보조 자료와 음성 전사를 사용한 강의 검색 시스템

이동현* 이근배
포항공과 대학교 컴퓨터 공학과
{ semko, gblee }@postech.ac.kr

A LECTURE SEARCH SYSTEM USING RELEVANT INFORMATION AND SPEECH TRANSCRIPTION

Donghyeon Lee* Gary Geunbae Lee
Department of Computer Science and Engineering
Pohang University of Science and Technology

요 약

음성 오디오 검색 시스템을 구축하기 위해서는 몇 가지 과정이 필요하다. 첫 번째 과정이 음성 인식기를 이용하여 음성 오디오를 텍스트 형태로 표현하는 것이다. 하지만, 음성 인식기에서 수반되는 음성 인식 오류를 피할 수는 없다. 음성 인식 오류를 최소화하기 위해서 음성 인식 출력의 lattice를 색인(index)해야 하는데, 보다 효과적인 처리를 위하여 압축된 형태를 사용한다. 본 연구에서는 특별히 한국어 강의를 대상으로 검색 시스템을 구축했다. 강의에서는 특별히 관련된 자료를 쉽게 구할 수 있는 데, 이런 자료를 언어 모델에 이용하여 음성 인식 성능을 향상 시킬 수 있다. 또한, 강의 자료를 이용한 추가 색인 테이블(index table)을 생성하여 검색 성능 향상에 도움을 준다. 실험에서 고등학교 과정 수학 강의 동영상을 이용하여 자동화된 강의 검색 시스템을 구축하고, 보조 자료를 이용해 성능을 향상 시키는 것을 보인다.

1. 서 론

최근에 들어, 컴퓨터 파워와 네트워크 대역폭이 향상됨과 동시에 저장 공간의 가격이 하락함에 따라 웹상에서의 멀티미디어 데이터들은 급격하기 증가하고 있다. 하지만, 이런 자료들은 기존의 텍스트 중심의 검색 기술을 적용하기에는 큰 거리감이 있다. 멀티미디어 데이터 중 특히 음성 정보는 중요한 역할을 하고 있는 데, 이에 따라 음성 문서 검색(Spoken Document Retrieval, SDR) 기술에 대한 연구도 활발히 진행되고 있다.

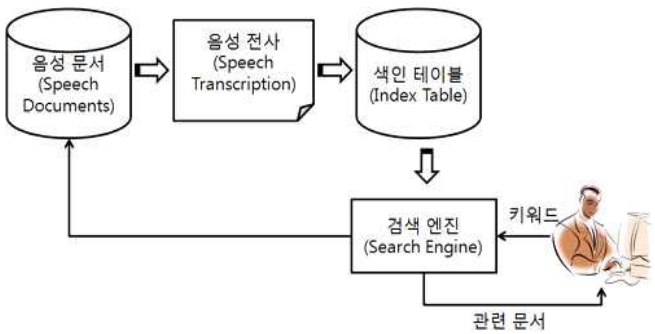


그림 1 음성 문서 검색의 과정

그림 1 은 음성 문서 검색의 전반적인 과정을 보여주고 있다. 음성 문서 검색은 크게 3가지 부분으로 나눌 수 있는데, 첫째는 음성 웨이브를 텍스트로 표현하기 위한 음성 전사(speech transcription) 단계, 둘째는 음성 전사된 결과를 이용해 음성 문서를 색인(index)하는 단계, 셋째는 색인 테이블을 참고하여 사용자가 원하는 키워드에 맞추어 연관된 문서를 보여주는 단계이다.

음성 문서 검색에 대한 연구가 진행됨에 따라, 특정 음성 문서 그룹에 대한 실험과 시스템 개발이 이루어졌다. 대부분의 경우 방송 뉴스를 대상으로 이루어졌으며, 강의 비디오 데이터를 바탕으로 한 경우도 있었다. 예를 들면, TREC (Text REtrieval Conference) Spoken Document Retrieval evaluation[1]에서 방송 뉴스를 이용하였고, MIT Lecture browser[2]에서 MIT의 대학 강의를 이용하였다.

