비디오의 오디오 정보 요약 기법에 관한 연구

Investigating the Efficient Method for Constructing Audio Surrogates of Digital Video Data

김현희(Hyun-Hee Kim)*

초 록

본 연구는 비디오의 오디오 정보를 추출하여 자동으로 요약하는 알고리즘을 설계하고, 제안된 알고리즘에 의해서 구성한 오디오 요약의 품질을 평가하여 효율적인 비디오 요약의 구현 방안을 제안하였다. 구체적인 연구 결과를 살펴보면 다음과 같다. 먼저, 제안 오디오 요약의 품질이 위치 기반 오디오 요약의 품질 보다 내재적 평가에서 더 우수하게 나타났다. 이용자 평가(외재적 평가)의 요약문 정확도에서는 제안 요약문이 위치 기반 요약문 보다 더 우수한 것으로 나타났지만, 항목 선택에서는 이 두 요약문간의 성능 차이는 없는 것으로 나타났다. 이외에 비디오 브라우징을 위한 오디오 요약에 대한 이용자 만족도를 조사하였다. 끝으로 이러한 조사 결과를 기초로 하여 제안된 오디오 요약 기법을 인터넷이나 디지털 도서관에 활용하는 방안들을 제시하였다.

ABSTRACT

The study proposed the algorithm for automatically summarizing the audio information from a video and then conducted an experiment for the evaluation of the audio extraction that was constructed based on the proposed algorithm. The research results showed that first, the recall and precision rates of the proposed method for audio summarization were higher than those of the mechanical method by which audio extraction was constructed based on the sentence location. Second, the proposed method outperformed the mechanical method in summary making tasks, although in the gist recognition task(multiple choice), there is no statistically difference between the proposed and mechanical methods. In addition, the study conducted the participants' satisfaction survey regarding the use of audio extraction for video browsing and also discussed the practical implications of the proposed method in Internet and digital library environments.

키워드: 비디오 요약, 오디오 요약, 영상 요약, 텍스트 초록, 멀티미디어 기반 요약, 텍스트 기반 요약, 소셜 메타데이터
tag, annotation, digital library

* 명지대학교 인문대학 문헌정보학과 교수(kimhh@mju.ac.kr)

■ 논문접수일자 : 2009년 8월 12일 ■ 최초심사일자 : 2009년 8월 14일 ■ 게재확정일자 : 2009년 8월 27일
■ 정보관리학회지, 26(3): 169-188, 2009. [DOI:10.3743/KOSIM.2009.26.3.169]

1. 서 론

1.1 연구의 배경과 목적

방송 및 정보 환경의 디지털화로 인하여 영상 정보를 주로 다루는 방송사 자료실, 한국영상자료원에서는 물론 일반 디지털 도서관에서 현재 가장 중요한 문제로 대두되는 것은 디지털 아카이빙으로 볼 수 있다. 디지털 아카이빙은 특히 방송 자료가 디지털 형태로 제작되고 방송사에서 최근 6개월분의 자료를 소장하고 있도록 하는 의무 조항을 가지고 있어 핵심적인 주제로 떠오르고 있다.

이러한 방송영상물은 디지털문화컨텐츠의 근간을 이룬다. 특히, 디지털방송이 2010년 이후 본격화되므로, 디지털컨텐츠의 전체산업을 주도할 것으로 전망한다. 하나의 콘텐츠가 다양한 매체와 포맷으로 제작되어 활용되는 원소스 멀티유즈(One Source Multi Use)개념에 따라 생성되는 다양한 형태의 영상물 생산 관행은 궁극적으로는 모든 형태의 영상물(방송영상, 영화영상, 광고영상, 게임영상, 인터넷상의 멀티미디어영상 등) 통합정보시스템 운영을 지향하게 할 것이다.

현재 대부분의 지상파 및 케이블 방송사는 보도영상자료(예, 아날로그 영상의 소스 테이프, 디지털영상의 디지털 파일)의 검색과 활용을 염두에 두고 관리하기 보다는 자료의 보존에 초점을 맞춰서 단순 집적 형태를 취해 오고 있다. 이러한 문제점을 해결하기 위해서 영상물의 기획단계에서부터 유사 영상물의 존재여부와 저작권법상의 유사성 저촉여부에 대한 전문가적인 판단은 물론 최종 이용자나 구매자의 자료에 대한 적합성 판정을 위해서 체계적인 영상 자료의 브라우징 방법이 요구된다. 더욱이 비디오는 의미를 결정하는 다양한 특성(예, 오디오 또는 영상 채널)을 갖고 있는 대용량 자원이기 때문에 비디오 전체 클립을 보기 전에 적합성 판정을 위한 좀 더 세밀한 브라우징 과정이 필요하다(Kristin et al. 2006).

현재 방송사 자료실 또는 디지털 도서관은 영상물 자료를 메타데이터(예, 표제)나 비디오의 스피치 내용 및 자막 등과 매칭시키는 키워드 검색을 이용한다. 이외에 이미지 프로세싱 기법을 적용한 키프레임 이미지 매칭 또는 프레임에 나타난 객체 인식 기법 등을 활용하여 비디오 자료를 일차적으로 검색한다(Smeaton and Browne 2006; Smeaton 2007). 이러한 일차 검색의 결과로 제시된 잠정적으로 관련이 있다고 판단되는 비디오 자료의 그룹에서 영상물 제작자나 판매자가 기술한 텍스트 기반 메타데이터나 요약문 또는 홍보 자료를 사용하거나 최근에는 키프레임들로 구성된 영상 초록(스토리보드 또는 슬라이드쇼), 키프레임 하나로 구성된 포스터 프레임(poster frame), 빨리보기(fast forwards) 등을 통해서 최종적으로 적합한 자료를 선정하고 있다.

비디오 요약은 크게 두 가지 즉, 스토리보드와 같은 정적인 요약과 비디오 스킴[1]과 같은 동적인 요약이 있다. 비디오와 오디오 정보를

1) 비디오 스킴은 정지된 상태에서 보는 것이 아니라 플레이시킨다는 개념에 가까우며, 예를 들어서 20분 정도의 영상과 오디오 정보를 통합하여 2분 정도의 요약으로 만드는 것이다. 비디오 스킴은 영화 예고편처럼 이용자의 주의를 끌게끔 의도된 긴 비디오의 일부를 짧게 요약한 것이라고 할 수 있다.

통합하여 긴 비디오의 일부를 짧게 요약한 비디오 스킴은 비디오 요약의 가장 이상적인 방법이지만 이를 구현하는데 아직까지 기술적인 어려움과 비용이 많이 드는 단점이 있다. 또 다른 비디오 요약 방법으로 최근 많이 사용하고 있는 오디오가 배제된 영상 초록이 있는데 영상 초록은 이미지만으로 의미를 전달하는데 한계가 있다. 자연 다큐멘터리, 강의 또는 연설 비디오 자료와 같이 오디오를 통해서 주로 정보를 전달하는 비디오들은 이미지로 구성된 영상 초록만으로 비디오 내용을 파악하는 것이 더욱 어렵다. 더욱이 오디오 정보가 중복된 정보를 제공함으로써 영상 정보를 보완한다는 보고도 있고(Gunther, Kazman and MaccGregor 2004), 오디오 정보가 영상 정보 보다 정보를 이해하고 기억하는데 더 효율적임을 증명하는 연구도 있다(Schmandt and Mullins 1995).

또한 핸드폰이나 PDA와 같은 소규모 디스플레이 장치를 갖는 기기에서는 영상 정보 보다 오디오 정보가 더 편리하게 이용될 것이다. 최근에 많은 대학들이 강의 자료를 비디오 또는 오디오 파일로 일반에게 공개하고 있는데 특히 예일대는 오픈 예일 코스(Open Yale Courses) 프로젝트를 통해서 강의 자료를 오디오, 비디오 등 다양한 파일로 제공하고 있는 추세이다. 이러한 정보 환경의 변화에 따라서 오디오 요약의 필요성이 그 어느 때 보다 높아지고 있다고 생각된다.

