음향학적 자질을 활용한 비디오 스피치 요약의 자동 추출과 표현에 관한 연구*

Investigating an Automatic Method for Summarizing and Presenting a Video Speech Using Acoustic Features

김현희 (Hyun-Hee Kim)**

초 록

스피치 요약을 생성하는데 있어서 두 가지 중요한 측면은 스피치에서 핵심 내용을 추출하는 것과 추출한 내용을 효과적으로 표현하는 것이다. 본 연구는 강의 자료의 스피치 요약의 자동 생성을 위해서 스피치 자막이 없는 경우에도 적용할 수 있는 스피치의 음향학적 자질 즉, 스피치의 속도, 피치(소리의 높낮이) 및 강도(소리의 세기)의 세 가지 요인을 이용하여 스피치 요약을 생성할 수 있는지 분석하고, 이 중 가장 효율적으로 이용할 수 있는 요인이 무엇인지 조사하였다. 조사 결과, 강도(최대값 dB과 최소값 dB간의 차이)가 가장 효율적인 요인으로 확인되었다. 이러한 강도를 이용한 방식의 효율성과 특성을 조사하기 위해서 이 방식과 본문 키워드 방식간의 차이를 요약문의 품질 측면에서 분석하고, 이 두 방식에 의해서 각 세그먼트(문장)에 할당된 가중치간의 관계를 분석해 보았다. 그런 다음 추출된 스피치의 핵심 세그먼트를 오디오 또는 텍스트 형태로 표현했을 때 어떤 특성이 있는지 이용자 관점에서 분석해 봄으로써 음향학적 특성을 이용한 스피치 요약을 효율적으로 추출하여 표현하는 방안을 제안하였다.

ABSTRACT

Two fundamental aspects of speech summary generation are the extraction of key speech content and the style of presentation of the extracted speech synopses. We first investigated whether acoustic features (speaking rate, pitch pattern, and intensity) are equally important and, if not, which one can be effectively modeled to compute the significance of segments for lecture summarization. As a result, we found that the intensity (that is, difference between max DB and min DB) is the most efficient factor for speech summarization. We evaluated the intensity-based method of using the difference between max-DB and min-DB by comparing it to the keyword-based method in terms of which method produces better speech summaries and of how similar weight values assigned to segments by two methods are. Then, we investigated the way to present speech summaries to the viewers. As such, for speech summarization, we suggested how to extract key segments from a speech video efficiently using acoustic features and then present the extracted segments to the viewers.

키워드: 스피치 요약, 비디오, 피치, 강도, 내재적 평가, 스피치 속도
speech summarization, acoustic features, prosodic features, TED Talks, Praat

---

* 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 연구되었음 (NRF-2011-342- B00025).

** 명지대학교 문헌정보학과 교수(kimhh@mju.ac.kr)

■ 논문접수일자 : 2012년 11월 21일 ■ 최초심사일자 : 2012년 11월 23일 ■ 게재확정일자 : 2012년 12월 13일
■ 정보관리학회지, 29(4), 191-208, 2012. 〔http://dx.doi.org/10.3743/KOSIM.2012.29.4.191〕

1. 서 론

1.1 연구 배경과 목적

대학이나 기관은 방대한 스피치 형태의 교육 자료를 인터넷 사이트들을 통해서 제공하고 있지만, 이용자들은 적절한 메타데이터와 요약 정보의 부족으로 이들 자료를 효율적으로 활용하지 못하는 경우가 많아지고 있다. 따라서, 이러한 방대한 정보원에 대한 효율적인 접근을 가능하게 하는 스피치 요약의 자동 생성에 대한 연구가 필요한 시점이 되었다.

스피치에서 핵심 내용을 추출하는 것과 추출한 내용을 효과적으로 표현하는 것은 스피치 요약을 생성하는데 있어서 가장 중요한 작업이다(van Houten, Oltmans, & van Setten, 2000). 스피치 요약을 추출하기 위해서 사용하는 방법으로 변수간의 연관성이나 유사성에 기초한 특징 기반 방식이 있다. 특징 기반 방식은 단어의 출현 빈도 정보 등을 이용하는 어휘적 기법(lexical method), 문장 위치와 같은 구조 정보를 이용하는 구조적 기법(structural method), 문장에서 새로운 명사들의 출현 빈도와 같은 정보를 이용하는 담화적 기법(discourse method), 그리고 스피치의 속도(speaking rate), 피치(pitch, 소리의 높낮이), 강도(intensity, 소리의 세기) 등을 이용하는 음향학적 기법 등을 포함하고 있다.

최근 비디오 스피치 자막에 기초한 어휘적 기법 또는 담화적 기법 대신 스피치의 피치 또는 강도 등을 이용하는 음향학적 자질을 이용한 기법에 대한 많은 연구들이 진행되고 있다. Maskey와 Hirschberg(2006)는 음향학/운율적 자질을 스피치 요약에 효율적으로 활용하였고, Zhang과 Fung(2012)은 화자(speaker)의 정규화한 음향학적 자질을 이용하여 강의 자료 요약의 성능을 개선시킬 수 있다고 주장하였다. 이와 같이 음향학적 자질 기법에 대한 연구가 많아지고 있는 것은 비디오 자막의 인식과 분석이 스피치 자막의 자동 인식에서 생기는 에러, 스피치 자막의 세그먼트(문장)간의 명확한 구분의 어려움, 중복적이고 문법적으로 맞지 않는 세그먼트들 때문에 쉽지 않다는 점과 무관하지 않다(Liu & Hakkani-Tür, 2011).

음향학적 자질을 사용하는 기본 가정은 화자에 의해서 강조된 문장 또는 단어가 요약을 위한 중요한 정보를 제공할 것이라는 사실에 있다. 구체적으로, 소리의 세기를 판단하는 데 쓰이는 지표인 강도(단위: dB)는 사람들이 스피치에서 중요한 부분을 말할 때 좀 더 큰 소리로 말하는 경향이 있음을 가정한 것이다. 소리의 높낮이(성대의 초당 진동수)를 나타내는데 쓰이는 지표인 피치(단위: Hz)는 주제 전환(topic shifts)은 피치의 변동과 관련되며 새로운 주제는 종종 중요한 내용이 담긴 문장들과 함께 소개되고 있다는 사실에 근거한 것이다(Maskey, 2008; Hirschberg & Naktani, 1996). 이외에 스피치 속도는 사람들이 중요한 부분을 말할 때 스피치 속도가 빠르게 또는 느리게 바뀌는 경향이 있음을 가정한 것이다.

스피치의 주요 세그먼트(문장)를 추출한 다음에는 이를 오디오 또는 텍스트 형태로 표현할 수 있다. 오디오 요약은 비디오 스트림(stream)에서 중요한 오디오 부분들을 직접 추출하여 구성하거나 또는 텍스트 요약을 음성 합성기에 의해서 오디오 요약으로 변환시켜서 만들 수 있다.

이 두 종류의 요약은 각자 독자적인 특징을 갖고 있다. 텍스트 요약은 스피치의 주제를 명확하게 하면서, 브라우징과 처리를 더 용이하게 하는 반면 오디오 요약은 음향학적 정보(예, 화자의 감정)를 전달하기 때문에 스피치 내용을 좀 더 잘 이해할 수 있게 한다(Chen & Lin, 2012; Furui, Kikuchi, Shinnaka, & Hori, 2004).

본 연구는 스피치의 속도, 피치 및 강도의 세 가지 음향학적 자질을 이용하여 스피치 요약을 생성할 수 있는지 그리고 이 중 가장 효율적으로 이용할 수 있는 요인이 무엇인지 조사해 보고 추출된 스피치의 핵심 세그먼트를 오디오 또는 텍스트 형태로 표현했을 때 어떤 특성이 있는지 이용자 관점에서 분석해 봄으로써 스피치 요약을 효율적으로 추출하여 표현하는 방안을 제안하는 것을 연구 목적으로 하고 있다.