음성 문서 그룹을 방송 뉴스, 강의, UCC(User Created Contents) 등 여러 가지로 나누어 볼 수 있는 데, 몇몇 부분에서 특성의 차이점을 가지고 있다.

대부분의 음성 문서 검색 시스템이 다루고 있는 방송 뉴스의 경우를 보면, 우선 방대한 양의 데이터가 있다. 최근 방송사들은 대부분의 주요 뉴스 클립을 뿐만 아니라 방송에서 사용된 대본까지도 온라인상으로 제공해준다. 하지만, 이 대본을 이용하여 텍스트 기반의 검색 엔진을 바탕으로 방송 뉴스 검색을 수행하기도 한다. 음성 인식 관점에서 방송 뉴스는 비교적 문어체적인 특성과 훈련된 발음으로 인해 좋은 성능을 가져다 줄 수 있지만, 새로운 단어들이 지속적으로 추가됨으로서 미등록어 문제도 앉고 있다. 또, 방송 뉴스의 경우 텍스트 뉴스를

* "본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음" (IITA-2008-C1090-0801-0045)

이용하여 성능을 향상 시키는 데 도움을 줄 수 있다.

강의의 경우, 사람이 직접 받아쓰기를 하지 않으면 대본을 구하기 어려운 점 때문에 대량의 코퍼스를 수집하는 데 어려움이 있다. 대부분의 온라인 교육 사이트의 경우, 색인 과정에서 수작업으로 하고 있다. 음성 인식 관점에서 강의는 대화체적인 특성을 가지고 있으며, 음성 환경도 방송 뉴스에 비해 좋지 못하다. 하지만, 상대적으로 제한된 어휘를 사용한다는 장점이 있다. 또, 강의는 관련된 보조 자료를 이용하여 성능을 향상 시킬 수 있다. 최근에 주목 받고 있는 UCC에서도 음성 정보를 포함하기도 하는 데, 어휘 선택 및 음성 환경 등 여러 가지 측면에서 앞선 두 가지 경우 보다는 어려운 부분이 많다.

본 연구에서는 보조 자료와 음성 전사를 이용한 강의 검색 시스템을 개발하였다. 강의 보조 자료는 크게 두 가지 관점에서 사용하였다. 우선, 음성 인식 성능을 향상 시키기 위한 언어 모델링에 강의 보조 자료를 이용한다. 그리고 검색 성능 향상을 위해, 보조 자료를 바탕으로 색인 테이블을 추가적으로 생성한다.

본 논문은 다음과 같은 차례로 구성되어 있다. 2절에서는 음성 인식과 관련된 부분에 대해 소개하고, 3절에서는 음성 인식의 출력과 보조 자료를 바탕으로 색인을 어떻게 했는지 설명하고, 4절에서는 색인 테이블을 바탕으로 검색 과정을 논하고, 5절에서는 고등학교 수학 강의 데이터를 바탕으로 한 실험의 결과에 대해 기술하고 분석하도록 한다.

2. 음성 인식

2.1 음성 문서 검색과 음성 인식 성능

음성 문서 검색에 있어서 음성 인식기의 성능은 매우 중요하다. 실제로 TREC SDR evaluation에서는 10~20% 단어 오인식율(Word Error Rate; WER)의 음성 인식 환경에서 성공적인 음성 문서 검색을 수행했다고 평가했다. 그 이유는 중요한 단어는 반복되는 경향이 있고, 의미적으로 도움을 주는 단어들도 자주 나타나기 때문이다. 즉, 음성 인식 성능이 일정 수준 이상만 되면 검색 성능에는 큰 악영향을 주지 않았다. 하지만, 실제 음성 인식 환경에서는 단어 오인식율이 30%에서 심지어 50%까지 이른다. 따라서, 음성 문서 검색에 앞서 음성 인식 성능을 높이기 위한 노력이 필요하다.