본 연구의 목적은 비디오의 오디오 요약을 자동으로 추출하기 위한 알고리즘을 제안하고, 제안된 알고리즘에 의해서 구성한 오디오 요약의 품질을 표준 요약과 비교하여 측정하는 내재적 평가와 이용자의 의미 분석의 정확도에 따라 측정하는 외재적 평가를 수행하여 효율적인 오디오 요약 방안을 제안하는데 있다.

1.2 연구 문제와 방법

본 연구에서 조사하고자 하는 문제는 다음과 같다. 첫째, 텍스트 요약에 적용된 이론과 방법들이 오디오 요약에도 그대로 적용될 수 있을 것인가? 둘째, 본 연구에서 제안된 알고리즘에 의해서 구성된 오디오 요약의 품질이 위치 기반 오디오 요약의 품질 보다 내재적 평가(재현율과 정확률)에서 더 우수하게 나타날 것인가? 셋째, 이러한 내재적 평가 결과가 외재적 평가에도 영향을 미칠 것인가? 즉, 제안 오디오 요약이 위치 기반 오디오 요약 보다 외재적 평가(이용자의 비디오 의미 파악의 정확도)에서 더 우수하게 나타날 것인가? 넷째, 이용자 관점에서 본 오디오 요약에 대한 만족도는 어떠한가이다.

이러한 네 가지 연구 문제들을 조사하기 위해서 Money와 Agius(2008, 2009), Kristin et al.(2006) 등의 선행 연구와 국내외 방송사의 영상물 관리 현황을 조사, 분석하여 비디오 요약에 대한 개념적 모형을 구성한다. 오디오 요약를 구성하기 위한 표본 비디오 자료의 장르는 음성으로 많은 정보를 표현하는 교육 및 연설 비디오로 정하고 유투브 사이트에서 8개의 비디오들을 선정한다. 그런 다음 8개의 표본 비디오의 오디오 내용을 분석한 후 본 연구에서 제안한 알고리즘을 이용하여 오디오 요약을 구성한다.

오디오 요약의 효율성을 평가하기 위해서 다음의 두 가지 방법을 이용한다. 즉, 오디오 요약

의 품질 평가를 통해서 요약 기법의 성능을 평가하는 내재적 평가와 이용자들로 하여금 오디오 요약을 통해서 전체 비디오가 표출하고자 하는 의미를 얼마나 정확하게 파악했는지를 실험을 통해서 파악해 봄으로써 요약 기법의 성능을 평가하는 외재적 평가를 한다(정영미 2005). 내재적 평가를 위해서 연구팀은 오디오 정보에서 비디오의 의미를 가장 잘 나타내는 문장들을 추출하여 표본 요약을 구성한다. 외재적 평가를 위해서 제안된 오디오 요약과 위치 정보에 기초하여 자동으로 구성한 오디오 요약간의 성능을 비교, 분석한다. 이 두 오디오 요약 기법의 성능을 비교하기 위해서 실험 시스템을 구현하고, 대학원생과 학부생으로 구성된 20명의 피조사자를 이용한다. 평가 결과의 통계 분석을 위해서 SPSS 통계 패키지를 사용한다.

2. 선행 연구

선행 연구로 비디오 요약과 텍스트 요약 이론에 연구들을 살펴본다.

2.1 비디오 요약

Song과 Marchionini(2007)는 오디오 초록과 영상 초록의 적합성 판정을 위한 효용성을 테스트해 보기 위해서 13개의 비디오 자료와 36명의 피조사자들을 이용하였다. 이들은 영상 초록과 오디오 초록을 결합한 영상/오디오 초록의 효과가 오디오 또는 영상 초록을 단독으로 사용했을 때 보다 더 효율적이라고 주장하고 있다. 이들은 오디오 초록이 영상 초록 보다 비디오 세그먼트에 대한 더 나은 이해를 가져다 주지만 이용자들은 영상 초록을 오디오 초록 보다 더 선호한다고 기술하였다. 아울러 이들 연구에서 이용자들은 오디오 초록을 통해서 비디오 내용을 확인하고 추가적인 가치를 얻는 것으로 나타났으며, 이러한 사실은 Over et al.(2005) 및 Yang과 Marchionini(2005)의 연구에서도 이미 확인되었다. Song과 Marchionini가 제안한 오디오 초록은 Open Video Digital Library 연구원들에 의해서 작성한 텍스트 초록을 음성 합성기를 이용하여 오디오로 변환하여 사용하고 있다. 이러한 방법은 오디오 초록 작성의 단가를 낮추는 방법이긴 하지만 오디오 내용과 키프레임에서 나타내는 내용과 일치하지 않는 경우가 많아서 비디오 내용을 파악하는데 용이하지 않는 문제점이 있을 것으로 생각된다.

Furini와 Ghini(2006)는 비디오에서 사운드가 없는 부분이 상당히 많은 비중을 차지하는 것을 인지하고 전체 비디오 클립을 사운드가 없는 부분과 사운드가 있는 부분으로 구분한다. 그런 다음 사운드가 있는 비디오 프레임들만을 모두 선정하거나 사운드가 있는 부분은 물론 사운드가 없는 부분에서도 무작위로 프레임들을 추출한 후 선정된 키프레임들과 사운드를 재코딩한 후 비디오 스킴을 구성할 수 있는 3개의 서로 다른 휴리스틱 알고리즘을 제안하였다. 영화, TV 뉴스, TV 쇼, 축구, 토크쇼의 서로 다른 5개의 장르의 비디오들을 제안한 세 개의 알고리즘을 이용하여 비디오 요약을 하였는데 최대 50%까지 축약할 수 있었다. 비디오 요약에 대한 만족도를 30명의 피조사자에게 5점 척도를 사용하여 조사하였다. 조사 결과 세 개의

알고리즘에 의해서 구성된 TV 뉴스 요약의 평균 만족도(4.5)가 가장 높게 나타났으며, 영화의 평균 만족도(2.8)가 가장 낮게 나왔다. 이 연구는 전반적으로 비디오 요약이 비디오 전체 클립의 내용을 어느 정도 대체할 수 있다는 결론을 얻었다. 그러나 이들이 사용한 영상 및 오디오 요약은 최대 전체 클립의 50%만을 축약할 수 있기 때문에 적합성 판정을 위한 브라우징에는 사용할 수 없을 정도로 비디오 요약이 너무 긴 것이 문제이다.

Money와 Agius(2008)는 비디오 요약에 대한 개념적인 틀을 제안하였다. 그들이 제안한 개념적인 틀은 비디오 요약 기술(비디오 스트림의 요약을 얻기 위해서 비디오 내용 처리에 사용된 방법)과 비디오 요약(비디오 요약 기술의 결과물)으로 구분된다. 비디오 요약에 사용되는 기술을 세 가지 범주 즉, 첫째, 비디오 스트림에 있는 이미지, 오디오 또는 텍스트와 같은 내적 정보를 비디오 요약에 이용하는 방법, 둘째, 비디오 스트림에 있지 않는 이용자 이용 행태와 관련된 이용자 기반 정보, 문맥 정보 등과 같은 외적 정보를 이용하는 방법, 셋째, 내적 및 외적 정보를 모두 이용하는 하이브리드 방식으로 나눌 수 있다. 또한 비디오의 내용 유형을 객체, 이벤트, 자각 및 특징 기반으로 구분하기도 하며 비디오 유형을 상호작용, 개인화로 구분하기도 한다. 이 연구는 비디오 요약은 시맨틱갭과 같은 오래된 도전을 극복하고 각 개인에게 더 적합한 관련성을 갖는 비디오 요약을 제공할 수 있는 외적 정보 즉, 이용자 기반 정보 등을 비디오 요약에 도움이 되는 핵심 정보라고 제안하고 있다.

이들은 이러한 비디오 요약에 대한 개념적인 틀을 구성하는 것 외에 감정적인 비디오 요약을 위해서 피부의 전기적 반응, 호흡 진폭, 호흡 횟수, 혈액양 펄스, 심박수의 5개의 기준을 활용하여 이용자의 생리학적 반응을 분석하여 이용하였다(Money and Agius 2009). 실험 결과, 호러(horror) 콘텐츠는 전기적 반응, 호흡 진폭, 호흡 횟수 및 혈액양 펄스에서 중요한 반응을 이끌어냈다. 코메디 콘텐츠는 전기적 반응에서는 상대적으로 낮은 수준의 반응을 이끌어 냈지만, 나머지 네 개의 기준에서는 중요한 반응을 이끌어 냈다. 한편, 드라마 콘텐츠는 상대적으로 낮은 수준의 생리학적 반응을 보였고, 과학 소설과 액션 콘텐츠는 중요한 전기적 반응을 이끌어 내는 것으로 나타났다. 끝으로 이들은 이러한 실험 결과를 비디오 요약에 적용할 수 있는 방안들에 대해서 언급하고 있다.

