient noise including other people's voice and background noise, while anchor's voice produced in the studio does not.

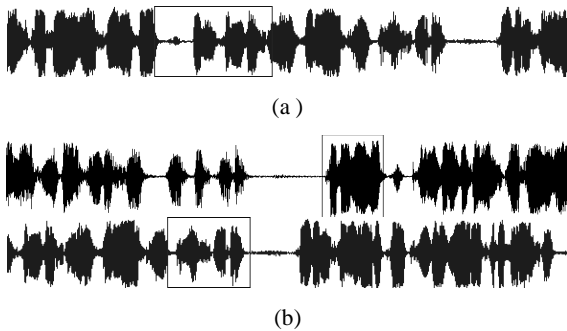


Figure 4. Examples of keyword and key sentence detection from a male anchor's voice in TV news (a) keyword (b) the beginning and end words of the key-sentence.



Figure 5. An example of keyword detection from an anchorwoman's voice in TV news.

4. CONCLUSION

A new approach for content-based video segmentation has been proposed. This approach uses closed caption directly in the speech recognition process to extract speech segment of keyword/key-sentence from the audio signal. Since the audio signal in the sound track is synchronized with image sequences, the time information for meaningful video segmentation can be obtained from that of the extracted speech segment. The proposed scheme has proven to be very promising for content-based video segmentation as well as indexing and retrieval of multimedia database with closed caption.

In this experiment closed caption and audio signal were obtained manually from the MPEG1 formatted data. Further studies will include constructing the system for automatic separation of audio signal and closed caption with its time index, and implementation of an automatic and/or semiautomatic video indexing system.

5. REFERENCE

- [1] John S. Boreczky and Lynn D. Wilcox, "A Hidden Markov Model Framework for Video Segmentation Using Audio and Image Features," *Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. VI, pp. 3741-3744, 1998.
- [2] Claude Montacie and Marie-Jose Caraty, "Sound Channel Video Indexing," *Proc. European Conf. on Speech Communication and Technology*, vol. 5, pp. 2359-2362, 1997.
- [3] Howard D. Wactlar, Alexander G. Hauptmann and Michael J. Witbrock, "IFORMEDIA™ : News-On-Demand Experiments In Speech Recognition," *Proc. of ARPA Speech Recognition Workshop*, Feb. 18-21, 1996.
- [4] C.-H. Lee, L.R. Rabiner, R. Pieraccini, and Jay G. Wilpon, "Acoustic Modeling of Subword Units for Speech Recognition," *Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 721-724, 1990.
- [5] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. of IEEE*, vol. 77, no. 2, pp.257-286, Feb., 1989.
- [6] B.-H. Juang, L.R. Rabiner, "The Segmental K-Means Algorithm for Estimation Parameters of Hidden Markov Models," *IEEE Trans. On ASSP*, vol. 38, no. 9, pp. 1639-1641, Sep., 1990.
- [7] Mosur K. Ravishankar, "Efficient algorithms for Speech Recognition," *Ph.D. thesis, CMU*, 1996.