캡션정보 및 음성인식을 이용한 내용기반 비디오 정보 색인 및 검색에 관한 연구

손종목*, 김진웅**, 배건성*
*경북대학교 전자전기 공학부
**한국전자통신연구원 방송기술 연구부

A Study on the Content-Based Video Information Indexing and Retrieval Using Closed Caption and Speech Recognition

Jong Mok Son*, Keun Sung Bae*, Jin Woong Kim**
*School of Electronic and Electrical Engineering, Kyungpook National University.
**Broadcasting Technology Department, Electronics and Telecommunications Research Institute

요약

뉴스나 드라마, 영화 등의 비디오에 대한 검색 시 일반 사용자의 요구에 가장 잘 부합되는 결과를 얻기 위해 비디오 데이터의 의미적 분석과 색인을 만드는 것이 필요하다. 일반적으로 음성신호가 비디오 데이터의 내용을 잘 나타내고 비디오와 동기가 이루어져 있으므로, 내용기반 검색을 위한 비디오 데이터 분할에 효율적으로 이용될 수 있다. 본 논문에서는 캡션 정보가 주어지는 방송뉴스 프로그램을 대상으로 효율적인 검색, 색인을 위한 비디오 데이터의 분할에 음성인식 기술을 적용하는 방법을 제안하고 그에 따른 실험결과를 제시한다.

Ⅰ. 서 론

오늘날에 있어서 정보통신 기술의 발전으로 텍스트 뿐만 아니라 비디오, 오디오 등의 디지털 멀티미디어 정보에 대한 요구가 크게 증가하고 있다. 따라서, 멀티미디어 정보를 효율적으로 표현하고, 저장, 검색하는 것에 대한 필요성이 증대되고 있으며, 다양한 사용자의 요구를 수용하기 위해서 비디오 데이터의 내용에 기반해 데이터를 분할하고 색인을 만드는 것이 필요하다.

멀티미디어 정보의 검색을 위해서 사람이 직접 데이터를 분할하고 색인을 만드는 것은 시간과 비용을 많이 소모하는 작업이기 때문에, 방대한 비디오 데이터의 처리를 위해서 이를 자동화하는 것이 요구된다. 이를 위해서 비디오 데이터의 효율적인 검색 및 색인 방법이 필요하며, 영상특징이나 오디오특징을 사용한 색인, 검색, 분할 기법들이 활발히 연구되고 있다[1,2]. 또한, 연속음성인식 기법이나 화자인식 기법 등을 오디오 데이터의 색인과 검색에 이용하려는 연구도 활발히 진행되고 있다 [3,4,5].

일반적으로 음성신호가 비디오 데이터의 내용을 잘 표현하기 때문에, 이를 사용하여 내용기반 검색 및 색인을 용이하게 할 수 있으며, 연속음성인식 기법을 사용하여 비디오 데이터 분할, 색인 검색에 사용하려는 연구가 많이 이루어지고 있다. 하지만, 현재 연속음성에 대한 인식 기술이 불완전하므로 신뢰성 있는 색인과 검색을 위해서는 보완이 필요하다.

드라마 및 방송뉴스 비디오 프로그램에 주어지는 캡션정보는 프로그램에 포함된 음성정보 및 오디오 정보를 문자로 표현하고 있으므로 이를 비디오 내용기반 검색 및 색인에 이용할 수 있으며, 현재 드라마, 뉴스 등의 비디오 정보에 캡션 정보가 포함되는 경향이 증가하고 있다. 또한, 캡션정보는 음성신호에 대응되는 문자정보를 제공해주므로 연속음성인식 기술의 불완전한 문제점을 상당히 줄여 줄 수 있다. 따라서, 본 연구에서는 캡션 정보 및 음성인식기술을 이용하여 비디오 데이터를 분

할, 검색하는 방법을 제안하고 그에 따른 실험 결과를 제시한다. 본 논문의 구성은 다음과 같다. 2 장에서는 캡션 정보 및 음성인식 기술을 이용한 내용기반 비디오 분할 방법에 대해 설명하고, 3 장에서 실험결과를 제시하고 검토한다. 마지막으로 4 장에서 결론을 내리고 향후 연구방향을 제시한다.