Kristin et al.(2006)은 디지털 비디오의 오디오 요약물을 위한 설계안을 제안하였다. 이들은 오디오 요약물의 다섯가지 유형(예, 오디오를 위한 영상 요약, 오디오 요약 등)을 정의하고, 각 유형의 장단점을 이용자 노력의 정도, 비용 및 전반적인 압축 비율의 측면에서 기술하고 있다.

국내 연구를 살펴보면, 김재곤 등(2000)은 비디오 요약에는 영상 요약을 구성하는 정적인 요약과 비디오 스킴과 같은 동적인 요약이 있다고 기술하고 있다. 이들은 비디오 스킴은 전체 비디오 내용을 효과적으로 표현할 수 있는 주요 구간들을 선별하는 방식의 요약으로 오디오를 함께 포함하므로 이용자의 이해도에 있어서 더 유리한 요약이라고 제안하였다. 진성호 등(2005)은 개인화된 비디오 요약 서비스를 제

공하기 위해, TV-Anytime에 기반한 사용자 선호도 정보를 이용하는 시스템 스킴(scheme)을 제안하였다. 제안한 방송 서비스의 유효성을 테스트하기 위해 영화 장르의 비디오에 대해서 사건(event)들을 분할하고, 해당 선호도 정보에서 추론된 선호 사건 정보에 따라 비디오 요약을 생성하고, 시청자 단말기에 제공하는 실험을 수행하였다.

끝으로, 이러한 비디오 요약 이론과 방법을 기초로 하여 실험 시스템을 구현한 사례들을 살펴보면 다음과 같다. 비디오 스킴 연구 중 가장 알려진 연구에는 미국의 카네기 멜론 대학에서 개발한 비디오 도서관 시스템인 인포미디어(Informedia) I과 II가 있다. 인포미디어는 비디오내의 텍스트 정보, 객체 인식, 오디오의 내용 인식 등을 통해서 고수준의 특정값을 사용한 비디오 색인 기법을 사용함으로써 데이터베이스 구축과 검색의 효율을 높였다. 즉, 음성으로 질의를 하면 관련 자료들의 비디오 내용을 요약하여 보여주는 비디오 스킴을 보여 주며 이러한 메타데이터들을 통해서 찾고자하는 비디오인가를 판단하게 된다(Hauptmann 2005; Witbrock and Hauptmann 1998).

더블린 시립 대학에서 개발한 Fischlar 비디오 도서관 시스템은 오디오 정보를 비디오 자료를 브라우징하는데 사용하고 있다. 이 시스템은 폐쇄 자막에서 분석한 텍스트 다이아로그를 전체 TV 뉴스를 개별적인 뉴스 스토리로 구분하는데 사용하고 이는 브라우징에도 활용한다. 즉, 이용자들은 브라우징을 위해서 구두 다이아로그에서 텍스트 검색을 수행함으로써 관련 뉴스 스토리를 검색할 수 있다(Smeaton 2007). 일반적으로 비디오 요약을 위해서 스토리보드를 사용하는 시스템들이 많은 편이다. 예를 들면, Open Video 프로젝트(http://www.open-video.org/)가 있다. 이 시스템에서의 비디오 요약 방법은 스토리보드와 처음 7초 듣기(7-second excerpt) 등을 사용하고 있다(Marchionini, Wildemuth and Geisler 2006). 이외에 인터넷 아카이브(http://www.archive.org/index.php)는 비디오자료의 스토리보드를 브라우징 자료로 사용하고 있다(Smeaton 2007).

2.2 텍스트 요약

Sparck Jones(2007)는 지난 10년 동안 발전해 온 자동 텍스트 요약 이론 및 기법 그리고 실험 시스템을 소개하고 있다. 이 연구는 최근 텍스트 요약 연구에서 탐구된 입력, 목적 및 출력 요인들을 살펴보고 그리고 요약에 적용된 평가 전략들을 논의하고 있다. 텍스트 요약에는 담화 구조 또는 지식기반 요약과 같은 정교한 요약 기법들이 있는데 이러한 기법들은 방대한 양의 비디오 자료에 적용하기에는 그 실효성이 낮다고 판단되어 포함시키지 않았다.

다음은 방대한 비디오 자료에 비교적 쉽게 적용할 수 있는 문장 추출에 의한 텍스트 요약에 대한 연구들을 살펴본다. 초기 연구로 Luhn(1958)은 문장의 중요도는 문장 내 각 단어의 중요도와 단어의 상대적 위치에 의해서 결정된다고 보았다. 따라서 단어의 출현빈도에 의해서 핵심어를 선정하고 핵심어들이 집중적으로 출현한 문장을 선택하여 요약문을 작성하였다.

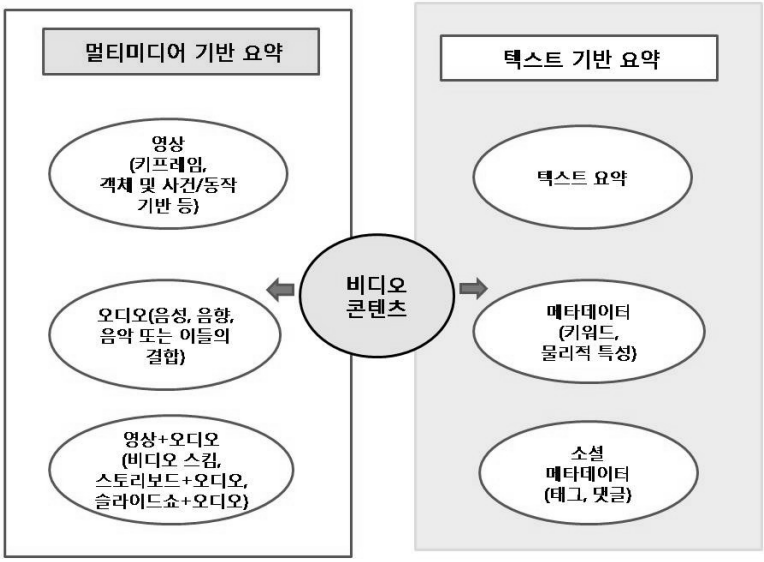
Edmunson(1969)은 문장의 중요도를 측정하는 네 가지 기준 즉, 단서어, 주요어, 표제어 및

문장 위치를 사용하였다. Kupiec et al.(1995)과 Myaeng과 Jang(1999)은 기계학습을 통한 요약 기법을 제안하였다. Kupiec 등이 문장 추출 기준으로 사용된 자질에는 문장 길이(5단어 이하의 문장은 선정하지 않음), 고정 단어구(예, in summary), 문단 내 문장 위치, 주제어 및 대문자 단어를 사용하였다. 한편 맹성현과 장동현은 문장 추출 기준으로 단서어, 문장 위치, 주제어, 중심성 및 표제어를 사용하였다. 이와 같이 살펴본 선행 연구에서 '위치' 그리고 '표제어와의 유사도'가 중요한 자질로 사용되고 있다.

3. 비디오 요약에 대한 개념적 모형

Money와 Agius(2008, 2009), Kristin et al.(2006) 등의 선행 연구들과 국내외 영상물 관리 시스템을 조사, 분석하여 비디오 요약에 대한 개념적 모형을 구성하였다. 〈그림 1〉은 비디오 요약의 개념적 모형이다.