2.2 음성 인식 시스템

본 연구에서 사용한 한국어 연속 음성 인식 시스템은 HTK(Hidden Markov Model Toolkit)[3]를 기반으로 하였다. 이 시스템은 발성 화자에 관계없는 음성 인식을 하는 화자 독립 시스템이다. 인식 단위로는 음소 기반의 유사음소단위(PLU, Phoneme Like Unit)를 사용하며, 48개로 구성된 유사음소단위 세트를 이용하였다.

음성 인식 시스템에서는 크게 음향 모델(Acoustic Model), 발음 모델(Pronunciation Model), 언어 모델(Language Model) 등 3가지 모델을 사용한다. 음향 모델은 상태 공유의 연속 히든 마코프 모델을 사용하였다. 모든 히든 마코프 모델은 각각 세 개의 상태를 가지고, 각 상태의 출력 확률 값은 다수의 가우시안 혼합분포로부터 계산된다. 발음 모델은 연관 규칙(Association Rules)을 사용한 자소열-음소열 변환기(Grapheme to Phoneme Converter; G2P)[4]를 이용하여 생성했다. 언어 모델은 SRILM toolkit[5]을 이용하여 바이그램(bigram) 모델을 기본적으로 적용했고, 트라이그램(trigram) 모델도 점수 재조정(re-scoring)에 이용하였다.

음향 모델, 발음 모델, 언어 모델을 이용하여 네트워크를 생성한 뒤 비터비(viterbi) 알고리즘을 이용하여 탐색을 수행한다. 음성 인식의 최종 출력은 크게 1-best와 lattice로 나누어 볼 수 있다. 간편하게 1-best를 사용할 수 있지만, 음성 인식 성능이 부족할 경우 문제가 된다. lattice는 상대적으로 복잡한 구조를 가지고 있지만, 잘 활용한다면 음성 인식 성능의 부족함을 조금은 만회할 수 있다. 실제로 단어 오인식율이 상당히 높은 응용에서는 lattice를 많이 활용하며, 음성 문서 검색에 있어서도 기본적으로 lattice를 바탕으로 색인 과정을 거친다.

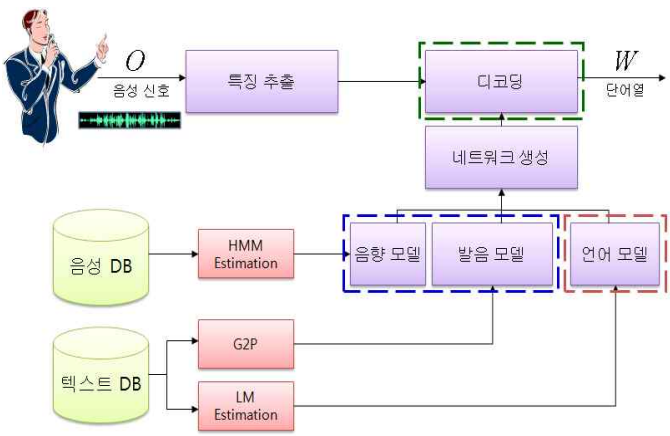


그림 2 음성 인식 시스템의 구조

2.3 음성 문서 검색을 위한 음성 인식 모델링

음성 인식기 디코딩 및 음향 모델 훈련에 앞서 음성 문서에 대한 전처리가 필요하다. 강의와 같은 음성 파일은 길이가 1시간 정도 되는데, 효과적인 음성 인식을 위해서 문장 단위로 잘라내는 것이 필요하다. 여기서는 음성 툴인 Praat[6]을 이용하여 잘라낸 뒤 수작업을 거쳐 완성 했다.

음성 인식이 가지는 기본적인 한계점은 훈련한 어휘에서 벗어난 단어에 대해서는 인식을 하지 못하고 다른 단어들로 대체가 된다는 것이다. 음성 인식 성능을 향상시키기 위해서 어휘 선택과 언어 모델링은 중요한 역할을 한다. 기본적으로 언어 모델링 적응(adaptation) 기술을 사용하는 데, 일반적인 언어 모델에 도메인에 특화된 언어 모델을 결합한 형태를 가진다.