Ⅱ. 캡션정보 및 음성인식 기술을 이용한 내용기반 비디오 분할

2.1 비디오 데이터 분할

드라마 및 방송뉴스의 비디오 프로그램에 주어지는 캡션정보는 프로그램에 포함된 음성정보 및 오디오 정보를 문자로 표시해준다. 때문에, 캡션정보의 문자열을 이용할 경우 자연어 처리 기법을 적용하여 내용기반 비디오 프로그램의 분할이 용이하며, 이를 이용하여 비디오 프로그램의 색인 및 검색을 효율적으로 수행할 수 있다.

비디오 프로그램에서 비디오 데이터와 음성 데이터는 동기가 이루어져 있지만, 캡션정보는 그러하지 못하다. 특히, 방송뉴스 비디오 프로그램의 경우에서는 음성신호에 비해 상당한 시간이 경과한 후에 나타나는 경향을 보인다. 때문에, 캡션정보를 내용기반 비디오 검색에 사용하기 위해서는 음성인식 기술을 사용

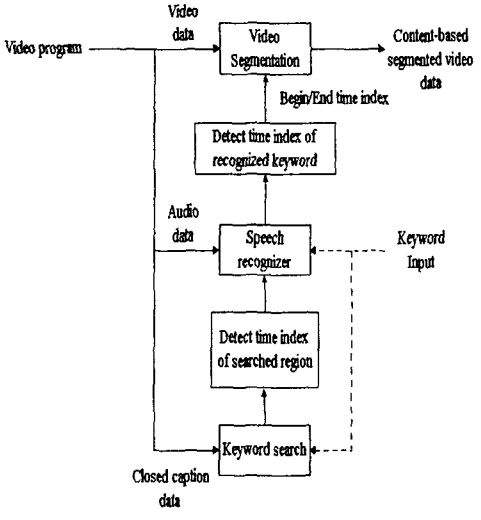


그림 1. 비디오 색인 및 검색시스템 구성도

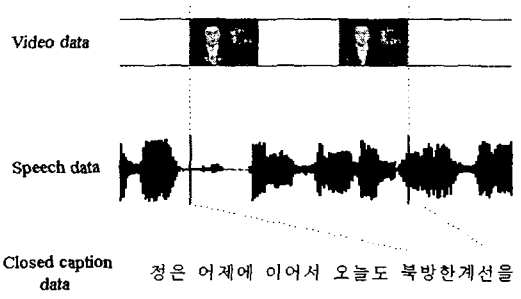


그림 2. 캡션정보를 이용한 비디오 분할

하여 사용자가 요구하는 캡션정보가 포함되는 비디오/오디오 구간을 정확하게 찾아내어야 한다.

캡션정보 및 음성인식 기술을 사용하여 본 연구에서 제안한 내용기반 비디오 색인 및 검색 시스템을 그림 1 에 나타내었다. 또한 주어진 키워드에 대응되는 음성구간을 검출하여 비디오 프로그램을 분할하는 예를 그림 2 에 나타내었으며, 그 과정은 다음과 같다.

STEP 1. 색인 또는 검색하고자 하는 내용의 키워드를 입력한다.

STEP 2. 캡션정보의 문자열에서 키워드를 검출하여, 키워드에 해당하는 캡션정보가 나타난 비디오 프레임의 시간을 찾는다.

STEP 3. 캡션정보가 오디오 데이터와 동기되어 있지 않으므로, 키워드에 해당하는 캡션정보가 나타난 비디오 프레임의 시간을 기준으로 오디오 데이터의 탐색 영역을 정한다.

STEP 4. 캡션정보에서 키워드가 나타난 주위의 문자정보를 이용하여 인식하고자 하는 음성모델을 구성한다.

STEP 5. 4)번 과정에서 구한 음성모델을 이용하여 3)번 과정에서 정한 탐색영역에서 인식기를 사용하여 키워드 또는 문장에 해당하는 음성신호 구간을 검출한다.

STEP 6. 5)번 과정에서 구한 음성신호의 time index 를 이용하여 비디오 데이터를 분할한다.