1.2 연구 문제와 방법

음향학적 자질을 이용하여 효율적인 스피치 요약을 구현하는 방안을 모색하기 위해서 다음과 같은 다섯 가지 연구 문제들을 조사한다.

- 연구 문제 1: 스피치 속도, 피치 및 강도가 스피치에서 중요한 내용을 추출하는 기준으로 얼마나 식별력이 있는가?
- 연구 문제 2: 연구 문제 1에서 식별력이 있다고 확인된 요인들이 스피치 요약을 위한 세그먼트의 중요도를 계산하기 위한 기준으로 동일한 효과가 있는지 만약 동일하지 않다면 어떤 요인이 더 효율적인가?
- 연구 문제 3: 연구 문제 2에서 가장 효율적이라고 분석된 요인을 이용한 방식과 전통적인 본문 키워드 방식간에 요약문의 품질과 이 두 방식에 의해서 각 세그먼트에 할당된 가중치간의 상관도 측면에서 어떤 차이가 있는가?
- 연구 문제 4: 피치와 강도간에 어떤 상관관계가 있으며 피치내 또는 강도내의 여러 요소간(예, 피치 최대값 vs. 피치 최대값 및 최소값간의 차이)에 어떤 상관 관계가 있는가?
- 연구 문제 5: 스피치 요약의 오디오 및 텍스트 형태가 브라우징 단계에서 이용자에서 어떤 영향(효과)을 주는가?

이러한 연구 문제들을 조사하기 위한 표본 비디오 자료는 음성으로 많은 정보를 표현하는 강의, 교육 및 연설 비디오로 정하고 유투브 사이트에서 28개의 TED Talks를 포함한 총 40개의 영어로 된 비디오들을 선정하였다. 스피치 요약의 효율성을 평가하기 위해서 요약 기법의 성능을 평가하는 내재적 평가를 하였다(정영미, 2005). 내재적 평가를 위해서 스피치 대본에서 비디오의 의미를 가장 잘 나타내는 문장들을 추출하여 표준 요약을 구성하였고(자세한 설명은 "4.2.1 표준 요약 및 본문 키워드 기반 요약의 구성" 참조), 통계 분석을 위해서 SPSS 통계 패키지를 사용하였다.

2. 선행 연구

본 장은 먼저 음향학적 자질을 활용한 스피치 요약의 추출에 관한 연구를 다룬 후, 오디오, 텍스트 또는 이미지 형태로 표현되는 비디오 요약의 표현에 관한 연구를 기술한다.

2.1 추출

스피치 요약을 추출하기 위해서 사용되는 방법에는 앞에서 설명한 특징 기반 방식과 규칙 기반 방식이 있다(〈표 1〉 참조). 규칙 기반 방식은 중복은 최소화하면서 적합성과 다양성을 최대화하는 최대 한계 적합성(Maximum Marginal Relevance, MMR) 모형 기반 방식, 단어와 단어, 단어와 문장 및 문장과 문장간의 의미 유사성 추정치를 제공하는 분석 기법인 잠재의미 분석(Latent Semantic Analysis, LSA), 벡터간의 유사도에 기초한 벡터공간모형(Vector Space Model, VSM) 등을 포함하고 있다. 다음은 스피치 요약의 추출에 관한 연구를 특징 기반 요약에 대한 연구와 특징 기반 요약과 규칙 기반 요약간의 비교 연구로 구분하여 설명한다.

2.1.1 특징 기반 요약

Maskey와 Hirschberg(2005, 2006)는 음향학/운율적 자질을 스피치 요약에 효율적으로 활용하였다. 구체적으로, Maskey와 Hirschberg(2005)는 방송 뉴스를 요약하기 위해서 음향학적 자질을 어휘적, 구조적 및 담화적 특성들과 결합하였을 때 가장 좋은 성능을 나타냈다고 보고하였다. 또한 스피치 자막을 이용할 수 없을 때에는 음향학적 자질과 구조적 특징을 결합하였을 때 가장 좋은 결과를 나타냈다고 기술하고 있다. Maskey와 Hirschberg(2006)는 음향학적 자질 기법이 방송 뉴스에서 중요한 문장을 추출하는데 적합하다고 보고하였다. 유사하게, Zhang과 Fung(2007)은 어휘적 특징의 활용 없이 음향학적 및 구조적 특징의 활용만으로 방송 뉴스 요약이 가능하다고 제안하고 있다.

Zhang, Chan 및 Fung(2007)은 강의 자료를 요약할 때 어휘적 특징이 음향학적 자질 보다 더 많은 기여를 한다고 기술하고 있다. 또한, 이 연구는 도입, 본문 및 결론으로 구성되는 수사학적 구조가 강의 자료에 내재되어 있으며,

〈표 1〉 선행 연구 목록

연 구	특징 기반 방식				규칙 기반 방식 (MMR, LSA, VSM)	장르
	어휘적 (Le)	구조적 (St)	음향학적 (Ac)	담화적 (Di)		
Maskey & Hirschberg(2005)	Le	St	Ac	Di		방송 뉴스
Maskey & Hirschberg(2006)		St	Ac			방송 뉴스
Zhang & Fung(2007)	Le	St	Ac			중국어 방송 뉴스
Zhang, Chan, & Fung(2007)	Le	St	Ac	Di		강의 자료
Zhang, Chan, Fung, & Cao(2007)	Le	St	Ac			방송 뉴스와 강의 자료
Xie et al.(2009)	Le	St	Ac	Di		회의 자료
Zhang & Fung(2012)	Le		Ac			중국어 강의 자료
Murray et al.(2005)	Le		Ac		MMR, LSA	회의 자료
Fujii et al.(2008)	Le		Ac		MMR	일본어 강의 자료
Lin et al.(2009)	Le	St	Ac		MMR, LSA, VSM, etc.	중국어 방송 뉴스
Zhu et al.(2009)			Ac		MMR	방송 뉴스

요약문 생성의 성능을 개선시키기 위해서 음향학/운율적 자질이 이러한 수사학적 정보를 모형화 하는 것을 가능하게 한다고 제안하고 있다. Zhang, Chan, Fung 및 Cao(2007)는 방송 뉴스와 강의 자료라는 두 개의 장르를 비교하여 스피치 요약을 위한 음향학/운율적, 언어학적 및 구조적 특징에 대한 비교 연구를 수행하였다. 비교 결과, 방송 뉴스의 경우, 음향학/운율적 및 구조적 특징이 어휘적 특징보다 더 중요하게 나타났다. 이와 반대로, 강의 자료의 경우, 어휘적 특징이 음향학/운율적 및 구조적 특징보다 더 중요하게 나타났다. 이와 같은 실험 결과는 방송 뉴스의 앵커와 리포터의 스피치 스타일은 상대적으로 일관성이 있고 전형적인 뉴스 스토리의 흐름이 있는 반면, 강의 진행자의 스피치 스타일은 매우 다양하다는 사실에 기인한다고 기술하고 있다.

Xie, Hakkani-Tur, Favre 및 Liu(2009)는 적절하게 정규화한 음향학적 자질의 채택은 어휘적 특징과 같은 또는 더 나은 성능을 나타낸다고 보고하였다. Zhang과 Fung(2012)은 화자의 음향학적 자질을 정규화 함으로써(예, 특정 피치값을 피치값의 전체 범위로 나누어줌) 강의 자료 요약의 성능을 개선시킨다는 것을 보여 주었다.