비디오 요약을 멀티미디어 기반 및 텍스트 기반 요약으로 구분하였다. 멀티미디어 기반 요약은 비디오의 실제 내용물을 요약한 것으로 영상(또는 이미지) 요약, 오디오 요약 그리고 영상/오디오 요약이 있다. 영상 요약에는 기본적으로 비디오의 의미를 나타내는 키프레임(이미지)들을 샷이나 장면 전환에 준하여 추출하여 구성한 스토리보드 또는 슬라이드쇼가 있다. 이외에 키프레임을 추출하는 기준을 핵심적인 객체, 사건, 동작 또는 자막에 두고 이를 표현하거나 포함한 프레임들을 추출하여 구성한 스토리보드 또는 슬라이드쇼가 있다. 오디오 요약에는 음성, 음향 또는 음악 요약 또는 이들을 결합한 요약이 있다. 끝으로 영상/오디오 요약에는 비디오 전체 내용을 압축한 비디오 스킴



〈그림 1〉 비디오 요약의 개념적 모형

그리고 스토리보드(또는 슬라이드쇼)와 오디오를 동시적 또는 비동시적으로 결합한 요약이 이 범주에 속한다.

텍스트 기반 요약에는 전통적인 초록으로 알려진 텍스트 요약, 키워드 및 물리적 특성(비디오의 재생시간, 파일 크기 및 비디오 포맷 등) 등을 포함한 전통적인 메타데이터 그리고 비디오에 실제 이용자들이 코딩한 태그, 댓글 및 비디오평과 같은 소셜 메타데이터가 있다.

4. 오디오 요약 설계

4.1 표본 비디오

오디오 요약을 구성하기 위해서 표본 비디오 자료의 장르는 음성으로 많은 정보를 표현하는 교육 및 연설 비디오로 정하고 종수는 8개로 선정하였다(표 1 참조). 비디오의 재생 시간은 4~23분이다. 8개 비디오 중 4개는 4~6분 사이이고 나머지 4개는 10~23분 사이이다. 표본 비디오는 유투브 사이트(www.youtube.com)에서 선정하였다.

영어 비디오들을 표본 비디오로 선정한 이유는 태그와 댓글 등의 소셜 메타데이터를 포함하고 있으면서 피조사자들에게 비디오 내용이 비교적 많이 알려지지 않았기 때문이다. 또한 본 연구에서 제안한 오디오 요약 알고리즘은 오디오의 언어에 관계없이 모두 적용될 수 있기 때문이다.

4.2 제안된 알고리즘에 의한 오디오 요약 구성

4.2.1 알고리즘 설계

오디오 정보는 음성, 음향 및 음악 정보를 포함하여 오디오 정보를 정의하는 것이 일반적이나 본 연구에서 구현된 오디오 요약은 음성 정보에 초점을 맞춘다. 오디오 요약 알고리즘을 구성하기 위해서 선행 연구에서 기술한 기존의 텍스트 요약 이론들을 활용하였으나 이러한 텍스트 요약 이론은 논문 형식에 맞춘 것이라 비디오의 오디오를 축약하는 데에 그대로 적용하기에 무리가 따랐다. 따라서 내용 일부를 수정하였다.

〈표 1〉 표본 비디오

번호	표제	재생 시간 (분 : 초)	번호	표제	재생 시간 (분 : 초)
1	Washtenaw Community College School of Culinary Arts	04:41	5	Learning English-Lesson Forty Three(Superstition)	10:41
2	Kate Lundy: What I do for Open Government	05:01	6	Disability Services at ASU Libraries	10:05
3	Meet the Mentor	05:09	7	Steve Jobs' 2005 Stanford Commencement Address	15:04
4	Beating Holiday Stress	04:13	8	President Clinton 1997 Inaugural Address	22:57

다음 공식(W)은 각 문장의 가중치를 구하기 위해서 본 연구에서 제안한 것이다. 공식(W)은 위치와 표제와의 유사도라는 두 개의 자질을 사용하고 이외에 각 비디오에 이용자들이 접근점으로 할당하는 태그와의 유사도를 새롭게 포함시켜 총 3개의 자질(위치, 표제어, 태그/댓글)을 사용하였다. 만약 태그수가 4개 이하인 경우에는 댓글을 분석하여 댓글에 자주 출현하는 단어를 태그로 간주하여 태그수가 최소 5개가 되도록 하였다. 이외에 텍스트 요약 이론에서는 서론이나 결론 등에 나타나는 문장을 주제 문장으로 보았으나, 비디오의 오디오 자료는 끝 부분 보다는 도입부 부분에서 주제 문장을 더 많이 포함하는 경향이 있었다. 따라서 첫 6개 문장의 가중치는 '1'로 하였고 나머지는 모두 '0.8'로 처리하였다. 이외에 문장에 출현하는 표제어와 태그의 가중치는 각각 '1'과 '0.7'로 하였다. 제안된 알고리즘의 특성은 오디오 내용을 압축하기 위해서 위치 정보, 전통적인 메타데이터인 표제 외에 소셜 메타데이터인 태그와 댓글 정보를 이용하였다는 점이다.

W = α +(ti_num * 1)+(tag_num * 0.7)

W = 총 가중치

α = 문장 위치 가중치(첫 6개 문장의 가중치 -〉 1, 기타 -〉 0.8)

ti_num = 해당 문장에 포함된 비디오 표제어 수(비디오의 표제, 부제 및 소제목 등에 출현한 단어들 가운데 불용어를 제외한 단어수)

tag_num = 해당 문장에 포함된 비디오에 부여된 태그수(또는 댓글의 키워드 수)

4.2.2 오디오 요약 절차

제안된 알고리즘에 의한 오디오 요약 절차는 다음과 같다.

첫째, 각 표본 비디오의 오디오는 스크립트를 이용하여 영문 텍스트로 변환하였다. 변환된 텍스트를 마침표나 물음표와 같은 구두점을 기준으로 하여 구분하였다. 그런 다음 각 문장에 순서대로 번호를 매겼다.

둘째, 각 문장에 위에서 기술한 공식(W)을 이용하여 가중치를 계산하여 할당하였다. 예를 들어서 특정 문장이 3번째 문장(가중치: 1)이며 2개의 표제어가 포함되어 있고(가중치: 2 × 1=2), 1개의 태그가 포함되어 있다면(가중치: 1 × 0.7=0.7), 총 가중치(W)는 '3.7'이 된다.

셋째, 문장의 가중치가 정해지면 가중치가 높은 순서대로 비디오의 재생시간에 따라서 4~12개의 문장을 선정하였다. 단 2단어 이하로 구성된 짧은 문장은 선정하지 않았다. 기준치 이상의 중요도를 갖는 문장들을 선택한 후 중복된 문장이 있으면 이를 제거하였다. 중복된 문장을 제거하기 위하여 두 문장 벡터 간 유사도를 산출하여 유사도 값이 기준치를 넘으면 두 문장 중 가중치가 높은 것을 선택하였다. 최종적으로 선택된 문장들을 텍스트에 출현 순서대로 요약문을 구성한 후 한글로 번역하고 이를 음성 합성기인 매직 잉글리쉬 플러스를 이용하여 오디오 파일로 변환하여 오디오 요약을 구현하였다.

4.2.3 오디오 요약 예

다음은 〈표 1〉의 일곱 번째 비디오인 스티브 잡스의 연설 내용을 제안한 알고리즘에 의해서 요약하는 과정을 설명한다. 각 문장에 가중치를 부여한 후 1.7 이상인 문장들을 추출한 결과

가 〈표 2〉에 나타나 있다.

〈표 3〉은 제안 요약문과 표본 요약문을 기술하고 있다. 제안 요약문에서 밑줄친 문장들은 표본 요약문과 일치하는 문장들이다. 제안된 요약문의 재현율과 정확률은 각각 0.31(4/13), 0.67(4/6)이다(표 5 참조).