본 연구에서 개발한 시스템의 경우 고등학교 수학 강의를 검색하기 때문에, 관련된 보조 자료를 활용하면 어휘 선택과 언어 모델링에 큰 도움이 될 수 있다. 강의에

서는 주로 대화체의 형태로 발화가 이루어지기 때문에 기본적으로 언어 모델링을 대화체에 맞추어야 한다. 여기서는 강의에 사용된 대화체 내용을 기본으로 하여, 강의 내용에 맞는 보조 자료 내용을 결합하여 언어 모델링을 했다.

3. 음성 문서 색인

3.1 음성 문서의 분할

음성 문서는 강의 비디오 파일에서 오디오 데이터를 추출하여 만들어졌다. 그 이후 음성 인식기를 거쳐 오디오 신호는 텍스트로 전사되는 과정을 거쳤다. 이렇게 생성된 음성 전사로부터 음성 문서를 색인하고, 색인 테이블로부터 검색을 수행한다. 사용자는 음성 검색 시스템에 키워드를 입력하면, 시스템은 관련된 음성 문서를 사용자에게 제공해준다.

하지만, 강의 오디오의 경우에는 음성 문서가 너무 긴 경향을 가지고 있다. 예를 들어 '지수법칙'이라는 키워드를 입력했을 때, '로그'에 대해 전반적으로 다룬 음성 문서가 제공될 것이다. 이러한 불편함을 최소화하기 위해 음성 문서를 분리하는 것이 도움이 된다. 이런 과정은 토픽 분할 기술(Topic Segmentation)을 통해서 진행 될 수 있지만, 여기에서는 수작업을 통해서 1시간 분량의 강의 오디오를 좀 더 짧은 단위로 분할했다. 즉, 사용자가 키워드를 입력했을 때, 시스템은 문서 전체 대신 분할한 단위로 검색 결과를 제공해준다.

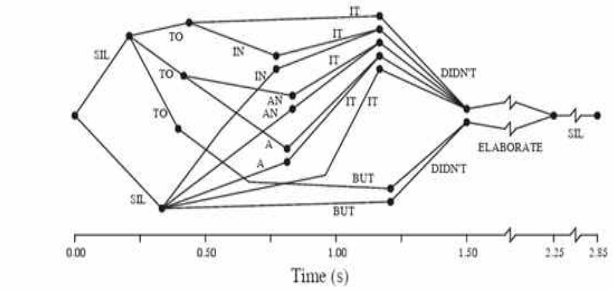
3.2 음성 문서의 색인 테이블 생성

텍스트 색인하는 방법에도 TF-IDF, 확률 모델 등 여러 가지가 있다. 여기에서는 Ad-hoc 모델(일명, early google model)[7]을 사용했다. Ad-hoc 모델은 Hit이라는 개념을 사용하는 데, Hit은 특정 문서에서 특정 단어가 나타난 것을 의미한다. 색인 테이블에는 Hit이 문서에서 발생한 위치와 함께 저장한다.

음성 문서의 색인 과정에서 음성 인식기의 출력으로 생성된 1-best 만을 사용한다면, 텍스트에서 하던 방식과 같다. 하지만 앞서 언급했듯이 lattice를 이용하는 것이 상대적으로 오인식률의 위험을 줄일 수 있다. 워드 그래프의 형태를 가진 lattice는 시간 정보, 스코어 정보, 링크 정보 등을 포함하고 있다. 색인 과정에서 lattice를 그대로 활용하기에는 단어 순서의 불확정성과 근접성 평가의 어려움 때문에 한계를 가지고 있다.

따라서 lattice를 색인 과정에서 처리하기 위해서는 좀 더 간략화한 형태로 변환할 필요가 있다. Chelba는 이러한 목적으로 PSPL(position specific posterior probability) [8]을 제안했는데, PSPL은 위치 정보와 함께 확률 정보를 포함한 구조를 가지고 있다. 즉, lattice에서 특정 단어가 특정 위치에 어떤 확률로 나타나는지를 나타낸다.