2.1.2 특징 기반 요약과 규칙 기반 요약간의 비교

Murray, Renals 및 Carletta(2005)는 회의 비디오 요약에서 MMR 방법, 음향학적 방법 및 잠재의미분석 방법의 성능을 비교하였다. 비교 결과, 잠재의미분석 방법이 가장 높은 성능을 보여 주었고, MMR 기법도 중복을 줄이는 기능 때문에 좋은 성능을 보여 주었다고 보고하였다. 이와 반대로, Fujii, Yamamoto, Kitaoka 및 Nakagawa(2008)는 일본어 강의 자료를 요약하는데 있어서, 어휘적 및 음향학적 자질을 이용하는 특징 기반 방법이 MMR 기법보다 더 우수하다고 보고하고 있다. 이들은 이러한 결과가 나오게 된 것은 강의는 보통 슬라이드를 이용하여 조직화하기 때문에 회의 자료 또는 방송 뉴스보다는 중복도가 상대적으로 낮기 때문으로 기술하고 있다. Lin, Chen 및 Wang(2009)은 특징 기반 방식의 성능이 MMR, LSA 및 VSM을 포함한 규칙 기반 방식의 성능보다 우수하다고 보고하고 있다. 특히, 이들은 중국어 방송 뉴스 요약에서 음향학적 자질이 어휘적 특징보다 더 많이 공헌하고 있다고 기술하고 있다.

스피치 요약에 대한 대부분 연구는 단일 문헌을 다루고 있지만, Zhu, Penn 및 Rudzicz(2009)는 복수의 문헌들을 다루고 있다. 이들은 중요한 발언(문장)을 확인하고 발언(문장)간의 유사성을 측정하기 위해서 스피치 자막을 사용하지 않고 복수 레코딩 자료에서의 음향학적 패턴의 발생 통계 정보를 사용하였다. 또한 스피치 요약에서의 중복를 제거하기 위해서 MMR 모형도 이용하였다.

2.2 표현

비디오 요약은 크게 동영상, 오디오, 이미지 또는 텍스트 형태로 표현될 수 있다. 이용자의 인지적 과정의 질적 분석을 적용한 비디오 요약의 표현 스타일의 유용성을 연구한 연구들을 살펴 보면 다음과 같다.

Ding, Marchionini 및 Soergel(1999)은 비디

오 요약의 세 가지 유형 즉, 이미지, 언어(키워드/구절) 및 이미지와 언어를 결합한 멀티 모달(multi-modal) 형식에 대해서 분석해 보았다. 분석 결과, 이용자는 비디오의 전체 의미를 쉽게 파악할 수 있게 하고 주제를 명확하게 하는 언어 정보와 감정, 정서 및 흥분을 전달하면서 주의를 끄는 이미지 정보가 서로 보완하는 멀티 모달 형식의 요약을 선호하였다. Turner(1994)는 텍스트와 이미지는 비디오 정보를 표현하는데 있어서 서로 보완적이며 상호 의존적인 측면이 있다고 하였다. Cawkell(1995)은 이미지가 항상 언어로 표현되는 텍스트보다 성능이 더 좋은 것은 아닌데 이는 이미지가 추상적 개념을 표현하는데 있어서 언어의 설명력(descriptive power)을 대체하지 못하는 경우도 있기 때문이라고 하였다. Marchionini, Song 및 Farrell(2009)은 오디오 요약의 품질이 비디오 의미 파악을 하는데 있어서 오디오와 이미지 정보가 결합된 멀티 모달 기반 요약의 품질과 거의 유사하다고 기술하고 있다. 이와 유사하게, 김현희(2011)는 이용자들이 오디오와 이미지가 비동시적으로 결합된 멀티 모달 요약을 시청한 후 비디오 내용을 파악할 때, 이미지 정보보다 오디오 정보가 더 유용하다고 보고하였다.

선행 연구들을 살펴본 결과, 강의 자료의 스피치 요약을 위해서 스피치의 속도, 피치 및 강도가 어떤 효과가 있는지에 대한 체계적인 연구가 부족한 편이다. 또한 추출된 스피치 요약이 어떤 형식으로 표현될 때 의미 파악을 더 정확하게 할 수 있는지에 대한 연구도 거의 없는 편이다. 따라서, 본 연구는 음향학적 자질을 이용하여 스피치 요약을 효율적으로 추출하여 표현하는 방안을 제안하고자 한다.

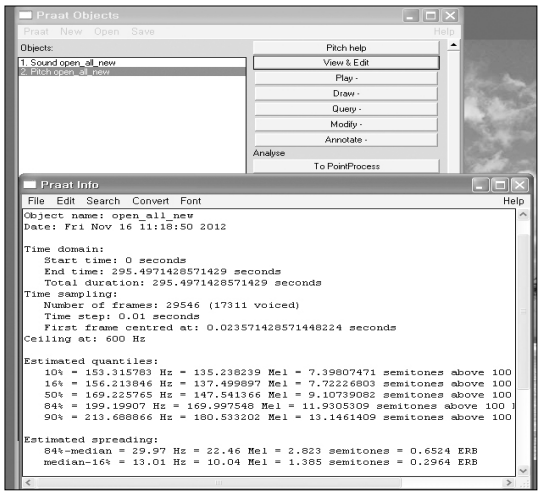
3. 스피치 속도, 피치 및 강도의 식별력 측정

본 장은 연구 문제 1(스피치 속도, 피치 및 강도의 식별력 측정)을 검증하기 위해서 표본 비디오와 분석 도구를 이용하여 측정한 내용과 결과를 기술한다.

3.1 표본 비디오와 분석 도구

스피치 내용 분석을 위해서 28개의 TED Talks를 포함한 총 40개의 영어로 된 표본 비디오를 유투브에서 선정하였다. 표본 비디오의 선정 기준은 태그를 3개 이상 가지고 있으면서, 음성으로 많은 정보를 표현하는 강의, 교육 및 연설 비디오로 4분~25분 사이에 있는 것들을 선택하였다. 특히 28개의 TED Talks 비디오는 비디오의 대본이 한글로 되어 있는 자료를 한글판 TED Talks 사이트(http://www.ted.com/translate/languages/ko)에서 선정하였다.

스피치의 음향학적 자질을 분석하기 위해서 Praat(5.2.42)(프라트) (http://www.fon.hum.uva.nl/praat/)을 이용하였다(Boersma & Weenink, 2006). 프라트는 음성 파일을 오픈한 후 음성파형을 보면서 스피치 속도, 피치 및 강도의 최대값, 최소값, 중앙값, 평균, 범위 등을 구할 수 있다. 이외에 프라트는 음성을 분석, 편집, 합성하는 기능도 가지고 있다. 〈그림 1〉은 프라트에서 비디오 3의 스피치 파일을 불러와 피치 분석을 한 후 여러 피치값의 통계(최대값, 최소값, 범위, 평균 등)를 화면에서 확인하는 인터페이스이다. 각 세그먼트의 통계값은 다시 세그먼트 파일을 불러와 같은 방법으로 실행하여 구한다.



〈그림 1〉 프라트(Praat) 사용 예

3.2 식별력 측정

프라트를 사용하여 표본 비디오의 스피치를 분석한 절차는 다음과 같다.

1) 스피치 파일 생성: 각 표본 비디오에서 곰 플레이어를 이용하여 스피치만을 분리하여 wav파일로 저장한다.