〈표 2〉 추출된 문장과 가중치

〈비디오 정보〉		
표제	Steve Jobs' 2005 Stanford Commencement Address	
태그	apple graduation education NeXT Pixar cancer computer Steve Jobs stanford address speech keynote commencement	
문장수	142개	
〈추출된 문장과 가중치〉		
문장 번호	내 용	가중치
1	I am honored to be with you today at your commencement from one of the finest universities in the world.	1(위치)+1 (표제어, commencement)=2
2	I never graduated from college.	1(위치)+0.7 (태그: graduated)=1.7
3	Truth be told, this is the closest I've ever gotten to a college graduation.	1(위치)+0.7 (태그: graduated)=1.7
4	During the next five years, I started a company named NeXT, another company named Pixar, and fell in love with an amazing woman who would become my wife.	0.8(위치)+0.7 * 2 (태그: NeXT, Pixar)=2.2
5	Pixar went on to create the worlds first computer animated feature film, *Toy Story*, and is now the most successful animation studio in the world.	0.8(위치)+0.7 * 2 (태그: Pixar, computer)=2.2
6	In a remarkable turn of events, Apple bought NeXT, I returned to Apple, and the technology we developed at NeXT is at the heart of Apple's current renaissance.	0.8(위치)+0.7 * 5 (태그: apple, NeXT)=4.3

〈표 3〉 제안 요약문과 표본 요약문

제안된 알고리즘에 의한 요약문(전체 문장수: 6개, 적합 문장수: 4개)(315자)
1) 먼저 세계 최고의 명문으로 꼽히는 이 곳에서 여러분들의 졸업식에 참석하게 된 것을 영광으로 생각합니다. 2) 사실 저는 대학을 졸업하지 못했습니다. 3) 태어나서 대학교 졸업식을 이렇게 가까이서 보는 것은 처음이네요. 4) 이후 5년 동안 저는 '넥스트'와 '픽사'를 만들고 지금의 아내와 사랑에 빠졌습니다. 5) 픽사는 세계 최초의 3D 애니메이션 토이스토리를 시작으로 지금은 세계에서 가장 성공한 애니메이션 제작사가 되었습니다. 6) 세기의 사건으로 평가되는 애플의 넥스트 인수와 저의 애플로 복귀 후 넥스트 시절 개발했던 기술들은 현재 애플의 르네상스의 중추적인 역할을 하고 있습니다.
표본 요약문(총문장수: 13개)
1) 먼저 세계 최고의 명문으로 꼽히는 이 곳에서 여러분들의 졸업식에 참석하게 된 것을 영광으로 생각합니다. 2) 저는 오늘 여러분께 제 인생의 세 가지 이야기를 해볼까 합니다. 3) 먼저 인생의 전환점에 관한 이야기입니다. 4) 배짱, 운명, 인생, 카르마(업) 등 그 무엇이든 믿음을 가져야만 합니다. 5) 두 번 째 이야기는 사랑과 상실에 관한 것입니다. 6) 이후 5년 동안 저는 '넥스트'와 '픽사'를 만들고 지금의 아내와 사랑에 빠졌습니다. 7) 세기의 사건으로 평가되는 애플의 넥스트 인수와 저의 애플로 복귀 후 넥스트 시절 개발했던 기술들은 현재 애플의 르네상스의 중추적인 역할을 하고 있습니다. 8) 세 번 째는 죽음에 관한 것입니다. 9) 그리고 가장 중요한 것은 마음과 영감을 따르는 용기를 가지는 것입니다. 10) 이미 마음과 영감은 당신이 진짜로 무엇을 원하는지 알고 있습니다. 11) 나머지 것들은 부차적인 것이죠. 12) 사진 밑에 '계속 갈망하라 여전히 우직하게'라는 메시지가 있었습니다. 13) 지금, 새로운 시작을 위해 졸업을 하는 여러분에게 동일한 바람을 가집니다.

4.3 위치 정보에 기반한 오디오 요약 구성

4.3.1 오디오 요약 절차

제안된 오디오 요약과의 성능 비교를 위해서 위치 정보에 기초한 오디오 요약을 구성하였다. 앞의 경우처럼 오디오 내용을 스크립트를 이용하여 영문 텍스트로 변환하였다. 변환된 텍스트를 마침표나 물음표와 같은 구두점을 기준으로 하여 구분한 다음 각 문장에 순서대로 번호를 매겼다. 그런 다음 비디오의 전체 내용을 포함시키기 위해서 비디오 오디오의 앞부분과 끝부분의 문장들을 동일한 비율로 추출하였다. 요약문의 전체 길이는 앞의 제안 요약문의 길이에 준해서 구성하였다.

4.3.2 오디오 요약 예

이와 같이 위치 정보만을 기준으로 추출한 〈표 4〉의 요약문은 14개 문장으로 구성되어 있다. 처음 6개 문장들(1~6)은 오디오의 앞부분에서 추출한 것이고 나머지 8개 문장들(7~14)은 오디오의 뒷부분에서 추출한 것이다. 문장들이 선택되면 텍스트에 출현 순서대로 요약문을 구성한 후 한글로 번역하고 음성 합성기를 이용하여 오디오 파일로 변환하였다. 자동 요약문의 전체 문장수는 14개이고 적합 문장수는 8번과 9번이 하나의 문장으로 결합되어 4개가 되었다. 이 두 문장을 결합한 이유는 〈표 3〉의 표준 요약에서 확인할 수 있는 것처럼 '사진 밑에 '계속 갈망하라 여전히 우직하게'라는 메시지가 있습니다.' 문장 속에 이 두 문장의 내용이 들어 있기 때문이다. 이에 따라서 자동 요약문의 재현율과 정확률은 각각 0.31(4/13), 0.29(4/14)이다(표 5 참조).

5. 제안된 오디오 요약의 평가

제안된 오디오 요약의 효율성을 평가하기 위하여 두 가지 방법, 즉 내재적 평가와 외재적 평가를 수행하였고(정영미 2005), 평가 결과의 통계 분석을 위해서 SPSS 통계 패키지를 사용한다.

5.1 내재적 평가

오디오 요약의 품질을 평가하기 위해서 2명의 연구자가 각 비디오의 표준 요약문을 구성하였다. 비교를 용이하게 하기 위해서 표준 요약문은 기존 스크립트에 있는 중요 문장들을

〈표 4〉 위치 정보에 기반한 요약문

위치 정보에 기반한 요약문(전체 문장수: 14개, 적합 문장수: 4개)(316자)
1) 먼저 세계 최고의 명문으로 꼽히는 이 곳에서 여러분들의 졸업식에 참석하게 된 것을 영광으로 생각합니다. 2) 사실 저는 대학을 졸업하지 못했습니다. 3) 태어나서 대학교 졸업식을 이렇게 가까이서 보는 것은 처음이네요. 4) 저는 오늘 여러분께 제 인생의 세 가지 이야기를 해볼까 합니다. 5) 단지 그뿐입니다. 6) 세 가지 이야기 입니다. 7) 그것이 그들의 마지막 작별 인사였습니다. 8) 계속 갈망하라. 9) 여전히 우직하게. 10) 제 자신에게도 항상 그러하기를 바랐습니다. 11) 그리고 지금, 새로운 시작을 위해 졸업을 하는 여러분에게 동일한 바람을 가집니다. 12) 계속 갈망하라. 13) 여전히 우직하게. 14) 감사합니다.

추출하여 순서대로 구성하였다. 그런 다음 각 오디오 요약의 내용이 표준 요약문과의 유사성 측정을 요약 재현율(제안된 요약문 내 적합문장 수 / 표준 요약문의 요약 문장 총 수)과 요약 정확률(제안된 요약문 내 적합문장 수 / 제안된 요약문의 요약 문장 총 수)로 측정하였다.

내재적 평가 결과는 〈표 5〉에 기술하였다. 예상한 대로 제안 요약문이 위치 정보 기반 요약문과 비교하여 재현율이 더 높은 것으로 나타났고 t-검증 결과 통계적으로 유의미한 차이를 나타냈다(0.44 vs. 0.31, $p(=0.03) < 0.05$). 또한 정확률도 제안 요약문이 더 높은 것으로 나타났고 이 역시 통계적으로 유의미한 차이를 나타냈다(0.63 vs. 0.37, $p(=0.01) < 0.05$).

5.2 외재적 평가

앞의 내재적 평가 결과 제안 오디오 요약의 품질이 위치 기반 요약 보다 더 우수한 것으로 나타났다. 다음은 실제 비디오 이용자들을 대상으로 하여 두 종류의 오디오 요약이 각 비디오의 의미 파악에 어떤 효과가 있는지 파악하기 위해서 실험을 실시하였다. 다시 말해서 오디오 요약의 품질이 실제 이용자들의 비디오 의미 파악에 어떤 영향을 미치는지 파악하고자 했다. 다음은 실험에 투입된 피조사자, 실험 시스템 및 실험 절차에 대해서 설명한다.