또한, 입력된 키워드가 포함된 문장의 시작 어절과 끝 어절에 대응되는 음성신호를 검출함으로써 문장단위의 검출을 할 수 있다. 앞

표. 1 음성데이터 분석 조건

Sampling Frequency	16 kHz
Quantization	16 bits
Hamming Window	20 ms(320 points)
Frame rate	10 ms(160 points)
Feature Parameters	1 order : energy 1 order : delta energy 12 order : MFCC 12 order : delta MFCC

에서 설명한 비디오 분할 과정으로 음성신호를 검출할 경우 내용기반 색인 및 검색에 효과적으로 응용할 수 있다.

2.2 음성인식 시스템 및 음성모델

오디오 트랙에서 얻어지는 음성 신호는 preemphasis 계수 0.95 로 전처리한 후, 20ms 길이의 해밍 윈도우를 10ms 간격으로 오버랩하여 구간단위로 분석하였다. 각 구간에서 1 차의 에너지와 12 차의 멜켑스트럼을 구하고, 현재 구간을 포함한 전후 각 3 구간(전체 7 구간)의 정보를 이용하여 1 차의 차분 에너지와 12 차의 차분 멜켑스트럼을 구하였다. 표 1 에 음성 데이터의 분석 조건을 나타내었다.

음성모델의 구성은 음소모델을 연결시켜 사용하였다. 음소모델은 문맥종속(Context dependent) SCHMM (Semi-Continuous HMM)을 사용하였다. 하나의 음소를 모델링하기 위하여 3 상태 Bakis(Left-to-Right) 모델을 사용하였으며, 그림 3 에 음소 모델 HMM 의 구성을 제시하고 있다. 각 음소 모델은 K-means 알고리듬과 Baum-Welch 재추정식을 사용하여 훈련하였으며, 훈련 데이터는 ETRI 의 445 DB 와 611 DB 를 사용하였다.

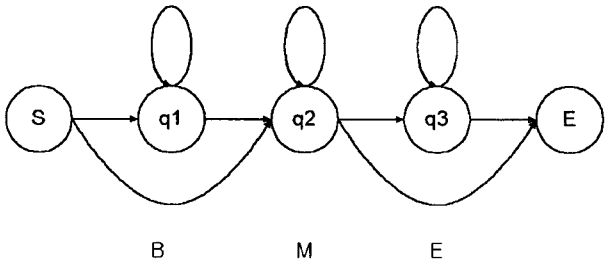


그림 3. 음소모델 HMM 의 구성

표 2. 발음사전 구성 예

어구	발음기호
긴장	g i n cl* j a ng > #
상황이	s a ng h wa ng i > #
이렇게	i r eo cl k e > #

음성모델은 캡션정보로 부터 키워드를 검출한 후 전후로 10 여개의 단어를 연결하여 구성한다.

주어지는 캡션 정보를 인식기에서 사용하기 위해서는 문자열을 발음 기호열로 표현해야 한다. 한글의 자소가 구체적인 음운현상을 반영한 것이 아니기 때문에, 한글 자소를 발음기호로 나타내는 것이 필요한데, 이를 위해서 자음동화, 된소리되기, 연음법칙, 음운 축약, 구개음화, 끝소리 규칙 등의 한글 읽기 규칙을 적용하였다. 이에 voiced closure, voiceless closure, voice offset 의 부가적 음운을 611DB 의 레이블링에 나타난 확률적 특성을 사용하여 첨가함으로써 발음 사전을 구성하였다. 표 2 에 발음사전 구성의 예를 나타내었다.

발음사전에 등록된 단어들의 연결 순서를 나타내기 위해 인식망을 구성하는데, 기본적으로 캡션 정보의 문자열이 나타나는 시간적인 순서대로 연결순서를 정하였으며, 단어간 또는 문장간에 나타나는 묵음 구간을 나타내기 위해 묵음 모델을 첨가 하였다. 또한, 단어간에 묵음구간 없이 이어지는 부분이 많기 때문에 첨가된 묵음 모델의 생략도 허용하였다. 예를 들어, 방송뉴스 프로그램의 /북한 경비정은 어제에 이어서 오늘도 북방한계선을 넘어오지