2) 측정: 프라트를 이용하여 40개의 표본 비디오 집합에서 무작위로 추출한 15개의 표본 비디오의 각 스피치 파일을 불러와 이 파일을 여러 개의 세그먼트(문장)별로 나눈 후 세그먼트파일들로 저장한다. 이때 작업을 용이하게 하기 위해서 표본 비디오의 자막을 참조하였다. 그런 다음, 프라트를 이용하여 각 스피치의 세그먼트 파일들을 불러와서 7개의 요소 즉, 스피치 속도, 3개의 피치값 통계(F0 최대값, F0 최소값 및 F0 최대값과 F0 최소값간의 차이) 및 3개의 강도값 통계(최대값 dB, 최소값 dB 및 최대값 dB과 최소값 dB간의 차이)를 추출하여 분석하였다. 그 결과, 7개 요소 중에서 4개 요소(F0 최대값, F0 최대값과 F0 최소값간의 차이, 최소값 dB 및 최대값 dB과 최소값 dB간의 차이)만이 세그먼트간의 차이를 구별해 주는 식별력이 있었고, 나머지 3개 요소(스피치 속도, F0 최소값 및 최대값 dB)는 식별력이 없었다.

예를 들어서, 대부분의 표본 비디오의 세그먼트들의 스피치 속도가 거의 유사하게 나타나 식별력이 없었다. 또한, 비디오 3(4장의 〈그림 3〉 참조)의 경우에서 확인할 수 있는 것처럼 각 표본 비디오에서 피치 최대값(F0 최대값)은 세그먼트간에 큰 차이를 보이고 있는데 반해서, 피치 최소값(F0 최소값)은 거의 차이를 나타내지 못했다. 다른 한편, 비디오 3(〈그림 4〉 참조)의 경우에서 확인할 수 있는 것처럼 각 표본 비디오에서 강도 최소값(최소값 dB)은 세그먼트간에 큰 차이를 보이고 있는데 반해서 강도 최대값(최대값 dB)은 큰 차이가 없는 것으로 나타났다. 이에 따라서 본 연구에서는 식별력이 있는 네 개의 요소만을 사용하여 스피치 요약을 구성하기로 한다.

4. 음향학적 자질을 이용한 스피치 요약 구성과 평가

본 장은 스피치 요약 구성과 평가에 대해서 기술한다. 스피치 요약 평가에서는 연구 문제 2~5를 검증하기 위해서 표준 요약, 본문 키워드 기반 요약 및 실험 시스템을 구성한 후 구성된 요약문에 대한 내재적 평가를 한다. 또한 스피치 요약의 오디오 또는 텍스트 형태의 장단점을 이용자 관점에서 조사해 본다.

4.1 스피치 요약 구성

세 단계로 구성된 요약문 구성 절차를 기술한 후, 이러한 절차에 따라서 구성된 비디오 3의 요약문 예를 보여준다.

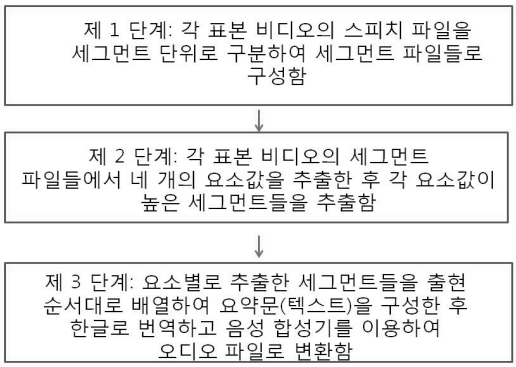
4.1.1 구성 절차

앞에서 기술한 식별력 측정 단계에서 네 개의 요소(F0 최대값, F0 최대값과 F0 최소값간의 차이, 최소값 dB 및 최대값 dB과 최소값 dB간의 차이)가 세그먼트간의 차이를 구별해 주는 식별력이 있는 것으로 나타났다. 다음은 이러한 네 개의 요소를 이용한 스피치 요약의 구성 절차를 세 단계로 구분하여 설명한다(〈그림 2〉 참조).

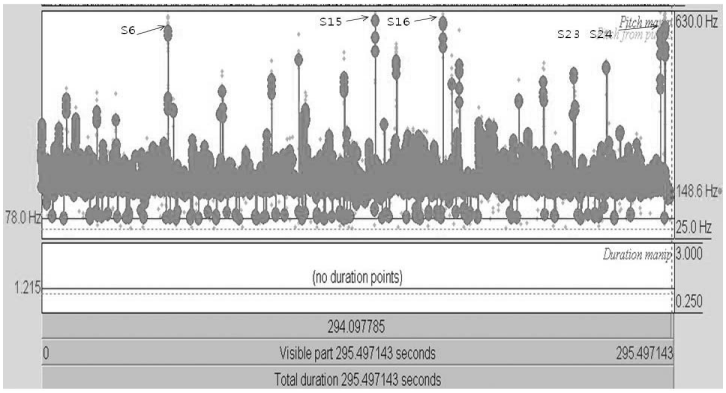
1) 제 1단계: 식별력 측정 단계에서 이미 분석한 15개의 표본 비디오들을 제외한 25개의 표본 비디오의 스피치 파일을 앞의 경우처럼 프라트를 이용하여 불러와 이 파일을 여러 개의 세그먼트(문장)별로 나누어 세그먼트 파일들로 구성하였다.

2) 제 2단계: 프라트를 이용하여 각 표본 비디오에 속한 세그먼트 파일들을 분석하여 네 개의 요소값(F0 최대값, F0 최대값과 F0 최소값간의 차이, 최소값 dB 및 최대값 dB과 최소값 dB간의 차이)을 추출한 후 각 요소값이 높은 4~6개의 세그먼트를 추출하였다.

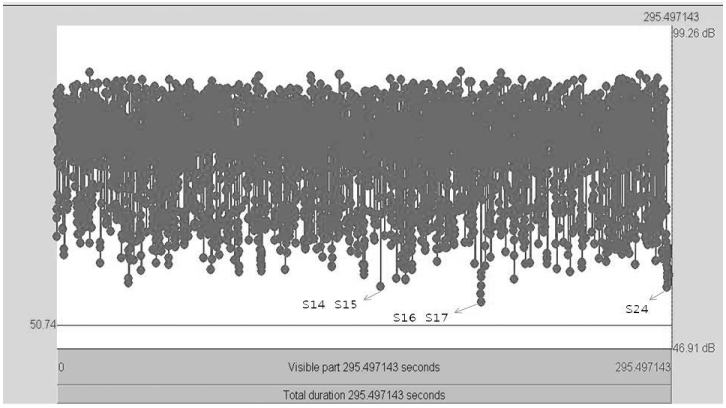
〈그림 3〉은 표본 비디오 3의 스피치 파일의 피치값을 그래프로 표시한 것이다. 예를 들어서, F0 최대값과 F0 최소값간의 차이값을 기준으로 하였을 때, 차이값이 0.77보다 큰 것으로 나타난 5개의 세그먼트(S_6, S_{15}, S_{16}, S_{23} 및 S_{24})가 최종적으로 선정되었다. 세그먼트 6(S_6)에 F0 최대값과 F0 최소값간의 차이값으로 0.95이 할당되었는데 0.95은 "(F0 최대값 - F0 최소값)



〈그림 2〉 스피치 요약의 구성 절차



〈그림 3〉 피치 분석(비디오 3)



〈그림 4〉 강도 분석(비디오 3)

/ F0 범위"라는 공식을 사용하여 산출하였다. 이 공식에 세그먼트 6(S_6)의 F0 최대값(573.9 Hz), 최소값(75.5 Hz) 및 범위값(524.5)을 대입하여 계산한 것이다. 이 때 차이값을 정규화하기 위해서 비디오 3의 스피치에 속한 세그먼트 파일 집단에서 F0 최대값(599.5)과 F0 최소값(75.0)의 차이를 F0 범위(524.5)로 사용하였다. 이와 같이 차이값을 정규화함으로써 다른 정규화된 요소값들과 비교하는 것이 훨씬 용이해 진다. 정규화된 4개의 요소값을 기준으로 하여 40개 스피치 비디오의 요약 총 160개(40 × 4)를 구성하게 된다. 다른 세 개의 요소들과는 달리 최소값 dB값을 이용하는 경우는 값이 낮을수록 가중치는 더 높게 계산되도록 하였다.