5.2.1 피조사자

피조사자는 학부생(16명)과 대학원생(4명)으로 구성된 20명을 활용하였다. 20명의 피조사자를 두 개의 집단으로 고르게 나눈 다음 그룹 1(10명)과 그룹 2(10명)로 구분하였다.

5.2.2 실험 시스템 구현

8개의 표본 비디오에 대한 두 개의 실험용 비디오 인터페이스(그룹1과 그룹2)를 구현하였다. 즉, 각 비디오에 대한 두 가지 종류의 오디오 요약을 구현하였다. 그룹 1에게는 제안된 알고리즘에 의해서 구성한 오디오 요약을 듣게 한 후 각 비디오에 대한 요약문을 먼저 입력하게 한다. 그런 다음 짧은 문장으로 구성된 10개의 항목들을 제시한 후 비디오의 자세한 내용 즉, 사건, 사실 등을 정확하게 설명하고 있는 항

〈표 5〉 재현율과 정확률

비디오 번호	제안 요약문		위치정보 기반 요약문	
	재현율	정확률	재현율	정확률
1	0.80	0.67	0.60	0.43
2	0.29	0.67	0.29	0.67
3	0.80	1.00	0.40	0.40
4	0.50	1.00	0.33	0.50
5	0.33	0.33	0.17	0.13
6	0.38	0.38	0.25	0.19
7	0.31	0.67	0.31	0.29
8	0.14	0.33	0.14	0.33
평균(표준편차)	0.44(0.24)	0.63(0.27)	0.31(0.14)	0.37(0.17)

목(들)을 선택하도록 한다. 그룹 2에게는 위치 기반 비디오 요약을 듣게 한 후 그룹 1과 동일한 작업을 하도록 한다.

〈그림 2〉는 두 실험 시스템들의 초기화면으로 표본 비디오 8개의 표제 목록을 보여준다. 〈그림 2〉에서 비디오를 선택하면 〈그림 3〉이 출력된다. 이 화면에서 이름과 비디오의 요약에 대해서 기술하고 저장하면 〈그림 4〉가 출력된다. 이 화면에서는 10개 항목들이 출력되는데 각 항목이 비디오 내용을 맞게 기술한 것인지 아니면 틀리게 기술한 것인지를 체크한다. 그리고 비디오 8을 기술할 때에는 실험에 걸린 총 시간과 오디오 요약에 대한 피조사자의 생각도 함께 기술하도록 하였다. 이 모든 답변은 웹상에서 데이터베이스(액세스)에 입력되도록 하였다.

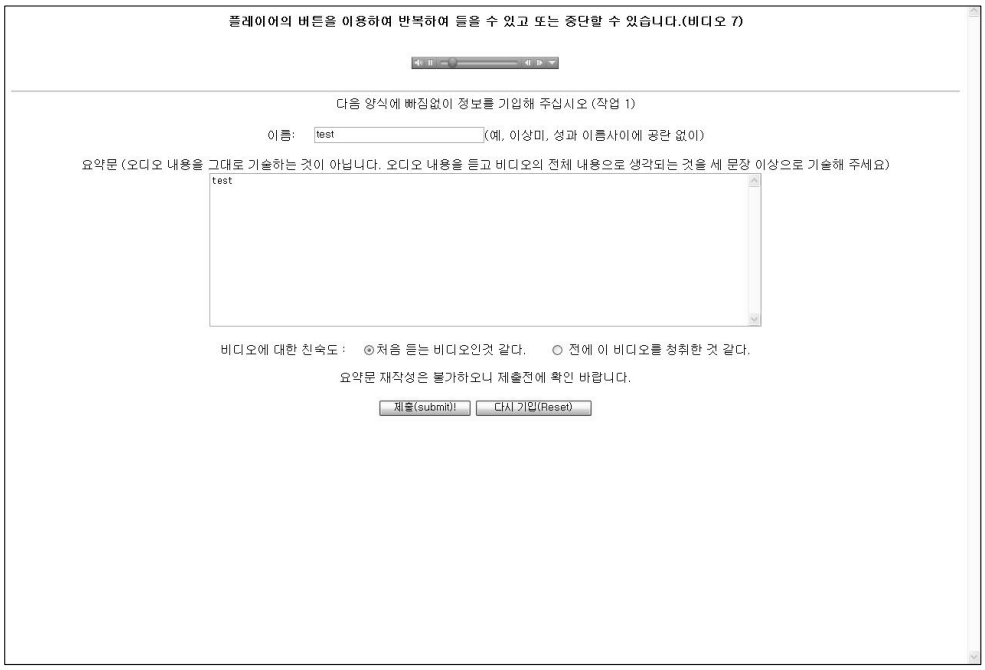
5.2.3 실험 절차

실험은 원격 테스트를 통해서 수행하였다. 20명의 피조사자에게 이메일로 입력할 때 주의사항과 실험 시스템의 주소를 알려주었다. 연구팀은 피조사자들이 이메일을 받은 후 4시간 이내에 시스템에서 작업을 마치도록 요청하였다. 시스템상으로 피조사자들이 각 비디오에 대한 오디오 요약을 듣고 요약문을 입력하게 하고, 그 다음 단계로 적합한 항목(들)을 선택하도록 하였다. 이때 요약문은 최소 세 문장 이상을 기술하도록 하였고, 피조사자들이 이전에 표본 비디오들을 들은 적이 있는지 비디오별로 체크하도록 하였다.

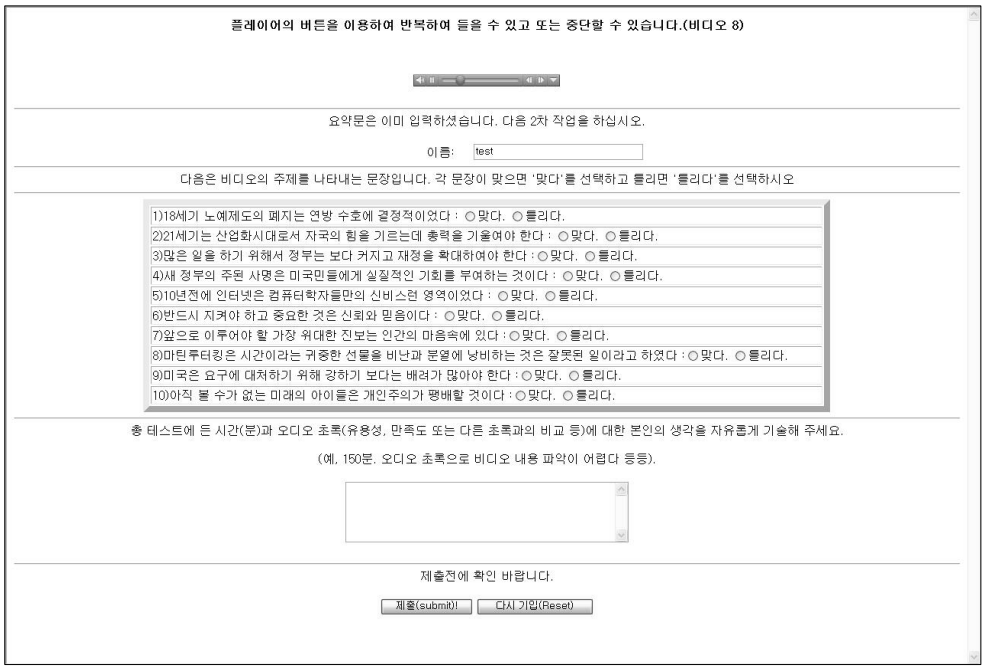
표본 비디오 수가 8개 이므로 각 집단별로 사례수는 80이 되고, 집단 1과 2의 총 사례수는 160이 된다. 이때 10개로 구성된 전체 항목을 보고 비디오 내용을 유추할 가능성이 높아진다고 판단하여 항목을 선택한 다음 이전 화면으로 돌아가 비디오 요약을 재작성하지 못하도록 시스템적으로 제한하였다. 따라서 비디오 요약을 완전하게 기술하지 못했는데 제출 버튼을 잘못

비디오코드	오디오 초록 듣기	비디오코드	오디오 초록 듣기
N0.1	클릭	N0.1	클릭
N0.2	클릭	N0.2	클릭
N0.3	클릭	N0.3	클릭
N0.4	클릭	N0.4	클릭
N0.5	클릭	N0.5	클릭
N0.6	클릭	N0.6	클릭
N0.7	클릭	N0.7	클릭
N0.8	클릭	N0.8	클릭

〈그림 2〉 시스템 인터페이스



〈그림 3〉 요약문 입력 화면



〈그림 4〉 항목 선택과 시간 및 만족도 입력 화면

입력한다든지 아주 특별한 경우에만 본인의 이름을 수정하여(예, 김수미2) 다시 입력하도록 하였다. 연구팀은 데이터베이스에 입력된 답변들을 체크하고 잘못된 점, 미진한 점이 있으면 다시 입력하도록 피조사자에게 요청하였다.