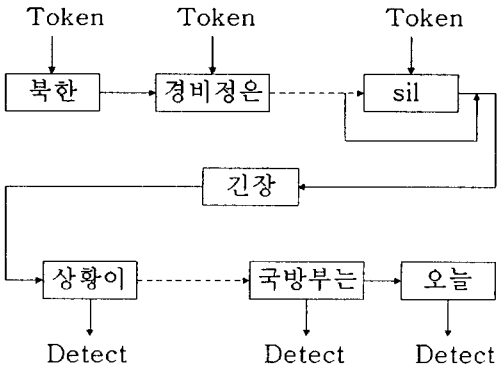


그림 4. 인식망 구성 예

않았습니다. 긴장 상황이 이렇게 진정 국면에 접어들면서 국방부는 오늘/이라는 캡션정보에서 /긴장/이라는 키워드의 음성신호를 검출하기 위한 인식망 구성 예를 그림 4 에 나타내었다.

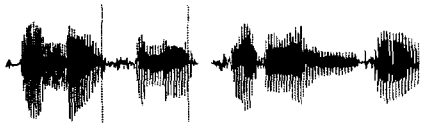
음성모델을 사용하여 오디오 데이터의 탐색영역에서 키워드 부분을 검출하는 과정은 다음과 같다. 캡션정보만으로는 오디오 데이터의 탐색 영역에서 시작음성을 정확히 알 수 없기 때문에, 음성모델에서 탐색영역의 시작이 될 수 있는 위치에 초기화한 Token 을 위치시킨다. Token 은 탐색영역에서의 시간정보와 확률적 유사도에 관한 정보, 그리고 Token 이 거쳐 온 경로에 대한 정보를 가지고 있다. 오디오 데이터의 탐색영역에 대해 음성모델을 사용하여 디코딩 과정을 거친 후 Token 의 확률적 유사도를 검사하여 확률이 가장 높은 Token 의 과거 경로 정보를 이용하여 키워드 구간을 검출하게 된다. 이때, 보통 탐색영역을 키워드가 포함될 정도로 크게 잡기 때문에, 키워드 이후의 위치에서 Token 을 검출한다.

디코딩 과정으로는 Viterbi 알고리듬을 사용하였으며, 계산량을 줄이기 위해 beam 탐색법과 *pruning* 기법을 적용하였다[6].

III. 실험 및 검토

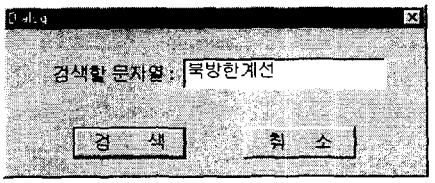
캡션정보 및 음성인식시스템을 이용하여 입력과 키워드를 이용하여 키워드 또는 키워드가 포함된 문장에 대응되는 음성신호를 검출하는 실험을 수행하였다.

(a) /여기는 신호처리 연구실입니다/

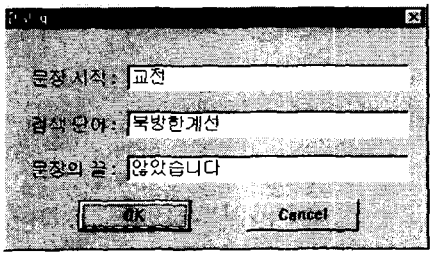


(b) /기는 신호처리 연구/

그림 5. /신호/ 검출 예



(a) 검색할 문자열



(b) 검색결과

그림 6. 키워드 입력 및 검색결과

우선, 탐색영역의 변동에 따른 영향을 살펴보기 위해 /여기는 신호처리 연구실입니다./라는 음성신호에 대해 /신호/라는 키워드의 검출실험을 하였다. 검출 결과를 그림 5 에 나타내었다. 그림 5(a)는 전체 문장에서 키워드에 대응되는 음성구간을 검출한 예를 보인 것이며, 그림 5(b)는 탐색영역이 키워드는 포함하고 있지만 그 영역이 임의로 주어졌을 경우의 검출 예이다. 탐색영역이 다르지만 성공적으로 /신호/라는 키워드를 검출하고 있음을 볼 수 있다.