〈그림 4〉는 표본 비디오 3의 스피치 파일의 강도값을 그래프로 표시한 것이다. 예를 들어서, 최대값 dB과 최소값 dB간의 차이값을 기준으로 하였을 때, 차이값이 0.47보다 큰 것으로 나타난 5개의 세그먼트(S_{14}, S_{15}, S_{16}, S_{17} 및 S_{24})가 최종적으로 선정되어 스피치 요약을 구성한다. 차이값은 앞의 피치의 경우와 같은 절차로 측정하였다.

3) 제 3단계: 최종적으로 선택된 세그먼트(문장)들을 스피치에서 출현 순서대로 요약문을 구성한 후 한글로 번역하고 이를 음성 합성기인 매직 잉글리쉬 플러스를 이용하여 오디오 파일로 변환하여 스피치 요약을 구성한다.

4.1.2 구성된 요약문의 예

피치(F0 최대값과 F0 최소값간의 차이)와 강도(최대값 dB과 최소값 dB간의 차이)를 기준으로 하여 선택된 비디오 3의 요약문은 〈표 2〉에 나타난 바와 같다. 이 두 가지 방식 이외에 비교를 위해서 본문 키워드 방식에 의해서 요약문을 구성하였다. 〈표 2〉에서 확인한 것처럼 본문 키워드 방식에 의해서 추출된 세그먼트 15와 16(S_{15}와 S_{16})이 강도 기반 방식에 의해서 구성된 요약문에 공통으로 포함된 것으로 미루어 강도가 스피치에서 중요한 부분을 말할 때 좀 더 큰 소리로 말하는 경향이 있다는 가정과 부

〈표 2〉 피치, 강도 및 본문 키워드에 의한 비디오 3의 요약문

피치에 기반한 요약
S_{6}. 이는 사람들이 그들의 질문에 대한 답 또는 정부의 결정 뒤에 숨겨진 이야기를 좀 더 쉽게 얻을 수 있도록 할 것이며, 사람들이 좀 더 효과적으로 개입할 수 있도록 다양한 채널들을 제공할 것이다. S_{15}. 당신이 이 사이트에서 볼 수 있는 것은 다양한 진취 정신이다. 이는 우리 유권자들의 생각이나 아이디어를 얻고 피드백을 수집하며 가공되어 적절한 채널을 통해서 하나의 강력한 정책 아이디어로 정부에 전달되도록 다양한 채널을 이용하고, 공론 영역을 이용하는 것이다. S_{16}. 정부가 많은 사람들의 좋은 아이디어를 수용한다면 유익한 일이 될 것입니다. 이는 효율적이며 우리가 이러한 아이디어를 발표하는 것 외에 이러한 아이디어를 트위터 피드와 블로그를 통해 공론 영역에 제시하면 동료에 의해서 평가되고 활용될 수 있을 것입니다. S_{23}. 그래서, 우리가 정부를 어떤 방식으로 개방할 것인지 그리고 디지털 혁명이 국민에게 의미 있도록 만드는 것이 제가 관심을 갖는 분야입니다. S_{24}. 만약 당신이 이것에 대한 좋은 생각이나 아이디어가 있다면 우리와 공유했으면 합니다. 이는 이 주제가 국가광역 네트워크를 중심으로 구축된 비즈니스를 지원하고 디지털 기술, 디지털 네트워크를 사회와 경제의 발전에 어떻게 이용할 것인지를 도와주기 때문입니다.
강도에 기반한 요약문
S_{14}. 우리가 하고 있는 것에 대해서 가장 흥분되는 것은 당신과 내가 소통을 하는데 사회 네트워크 환경을 이용한다는 점입니다. S_{15}. 당신이 이 사이트에서 볼 수 있는 것은 다양한 진취 정신입니다. 이는 우리 유권자들의 생각이나 아이디어를 얻고 피드백을 수집하며 가공되어 적절한 채널을 통해서 하나의 강력한 정책 아이디어로 정부에 전달되도록 다양한 채널을 이용하고, 공론 영역을 이용하는 것이다. S_{16}. 정부가 많은 사람들의 좋은 아이디어를 수용한다면 유익한 일이 될 것입니다. 이는 효율적이며 우리가 이러한 아이디어를 발표하는 것 외에 이러한 아이디어를 트위터 피드와 블로그를 통해 공론 영역에 제시하면 동료에 의해서 평가되고 활용될 수 있을 것입니다. S_{17}. 우리의 공론 영역 이벤트는 주어진 주제의 범위에서 그들의 생각을 스스로 인식하는 사람들을 위해 계획되었습니다. S_{24}. 만약 당신이 이것에 대한 좋은 생각이나 아이디어가 있다면 우리와 공유했으면 합니다. 이는 이 주제가 국가광역 네트워크를 중심으로 구축된 비즈니스를 지원하고 디지털 기술, 디지털 네트워크를 사회와 경제의 발전에 어떻게 이용할 것인지를 도와주기 때문입니다.
본문 키워드에 기반한 요약
S_{15}. 당신이 이 사이트에서 볼 수 있는 것은 다양한 진취 정신입니다. 이는 우리 유권자들의 생각이나 아이디어를 얻고 피드백을 수집하며 가공되어 적절한 채널을 통해서 하나의 강력한 정책 아이디어로 정부에 전달되도록 다양한 채널을 이용하고, 공론 영역을 이용하는 것이다. S_{16}. 정부가 많은 사람들의 좋은 아이디어를 수용한다면 유익한 일이 될 것입니다. 이는 효율적이며 우리가 이러한 아이디어를 발표하는 것 외에 이러한 아이디어를 트위터 피드와 블로그를 통해 공론 영역에 제시하면 동료에 의해서 평가되고 활용될 수 있을 것입니다. S_{19}. 우리가 계획하고 있는 다음 차례의 공론 영역은 정부 2.0에 관련된 것이 될 것입니다. S_{20}. 우리가 하고자 하는 것은 정부가 어떻게 효과적으로 사회 네트워크에 참여할 것인지에 대한 실제적이고 도움이 되는 아이디어를 얻는 것입니다. S_{21}. 정부가 사회 네트워크를 이용하는데 효과가 있기 위해서는 시민들에게 어느 정도의 권리를 부여해 주어야 합니다.

분적으로 일치한 것으로 보인다(자세한 설명은 "4.2.4 음향학적 자질 방식과 본문 키워드 방식 간의 비교" 참조). 비디오 3의 총 24개의 세그먼트들을 수작업으로 분석한 결과 주제를 전환시키는 세그먼트로 2개(S_{12}와 S_{19})가 있는데 피치를 이용하여 구성한 요약문에 이 두 세그먼트가 포함되지 않는 것으로 나타나 주제 전환을 피치의 변화에 의해서 확인할 수 있다는 사실은 이 사례에서는 확인되지 않았다.

4.2 스피치 요약 평가

스피치 요약의 효율성을 평가하기 위해서 요약 기법의 성능을 평가하는 내재적 평가를 하였다. 내재적 평가를 위해서 각 표본 비디오의 스피치 대본에서 비디오의 의미를 가장 잘 나타내는 문장들을 추출하여 표준 요약을 구성하였다. 그런 다음 네 개의 음향학적 요소(F0 최대값, F0 최대값과 F0 최소값간의 차이, 최소값 dB 및 최대값 dB과 최소값 dB간의 차이)에 기반하여 구성된 요약문을 표준 요약문과 비교하여 평가하였다. 또한 가장 효율성이 높은 것으로 나타난 강도(최대값 dB과 최소값 dB간의 차이)를 이용한 방식과 본문 키워드 방식을 표준 요약문을 기준으로 하여 비교하였다.