5.2.4 평가 결과

1) 비디오의 의미 파악 정확도

(1) 요약문 평가

각 요약문의 점수 범위를 '0-20'로 하고 5개의 기준(0-전혀 틀림, 5-조금 맞음 10-절반 맞음, 15-절반 이상 맞음 20-정확히 맞음)을 적용하여 2명의 연구자가 점수를 배정한 후 2개의 점수 평균값을 계산하여 최종 점수를 할당하였다. 전체 사례는 160개이나 이미 비디오를 본 적이 있다고 대답한 12개 사례(그룹 1->3개, 그룹 2->9개)를 제외시켜 총 사례수는 148개가 되었다. 제안 요약문과 위치 정보 기반 요약문의 기술 통계값과 t 검증 결과는 〈표 6〉에 나와 있다. 〈표 6〉에 나와 있는 것처럼 제안 요약문의 평균 정확도(15.32)가 위치 기반 요약문 평균 정확도(13.56) 보다 높았으며, 그 차이가 통계적으로 유의미한 것으로 나타났다($p(=0.00) < 0.05$).

(2) 항목 선택

각 비디오에 대한 10개 항목들의 점수 범위를 '0-20'로 하고 전체가 맞으면 20점, 한 개 틀릴 때 마다 2점을 감점하는 방식으로 계산하였다. 예를 들어서 맞는 항목을 틀리다고 하든지 아니면 반대로 틀린 항목을 맞는다고 하든지

〈표 6〉 요약문 정확도 분석 결과

평균(표준 편차) / 요약문 종류	평 균	표준 편차
제안 요약문	15.32	3.00
위치 기반 요약문	13.56	4.32
t-검증 결과		
대 응		유의확률
제안 요약문 vs. 위치 기반 요약문		0.00**
등분산 검정(F=22.63, $p(=0.00) < 0.05$)(등분산이 가정됨)		

〈표 7〉 항목 선택 정확도 분석 결과

평균(표준 편차) / 요약문 종류	평 균	표준 편차
제안 요약문	13.43	3.09
위치 기반 요약문	12.93	3.36
t-검증 결과		
대 응		유의확률
제안 요약문 vs. 위치 기반 요약문		0.35
등분산 검정(F=0.12, $p(=0.74) > 0.05$)(등분산이 가정되지 않음)		

하면 2점을 감점하였다. 〈표 7〉에 나와 있는 것처럼 제안 요약문의 평균 정확도(13.43)가 위치 기반 요약문의 평균 정확도(12.93) 보다 높았으나, 그 차이가 통계적으로 유의미하지 않은 것으로 나타났다($p(=0.35) > 0.05$).

2) 이용자 관점

오디오 요약에 대한 피조사자의 만족도가 높지 않은 것으로 나타났다. 특히 10명의 피조사자들(50.0%)이 문장 추출 기법에 의해 작성된 요약은 선택된 문장을 단순히 순서대로 나열한 것으로 요약문의 응집성이 떨어져 의미 파악이 힘들다고 언급하였다. 한편 4명의 피조사자(20.0%)들이 오디오 요약이 비디오의 전체 내용을 파악하는데 도움을 준다고 하였다. 5명(25%)의 피조사자들은 음성합성기로 구성한 오디오 요약의 기계음이 부자연스럽고 듣기 어려운 부분이 있어서 내용 파악이 힘들었다는 의견을 내놓았다. 이는 본 실험을 위해서 사용된 소규모 음성 합성기 대신 현재 상용으로 판매되는 고급 음성 합성기를 사용하면 얼마든지 감소시킬 수 있는 문제점이라고 생각한다.

비디오의 내용 파악 측면에서 볼 때, 4명의 피조사자들(20.0%)은 텍스트 초록이 오디오 요약 보다 효과적일 것으로 생각했다. 이때 피조사자들이 생각하는 텍스트 초록은 일반적으로 비디오 제작자 또는 정보 분석자가 수작업으로 가공한 초록을 의미하고 있어서 매체의 선호도 보다는 텍스트 초록을 통한 내용 파악의 정확도를 더 선호하는 것으로 유추해 볼 수 있다. 이외에 2명의 피조사자들(10.0%)은 영상 초록이 오디오 요약 보다 더 효과적일 것 같다는 의견을 내놓았다. 2명의 피조사자들(10.0%)은 오디오 요약을 텍스트 초록 또는 영상 초록과 병행하여 사용하면 이용자의 혼란을 줄이고 효율성을 발휘할 것 같다는 의견을 밝혔다. 이밖에 오디오 요약을 구성할 때 중간 중간 발췌하지 말고 일정 부분을 일정시간 동안 추출하여 제공하는 것이 더 효율적일 것 같다는 견해를 제시하는 피조사자도 있었다.

5.3 논의

본 연구의 첫 번째 연구 문제는 텍스트 요약에 적용된 이론과 방법들이 오디오 요약에도 그대로 적용될 수 있을 것인가이다. 요약문 정확도 분석 결과, 20점 만점에 두 방법의 평균값이 15.32, 13.56로 각각 나타나 대체적으로 비디오의 내용을 파악하고 있는 것으로 나타났다. 따라서 텍스트 요약에 적용된 이론과 방법이 오디오 요약에도 적용될 수 있을 것으로 생각되나 오디오 정보의 특성에 맞춘 요약 기법에 대한 심도 깊은 연구가 필요할 것으로 생각된다.

둘째, 제안 오디오 요약의 품질이 위치 기반 오디오 요약의 품질 보다 내재적 평가(재현율과 정확률)에서 더 우수하게 나타날 것인가이다. 예측한 대로 제안 오디오 요약의 재현율과 정확률이 더 높은 것으로 나타났다. 제안 요약문의 재현율(0.44)이 위치정보 기반 요약문의 재현율(0.31)과 비교하여 더 높은 것으로 나타났고 통계적으로 유의미한 차이를 나타냈다($p=0.03$). 또한 제안 요약문의 정확률(0.63)도 위치정보 기반 요약문의 정확률(0.37) 보다 더 높은 것으로 나타났고, 이 역시 통계적으로 유의미한 차이를 나타냈다($p=0.01$).

셋째, 이러한 내재적 평가 결과가 외재적 평

가에도 영향을 미칠 것인가이다. 외재적 평가 중 요약문 정확도는 제안 요약문의 평균 정확도(15.32)가 위치 기반 요약문 평균 정확도(13.56) 보다 높았다. 또한 그 차이가 통계적으로 유의미한 것으로 나타나(p=0.00) 내재적 평가 결과가 외재적 평가 결과에 영향을 미치는 것으로 나타났다. 그러나 항목 선택에서는 내재적 평가 결과가 외재적 평가 결과에 영향을 미치지 않은 것으로 나타났다. 즉, 제안 요약문의 평균 정확도(13.43)가 위치 기반 요약문 평균 정확도(12.93) 보다 높았으나, 그 차이가 통계적으로 유의미하지 않은 것으로 나타났다(p=0.35). 이러한 결과가 나오게 되는 원인은 먼저 오디오 요약에서 나오지 않은 내용을 항목 선택의 문항으로 구성한 점이다. 즉, 피조사자들은 오디오 요약을 통해서 비디오의 전체적인 내용을 파악할 수 있으나 요약에 언급되지 않은 자세한 사항들은 유추하기는 어려웠을 것으로 생각한다.