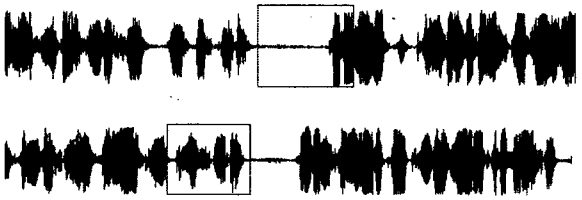
그림 6 은 아래에 보인 MBC 9 시 뉴스에서 얻은 캡션정보에서 입력된 키워드에 대해 키워드 및 키워드가 포함된 문장의 시작어절 및 끝어절을 검출한 결과를 나타낸 것이다.

캡션정보에서의 키워드 또는 문장 검출 결과를 바탕으로 한 음성신호 검출결과를 그림 7 에 나타내었다. 그림 8 은 여성기자의 데이터에 대한 키워드 검출 예이다.

캡션정보 :

/뉴스데스크입니다. 교전 이후 북한 경비정은 어제에 이어서 오늘도 북방한계선을 넘어오지 않았습니다. 긴장 상황이 이렇게 진정 국면에 접어들면서 국방부는 오늘 교전 상태는 사실상 끝났다고 말했습니다. 임태성 기자입니다./

(a) 키워드 /북방한계선/의 음성신호 검출결과



(b) 키워드가 포함된 문장의 시작 어절 /교전/과 끝 어절 /않았습니다./의 음성신호 검출결과

그림 7. 남성 앵커 뉴스에서 캡션정보에 대응되는 음성구간 검출 예

그림 8. 여성 기자 뉴스에서 키워드 /김정일/의 검출 예

그림 7과 8의 검출 결과를 보면 비교적 정확하게 음성구간을 검출해 낸 것을 볼 수 있다. 하지만, 방송뉴스 프로그램에서 기자나 인터뷰 데이터에는 다양한 종류의 배경잡음이 존재하기 때문에, 이로 인해 오검출이 발생하는 경우도 있었다. 실제 비디오 분할 시에는 의미를 가지는 문장단위의 분할이 더 타당한 것으로 보여진다. 그림 7에 나타나 있듯이 문장단위의 분할은 캡션정보에서 주어지는 문자열에서 문장의 첫 어절과 끝 어절 검출하여 음성신호에서 찾음으로써 할 수 있다.

Ⅳ. 결 론

본 연구에서는 캡션정보 및 음성인식기술을 이용하여 비디오 데이터를 분할, 검색하는 방법을 제안하고, 검출 실험을 통해 그 타당성을 확인하였다.

앵커의 경우와 같이 음성 데이터에 배경잡음이 없는 경우에 검출이 거의 정확하게 되었으나, 기자나 인터뷰 데이터에서는 다양한 배경 잡음의 존재로 검출 성능이 나빠졌다. 향후, 다양한 배경잡음의 처리방법에 대한 연구와 함께 음성인식기술을 이용한 비디오 색인 및 검색 시스템 구현에 대한 연구가 필요하다.

본 연구는 한국전자통신연구원 방송기술연구부의 지원으로 수행되었습니다. 지원에 감사드립니다.

참 고 문 헌

[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, Seech, 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] John Choi, Don Hindle, Julia Hirschberg, Ivan Magrin-Chagnolleau, Christine Nakatani, Fernando Pereira, Amit Singhal and Steve Whittaker, "An Overview of the AT&T Spoken Document Retrieval System," *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[4] Deb Roy and Carl Malamud, "Speaker Identification based Text to Audio Alignment for Audio Retrieval System," *Int. Conf. on Acoustics, Seech, and Signal Processing*, vol. Ⅱ, pp. 1099-1102, 1997.

[5] Ivan Magrin-Chagnolleau, Aaron E. Rosenberg and S.Parthasarathy, "Detection of Target Speakers in Audio Databases," *Int. Conf. on Acoustics, Seech, and Signal Processing*, vol. Ⅱ, pp. 821-824, 1999.

[6] mosur K. Ravishankar, "Efficient algorithms for Speech Recognition," Ph.D. thesis, CMU, 1996.