4.2.1 표준 요약 및 본문 키워드 기반 요약의 구성

각 비디오의 표준 요약은 두 명의 연구자들이 공동으로 작성하였다. 작성 방법은 연구자들이 비디오를 시청한 다음 유투브 사이트에 있는 텍스트 요약 및 메타데이터 내용을 세밀히 분석한 후 비디오 자막을 마침표를 기준으로 문장 단위로 구분한 후 비디오의 내용을 가장 잘 나타내는 문장들을 선정하였다. 선정된 문장 중에 두 명의 연구자가 똑같이 선정한 문장은 그대로 사용하고 서로 다른 문장을 선정한 경우는 서로 상의하여 최종적으로 적합한 문장을 선택하여 표준 요약을 구성하였다.

본문 키워드 방식에 의한 요약문 구성을 위해서 Extractor 시스템(http://www.extractor.com/)을 이용하였다. Extractor 시스템은 어휘적 기법과 담화적 기법을 이용하여 키워드를 추출하였다. 즉, 단일어를 어미를 제거한 어간으로 변환한 후, 각 어간의 출현빈도와 처음 어간이 출현한 위치 정보를 고려하여 핵심 어간을 추출하였다. 즉 높은 빈도를 갖고 있으면서 첫 번째 어간이 전체 텍스트의 앞부분에 출현하는 어간에 더 높은 가중치를 부여하는 방법을 사용하였다(Turney, 2000). 최대 3개까지 단일어가 결합되도록 한 복합어도 어간으로 변환한 후 단일어와 유사하게 처리하였다. 이와 같이 각 비디오 대본의 핵심 단일어와 복합어를 추출한 후 이를 용어 벡터로 표시한 후 이를 기준으로 각 문장과 본문 키워드간의 유사도를 코싸인 유사계수를 이용하여 계산하였다. 이때 용어의 가중치는 단일어, 복합어에 1.0, 2.0을 각각 할당하였다. 용어의 빈도에 대한 가중치는 빈도가 두 번 이상인 경우에는 일률적으로 원래 가중치에 2를 곱하였다. 최종적으로 코싸인 유사계수가 높은 문장 4~6개를 선정하여 요약문을 구성하였다.

4.2.2 실험 시스템과 피조사자

스피치 요약의 오디오 및 텍스트 형태의 장단점을 이용자 관점에서 조사해 보기 위해서 실험 시스템을 구현하였다. 스피치 요약의 구성 단계

에서 만들어진 40개의 표본 비디오에서 임의로 선택한 TED Talks 비디오 10개의 영문 요약문을 오디오 및 텍스트 형태의 한글 요약문으로 구성하여 실험 시스템으로 구현하였다. 모두 영어로 된 요약문이기 때문에 TED Talks 사이트(http://www.ted.com/OpenTranslationProject)에서 한국말로 번역된 내용을 가져와서 한글 요약문을 구성하였다. M대학교에서 문헌정보학을 전공하는 학부 학생 40명(남자: 18명, 여자: 22명)을 피조사자들로 모집하여 실험을 진행하였다.

4.2.3 음향학적 자질에 기반한 요약문 평가

연구 문제 2(음향학적 요소에 기반한 요약의 효율성 평가)를 검증하기 위해서 이전 단계에서 식별력이 있는 것으로 확인된 네 개의 음향학적 요소(F0 최대값, F0 최대값과 F0 최소값간의 차이, 최소값 dB 및 최대값 dB과 최소값 dB간의 차이)에 기반한 요약의 효율성을 평가하였다. 이를 위해서 각 요소에 기반하여 구성된 요약문을 표준 요약문과 비교하여 정확률과 재현율을 대체하는 하나의 척도인 F 측정((2 × 재현율 × 정확률) / (재현율 + 정확률))을 이용하여 유사도를 측정하였다(〈표 3〉 참조).

강도에 속한 최대값 dB과 최소값 dB간의 차이와 최소값 dB의 F값(0.17, 0.15)이 피치에 속한 F0 최대값과 F0 최대값과 F0 최소값간의 차이의 F값(0.10, 0.09)보다 상대적으로 더 높게 나와 강도가 피치보다 좀 더 높은 품질의 요약문을 구성하는 것으로 나타났다. 또한 이 네 가지 요소 중에서 최대값 dB과 최소값 dB간의 차이를 이용하는 것이 가장 효율적인 것으로 나타났다. 이러한 결과는 강도가 스피치에서 단어 중요도를 식별하는 가장 유용한 자질이라고 기술한 Wang과 Narayanan(2007)의 주장과 일치하는 결과이다.

〈표 3〉 음향학적 자질 방식의 평가

방 법	F 측정
강도(최대값 dB과 최소값 dB간의 차이)	0.17
강도(최소값 dB)	0.15
피치(F0 최대값)	0.10
피치(F0 최대값과 F0 최소값간의 차이)	0.09

4.2.4 음향학적 자질 방식과 본문 키워드 방식간의 비교

연구 문제 3(가장 유용한 음향학적 자질 방식과 전통적인 본문 키워드 방식간의 차이를 요약문의 품질과 이 두 방식에 의해서 각 세그먼트에 할당된 가중치간의 상관 관계 측면에서 분석)을 테스트하기 위해서 앞 단계에서 가장 효율성이 높은 것으로 나타난 최대값 dB과 최소값 dB간의 차이를 이용한 방식과 본문 키워드 방식을 표준 요약문을 기준으로 하여 비교하였다. 비교 결과, 본문 키워드 방식의 F값(0.20)이 최대값 dB과 최소값 dB의 차이를 이용한 방식의 F값(0.17)보다 높게 나타났으나 t-검증 결과 통계적으로 유의미한 차이는 없었다($p(=0.55) > 0.05$)(〈표 4〉 참조). 이는 스피치 요약을 하는데 있어서 음향학적 기법이 어휘적 기법과 유사한 성능을 갖고 있다고 할 수 있다.

또한 최대값 dB과 최소값 dB간의 차이에 의해서 구성된 세그먼트 가중치와 본문 텍스트 기반 방식에 의해서 구성된 세그먼트 가중치간에 강한 상관 관계가 있을 것으로 가정하고, 이 두 기법간의 상관도를 측정하였다. 측정 결과, 기대

와 달리 이 두 기법간의 상관도는 0.02(r=0.02)로 매우 낮게 나타났다. 이외에 F0 최대값에 의한 세그먼트 가중치와 본문 텍스트 기반 방식에 의한 세그먼트 가중치간에 상관도도 0.11(r=0.11)로 낮게 나타났다. 이는 음향학적 자질 기법은 어휘적 기법(키워드 방식)과는 다른 방식으로 핵심 세그먼트를 추출한다는 것을 의미한다.