넷째, 이용자 관점에서 본 오디오 요약에 대한 만족도는 어떠한가이다. 오디오 요약에 대한 만족도가 높지 않은 것으로 나타났다. 이는 오디오 요약의 요약문 정확도가 비교적 높게 나타난 것과 다른 결과이다. 자동으로 구성한 오디오 요약에 대하여 이용자의 가장 큰 불만은 요약문의 응집성이 떨어져 문맥이 자연스럽지 못한 경우가 많다는 점으로 나타났다. 따라서 요약문의 가독성을 향상시키기 위해서 요약문을 등위 또는 종속 접속사 등을 사용하여 문장들을 결합하는 표층적 수정 또는 텍스트로부터 생성된 초기 요약문에 수정 규칙들을 적용하여 배경 정보를 추가하는 심층적 수정 방안을 사용할 수 있다(정영미 2005; Mani 2001).

그러나 이러한 정교한 요약 기법들은 방대한 양의 비디오 자료에 적용하기에는 그 실효성이 높지 않다고 생각한다. 이러한 문제를 해결하는 방안 중 하나로 오디오 요약을 텍스트 기반 메타데이터나 영상 초록과 함께 사용하여 시너지 효과를 극대화시키는 방안이 있다. 이때 영상/오디오 초록을 구성할 때 영상과 오디오의 동시성을 최대한 높이기 위해서 오디오가 발췌된 위치에 출현하고 있는 프레임들을 추출하여 영상 요약을 구성한다면 비디오 내용의 의미 파악을 좀 더 용이하게 할 것으로 생각된다. 특히 영상/오디오 요약은 오디오는 물론 영상으로도 많은 정보를 전달하는 예술, 스포츠 분야 비디오에 유용하게 사용될 것이다.

다섯째, 실험 시간의 평균은 86.9분 이었고, 최소 40분, 최대 145분이 소요되었다.

6. 결 론

본 연구는 비디오의 오디오 요약을 자동으로 추출하기 위한 알고리즘을 제안하고, 제안된 알고리즘에 의해서 구성한 오디오 요약의 품질을 평가하여 효율적인 비디오 요약 기법을 제시하였다. 본 연구는 다음과 같은 연구 결과를 도출하였다.

첫째, 텍스트 요약에 적용된 이론과 방법이 오디오 요약에도 적용될 수 있을 것으로 판단된다. 그러나 좀 더 효율적인 오디오 요약을 구현하기 위해서는 오디오 정보의 특성에 맞춘 요약 기법에 대한 연구가 요망된다.

둘째, 제안 오디오 요약의 품질이 위치 기반 오디오 요약의 품질 보다 내재적 평가(재현율

과 정확률)에서 더 우수하게 나타났고, 통계적으로도 유의미한 차이를 보였다.

셋째, 외재적 평가의 요약문 정확도에서는 제안 요약문의 평균 정확도가 위치 기반 요약문 평균 정확도 보다 높게 나타났고, 통계적으로도 유의미한 차이를 보였다. 그러나 비디오의 내용 즉, 사건, 사실 등을 정확하게 기술하는 항목(들)을 선택하는 작업에서는 제안 요약문의 평균 정확도가 위치 기반 요약문 평균 정확도 보다 높았으나, 그 차이가 통계적으로 유의미하지 않은 것으로 나타났다. 이러한 결과가 나오게 되는 원인 중 하나는 오디오 요약에서 나오지 않은 내용을 항목 선택의 문항으로 구성한 점이다. 즉, 피조사자들은 제안 또는 위치 기반 오디오 요약을 통해서 비디오의 대략적인 내용을 파악할 수 있으나 요약에 언급되지 않은 자세한 사항들은 유추하기는 어려웠을 것이라고 판단된다.

넷째, 오디오 요약에 대한 만족도가 높지 않은 것으로 나타났다. 자동으로 구성한 오디오 요약에 대하여 이용자의 가장 큰 불만은 요약문의 응집성이 떨어져 문맥이 자연스럽지 못한 경우가 많다는 점으로 나타났다. 따라서 요약문의 가독성을 향상시키기 위해서 방대한 양의 비디오 자료에 적용하기에 그 실효성이 높지 않은 심층적 수정 방안 대신 오디오 요약을 텍스트 기반 메타데이터나 이미 인터넷 또는 디지털 도서관에서 활용되고 있고 영상 초록과 함께 사용하는 방안을 제안하였다.

참 고 문 헌

김재곤 등. 2000. 효율적인 비디오 브라우징을 위한 동적 요약 및 요약 기술구조. 『방송공학회논문지』, 5(1): 82-93.

정영미. 2005. 『정보검색연구』. 서울: 구미무역 출판부.

진성원 등. 2005. 개인화된 의미 기반 콘텐츠 소비를 위한 지능형 방송 시스템과 서비스. 『방송공학회논문지』, 10(3): 422-435.

Edmunson, H. P. 1969. "New methods in automatic extracting." *Journal of the ACM*, 16(2): 265-285.

Furini, M. and V. Ghini. 2006. "*An Audio-video smmarisation scheme based on audio and video analysis.*" *Proceedings of the IEEE Consumer Communications and Networking Conference*(CCNC '06), vol. 2, Las Vegas, NV, USA, 8-10 January, 2006, 1209-1213.

Gunther, R., R. Kazman, and C. MaccGregor. 2004. "Using 3D sound as a navigational aid in virtual environments." *Behaviour and Information Technology*, 23(6): 435-446.

Hauptmann, A. G. 2005. "Lessons for the future from a decade of informedia video analysis research." *Lecture Notes in*

Computer Science, Vol. 3568: 1-10. [cited 2006.6.25]. 〈http://www.informedia.cs.cmu.edu/documents/CIVR05_Hauptmann.pdf〉.

Kristin, B. et al. 2006. *Audio surrogation for digital video: A design framework*. UNC School of Information and Library Science(SILS) Technical Report TR 2006-21.

Kupiec, J., J. Pedersen, and F. Chen. 1995. "A trainable document summarizer." *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 68-73.

Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2): 159-165.

Mani, I. 2001. *Automatic summarization*. Amsterdam: John Benjamins Publishing Co.

Marchionini, G., B. M. Wildemuth, and G. Geisler. 2006. "The Open Video Digital Library: A Möbius strip of research and practice." *Journal of the American Society for Information Science and Technology*, 57(12): 1623- 1643.

Money, A. G. and H. Agius. 2008. "Video summarisation: A conceptual framework and survey of the state of the art." *Journal of visual communication and image representation*, 19(2): 121-143.

Money, A. G. and H. Agius. 2009. "Analysing user physiological responses for affective video summarisation." *Displays*, 30: 59-70.

Myaeng, S. H. and D. H. Jang. 1999. "Development and evaluation of a statistically-based document summarization system." In I. Mani and M. T. Maybury, eds. *Advances in automatic text summarization. Cambridge*, MA: The MIT Press, 61-70.

Over, P. et al. 2005. TRECVID, 2005: "An introduction." *Proceedings of the TRECVID, 2005*(Gaithersburg, MD), 1-14.

Schmandt, C. and A. Mullins. 1995. "AudioStreamer: Exploiting simultaneity for listening." *CHI '95: Conference companion on human factors in computing systems*, Denver, Colorado, United States, 218-219. from 〈http://doi.acm.org.libproxy.lib.unc.edu/10.1145/223355.223533〉.

Smeaton, A. F. 2007. "Techniques used and open challenges to the analysis, indexing and retrieval of digital video." *Information Systems*, 32: 545-559.

Smeaton, A. F. and P. Browne. 2006. "A usage study of retrieval modalities for video shot retrieval." *Information Processing and Management*, 42(5): 1330-1344.

Song, Y. and G. Marchionini. 2007. "Effects of audio and visual surrogates for making

sense of digital video." *Proceedings of CHI 2007*, San Jose, CA, USA. 867-876.

Sparck Jones, K. 2007. "Automatic summarising: The state of the art." *Information Processing and Management*, 43: 1449-1481.

Witbrock, M. and A. Hauptmann. 1998. "Speech recognition for a digital video library." *Journal of the American Society for Information Science and Technology*, 49(7): 619-632.

Yang, M. and G. Marchionini. 2005. "Deciphering visual gist and its implications for video retrieval and interface design." *Conference on Human Factors in Computing Systems(CHI)*. Portland, OR. Apr. 2-7.