〈표 4〉 음향학적 자질 방식과 본문 키워드 방식의 비교

방 법	F 측정
음향학적 자질 방식: 최대값 dB과 최소값 dB 간의 차이	0.17
본문 키워드 방식	0.20

4.2.5 피치와 강도의 상관 관계 측정

연구 문제 4(피치와 강도간의 상관 관계 분석 및 피치내 또는 강도내의 여러 요소간의 상관 관계 분석)를 조사해 보기 위해서 먼저 강도(최대값 dB과 최소값 dB간의 차이)와 피치(F0 최대값과 F0 최소값간의 차이)간의 상관도를 측정한 결과, 0.09(r=0.09)라는 매우 낮은 상관계수가 산출되었다(〈표 5〉 참조). 따라서 강도와 피치가 서로 다른 기준에 의해서 핵심 세그먼트를 추출한다는 것을 유추해 볼 수 있다. 앞으로, 피치의 변화가 주제 전환과 관련이 있는지 확인하는 것을 포함하여, 강도 또는 피치를 이용한 기법이 핵심 세그먼트를 어떤 기준 또는 패턴에 의해서 추출하는지 빅 데이터(big data)를 사용한 체계적인 분석이 필요해 보인다.

이외에 피치내 요소(F0 최대값과 F0 최소값간의 차이 vs. F0 최대값)간의 상관도는 0.97로 나타나 매우 높은 양의 상관 관계가 있고, 강도내의 요소(최대값 dB과 최소값 dB간의 차이 vs. 최소값 dB)간의 상관도는 -0.78로 나타나 비교적 높은 음의 상관 관계가 있음이 확인되었다. 따라서 피치값 또는 강도값을 사용할 경우에 이들의 모든 요소들을 사용하기보다는 대표적인 하나의 요소만을 사용하여 분석해도 큰 문제는 없어 보인다.

〈표 5〉 피치와 강도, 피치내의 요소들 및 강도내의 요소들의 상관 관계 분석 결과

관 계	상관계수
최대값 dB과 최소값 dB간의 차이(강도) vs. F0 최대값과 F0 최소값간의 차이(피치)	0.09
F0 최대값과 F0 최소값간의 차이(피치) vs. F0 최대값(피치)	0.97
최대값 dB과 최소값 dB간의 차이(강도) vs. 최소값 dB(강도)	-0.78

4.2.6 스피치 요약의 표현 형식에 대한 이용자 평가

연구 문제 5(스피치 요약의 표현 형식 평가)를 조사하기 위해서 피조사자 40명에게 실험 시스템에 저장된 10개 표본 비디오들의 텍스트와 오디오 형태의 요약문들을 브라우징하거나 청취하게 한 후 이 두 형태의 요약문의 장단점을 한 가지 이상 자유롭게 기술하도록 하였다. 조사한 결과는 다음과 같다. 16명(40%)은 오디오 요약을 듣는 것이 편안하다고 하였고, 13명(32.5%)은 오디오 요약에 집중하기가 용이하다고 답하였다. 7명(17.5%)은 오디오 요약이 일정하면서 신속하게 표현된다고 하였다. 5명(12.5%)은 오디오 요약을 듣는 동안에 멀티 태스킹 작업이 쉬워진다고 하였다. 즉, 오디오를 듣는 동안에 관련된 이미지를 보는 것과 같은 다른 작업을 대체로 용이하게 할 수 있다고 답

하였다. 4명(10%)은 오디오 요약에는 음향학/음율적인 정보를 포함하기 때문에 비디오 내용을 더 쉽게 파악할 수 있을 것 같다고 답하였다. 만약 실험 데이터로 음성 합성기로 구성한 오디오 요약 대신에 실제 비디오 스트림에서 오디오를 바로 추출하여 구성한 오디오 요약을 사용하였다면 이 항목에 대한 답변 비율이 더 높아졌을 것으로 생각된다.

반면에, 17명(42.5%)은 오디오 요약은 다시 듣고 싶은 부분으로 가고자 할 때 불편함이 있다고 말했고, 7명(17.5%)은 길이가 긴 오디오 요약은 집중하기가 어려웠다고 답하였다. 이외에 5명(12.5%)은 오디오 요약은 금방 사라지기 때문에 전체 내용을 기억하기가 어렵다고 답하였다.

텍스트 요약은 앞에서 오디오 요약의 단점으로 지적된 점들이 장점으로 언급되었다. 18명(45%)이 다시 읽어 보고 싶은 곳으로 이동하는 것이 용이하다고 하였고, 10명(25%)이 문장간의 문맥을 통해서 비디오의 전체적인 내용 파악을 쉽게 할 수 있다고 답하였다. 이외에 5명(12.5%)은 알려지지 않은 단어 또는 구를 이해하기가 용이했다고 답하였다. 반면에, 14명(35%)은 긴 텍스트 요약은 집중하기가 어려웠다고 답하였다. 6명(15%)은 텍스트 요약은 눈의 피로를 가져다 주었으며, 5명(12.5%)은 텍스트 요약을 적절히 브라우징하기 위해서는 오디오 요약과 달리 상당한 크기의 가상 공간이 필요하다고 답하였다.

이와 같이 오디오 또는 텍스트 형태의 스피치 요약은 고유한 특성을 갖고 있기 때문에 이용자 그룹의 선호도나 시스템 개발 환경에 따라서 적절한 표현 형식이 선택될 수 있다고 생각한다. 예를 들어서, 오디오 요약은 스마트폰과 같은 좁은 가상 공간을 갖고 있는 장치에 효과적으로 적용될 수 있을 것이다. 또한 오디오 요약은 텍스트 요약에 비해서 멀티 태스킹 작업이 더 용이하기 때문에 오디오 정보와 비주얼 이미지를 함께 표현해야 하는 경우에 더 유리하다고 할 수 있다. 다른 한편 텍스트 요약은 요약을 여러 번 반복적으로 브라우징해야 하는 경우 또는 문장간의 문맥을 통해서 비디오의 전체적인 의미 파악이 용이한 경우에 더 적합하다고 생각된다.

5. 결 론

본 연구는 스피치의 속도, 피치 및 강도의 세 가지 음향학적 자질을 이용하여 스피치 요약을 생성할 수 있는지 그리고 이 중 가장 효율적으로 이용할 수 있는 요인이 무엇인지 조사해 보았다. 그런 다음 이러한 조사 결과를 기초로 하여 효율적인 스피치 요약 방안을 구성하여 평가해 보았다. 또한 추출된 스피치의 핵심 세그먼트를 오디오 또는 텍스트 형태로 표현했을 때 어떤 특성이 있는지 이용자 관점에서 분석해 봄으로써 스피치 요약을 효율적으로 추출하여 표현하는 방안을 제안하였다. 핵심적인 연구 내용과 결과를 기술하면 다음과 같다.

첫째, 스피치 속도, 3개의 피치값 통계 및 3개의 강도값 통계를 추출하여 분석한 결과, 네 개 요소(F0 최대값, F0 최대값과 F0 최소값간의 차이, 최소값 dB 및 최대값 dB과 최소값 dB간의 차이)만이 세그먼트간의 차이를 구별해 주는 식별력이 있는 것으로 나타나 이 요소들을 이용하여 스피치 요약을 구성하였다.

둘째, 네 가지 요소에 기반한 요약문의 평가 결과, 최대값 dB과 최소값 dB간의 차이를 이용하는 것이 가장 효율적인 것으로 나타났다. 음향학적 자질 방식과 본문 키워드 방식을 요약문 품질의 측면에서 비교해 본 결과, 본문 키워드 방식의 F값(0.20)이 최대값 dB과 최소값 dB의 차이를 이용한 방식의 F값(0.17)보다 높게 나타났으나 통계적으로 유의미한 차이는 없었다. 따라서, 음향학적 기법이 방대한 비디오 스피치 자막의 처리가 필요한 어휘적 기법과 유사한 성능을 갖고 있다고 할 수 있다.

셋째, 음향학적 자질 방식과 본문 키워드 방식에 의해서 각 세그먼트에 할당된 가중치간의 상관도를 측정한 결과, 최대값 dB과 최소값 dB간의 차이에 의한 세그먼트 가중치와 본문 텍스트 기반 방식에 의한 세그먼트 가중치간에 상관도는 0.02로 매우 낮게 나타났다. 이는 음향학적 기법은 어휘적 기법(키워드 방식)과는 다른 방식으로 핵심 세그먼트를 추출한다는 것을 의미하는 것으로 앞으로 좀 더 체계적인 분석 작업이 필요해 보인다.

넷째, 음향학적 자질에 속한 강도(최대값 dB과 최소값 dB간의 차이)와 피치(F0 최대값과 F0 최소값간의 차이)간의 상관도를 측정한 결과, 0.09라는 매우 낮은 상관도가 산출되어, 강도와 피치도 서로 다른 기준에 의해서 핵심 세그먼트를 추출한다는 것을 확인하였다. 따라서, 앞으로 강도 또는 피치를 이용한 기법이 핵심 세그먼트를 어떤 기준 또는 패턴에 의해서 추출하는지 빅 데이터를 사용한 체계적인 분석이 필요해 보인다.

다섯째, 스피치 요약의 표현 형식(오디오 또는 텍스트 형태)은 각자 독자적인 특성을 가지고 있기 때문에 이용자 그룹의 선호도나 시스템 개발 환경에 따라서 적절히 선택될 수 있음을 확인하였다.

이와 같이, 본 연구는 영어 스피치 자료를 표본 자료로 사용하였다. 앞으로 한국어 스피치 자료를 대상으로 음향학적 자질을 분석하여 영어와 한국어의 스피치에서 스피치 속도, 강도 및 피치에서 어떤 차이를 보이는지 서로 비교하여 분석해 보는 것이 의의가 있을 것으로 생각된다. 본 연구의 분석 결과가 이용자의 선호도 또는 시스템 개발 환경을 고려한 강의 스피치 자료의 브라우징과 검색 시스템을 설계하는 단계에서 기초 자료로 활용될 수 있기를 바란다.

참 고 문 헌

김현희 (2011). 비디오 의미 파악을 위한 멀티미디어 요약의 비동시적 오디오와 이미지 정보간의 상호작용 효과 연구. 한국문헌정보학회지, 45(2), 97-118.
http://dx.doi.org/10.4275/KSLIS.2011.45.2.097

정영미 (2007). 정보검색연구. 서울: 구미무역출판부.

Boersma, P., & Weenink, D. (2006). Praat: Doing phonetics by computer. Retrieved from

http://www.praat.org/

Cawkell, A. (1995). A guide to image processing and picture management. Aldershot, Hampshire: Gower Publishing Ltd.

Chen, B., & Lin, S. (2012). A risk-aware modeling framework for speech summarization. IEEE Transactions on Audio, Speech, and Language Processing, 20(1), 211-222. http://dx.doi.org/10.1109/TASL.2011.2159596

Ding, W., Marchionini, G., & Soergel, D. (1999). Multimodal surrogates for video browsing. Proceedings of the Fourth ACM conference on Digital Libraries, 85-93.

Fujii, Y., Yamamoto, K., Kitaoka, N. & Nakagawa, S. (2008). Class lecture summarization taking into account consecutiveness of important sentences. Proceedings of Interspeech, 2438-2441.

Furui, S., Kikuchi, T., Shinnaka, Y., & Hori, C. (2004). Speech-to-text and speech-to-speech summarization of spontaneous speech. IEEE Transactions on Speech Audio Process, 12(4), 401-408. http://dx.doi.org/10.1109/TSA.2004.828699

Hirschberg, J., & Nakatani, C. (1996). A prosodic analysis of discourse segments in direction-given monologues. Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, 286-293.

Lin, S., Chen, B., & Wang, H. (2009). A comparative study of probabilistic ranking models for Chinese spoken document summarization. ACM Transactions on Asian Language Information Processing, 8(1), 1-23. http://dx.doi.org/10.1145/1482343.1482346

Liu, Y., & Hakkani-Tür, D. (2011). Speech summarization. In G. Tur & R. De Mori (Eds.), Spoken language understanding: Systems for extracting semantic information from speech (pp. 357-392). Chichester, UK: John Wiley & Sons, Ltd.

Maskey, S. (2008). Automatic broadcast news speech summarization. Unpublished doctoral dissertation, Columbia University.

Maskey, S., & Hirschberg, J. (2005). Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. Proceedings of Interspeech, 621-624.

Maskey, S., & Hirschberg, J. (2006). Summarizing speech without text using Hidden Markov Models. Proceedings of the Human Language Technology Conference of the NAACL (Companion Volume: Short Papers), Association for Computational Linguistics, 89-92. Retrieved from http://acl.ldc.upenn.edu/N/N06/N06-2023.pdf

Marchionini, G., Song, Y., & Farrell, R. (2009). Multimedia surrogates for video gisting: Toward combining spoken words and imagery. Information Processing and Management, 45(6), 615-630. http://dx.doi.org/10.1016/j.ipm.2009.05.007

Murray, G., Renals, S., & Carletta, J. (2005). Extractive summarization of meeting recordings. Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH), 593-596. Retrieved from http://www.cstr.ed.ac.uk/downloads/publications/2005/murray-eurospeech05.pdf

Turner, J. (1994). Determining the subject content of still and moving documents for storage and retrieval: An experimental investigation. Unpublished doctoral dissertation, University of Toronto.

Turney, P. (2000). Learning algorithms for keyphrase extraction. Information Retrieval, 2(4), 303-336. Retrieved from http://www.extractor.com/IR2000.pdf

van Houten, Y., Oltmans, E., & van Setten, M. (2000). Video browsing and summarization (Rep. No. TI/RS/2000/63). Enschede: Telematica Instituut. Retrieved from https://doc.telin.nl/dscgi/ds.py/Get/File-12409/

Wang, D., & Narayanan, S. (2007). An acoustic measure for word prominence in spontaneous speech. IEEE Transactions on Audio, Speech, and Language Processing, 15(2), 690-701. http://dx.doi.org/10.1109/TASL.2006.881703

Xie, S., Hakkani-Tur, D., Favre, B., & Liu, Y. (2009). Integrating prosodic features in extractive meeting summarization. Proceedings of the 11th Biannual IEEE Workshop on Automatic Speech Recognition and Understanding, 387-391. Retrieved from http://www.hlt.utdallas.edu/~shasha/papers/ASRU2009_xie.pdf

Zhang, J., & Fung, P. (2007). Speech summarization without lexical features for Mandarin broadcast news. Proceedings of NAACL HLT (Companion Volume), 213-216.

Zhang, Z., & Fung, P. (2012). Active learning with semi-automatic annotation for extractive speech summarization. ACM Transactions on Speech and Language Processing, 8(4), 1-25. http://dx.doi.org/10.1145/2093153.2093155

Zhang, J., Chan, H., & Fung, P. (2007). Improving lecture speech summarization using rhetorical information. Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, 195-200.

Zhang, J., Chan, H., Fung, P., & Cao, L. (2007). A comparative study on speech summarization of broadcast news and lecture speech. Proceedings of the annual conference of the international speech communication association, 2781-2784.

Zhu, X., Penn, G., & Rudzicz, F. (2009). Summarizing multiple spoken documents: Finding evidence from untranscribed audio. Proceedings of ACL/AFNLP, 549-557. Retrieved from http://www.aclweb.org/anthology-new/P/P09/P09-1062.pdf

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Kim, Hyun-Hee (2011). A study on the interactive effect of spoken words and imagery not synchronized in multimedia surrogates for video gisting. Journal of the Korean Society for Library and Information Science, 45(2), 97-118.
http://dx.doi.org/10.4275/KSLIS.2011.45.2.097

Chung, Young Mee (2007). Information retrieval research. Seoul: Gumi Trading Publisher.