기존 텍스트 기반의 ad-hoc 모델에서는 주어진 단어에 대해 해당 문서와 발생한 위치가 색인 테이블에 저장되었는데, PSPL을 이용할 경우 색인 테이블에 추가적으로 확률을 포함하면 된다.



0	1		2		3		4	
SIL	TO	0.5	IT	0.2	DIDN'T	0.5	ELABORATE	0.7
	IN	0.3	IN	0.2	IT	0.4	DIDN'T	0.2
	AN	0.1	A	0.1	ELABORATE	0.1	SIL	0.1
	BUT	0.1	BUT	0.1				
			DIDN'T	0.1				

그림 3 lattice와 PSPL의 예

이 밖에도 비슷한 구조로 CN(confusion network)[9]이 있는데, NULL 워드 처리 문제를 앉고 있고 PSPL과는 달리 CN은 lattice에서 발생한 모든 n-gram을 보장하지 못하는 단점이 있다.

3.3 보조 자료를 이용한 추가 테이블 생성

강의와 관련된 보조 자료는 쉽게 구할 수가 있을 뿐만 아니라, 잘 활용하면 성능 향상에도 도움이 될 수 있다. 보조 자료로부터 컨텐츠 테이블(Contents Table)과 매칭 테이블(Matching Table)을 생성했다. **그림 4**는 이 두 테이블이 생성되는 과정을 보여준다.

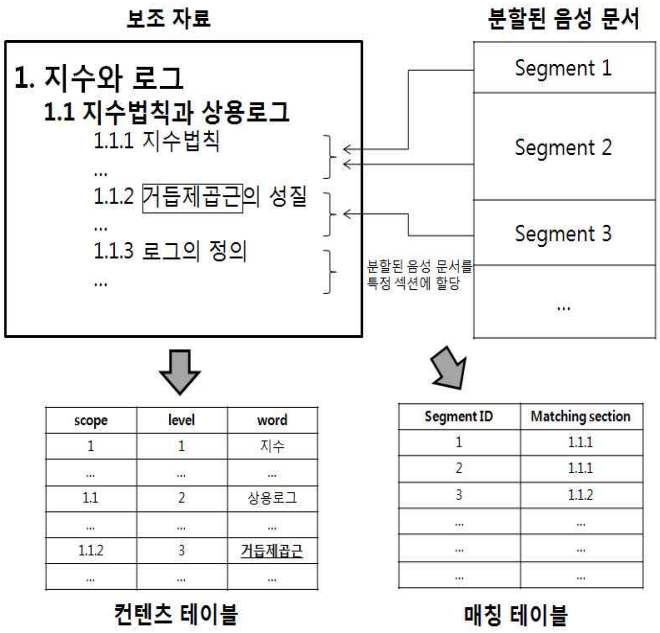


그림 4 컨텐츠 테이블과 매칭 테이블의 생성 과정

컨텐츠 테이블의 엔트리는 scope, level, word로 구성되어 있다. 각각은 보조 자료에서 나타나는 주요 단락의 제목으로부터 쉽게 추출할 수 있다. 즉, 제목으로부터 단어를 추출하고 이와 함께 그 제목이 포함하는 부분에 대한 정보를 저장한다. 예를 들어 '1.1.2 거듭제곱근의 성질' 이라고 표기된 제목에 대해서 각각의 단어 '거듭제곱근' 과 '성질' 은 scope '1.1.2'와 level 3과 함께 컨텐츠 테이블에 저장된다.

매칭 테이블은 특정 섹션에 분할된 음성 문서를 할당해준 뒤 그 정보를 저장한다. 이를 위해서 각 섹션마다 언어 모델을 생성한다. 분할된 음성 문서를 생성한 각각의 언어 모델에 적용해서 가장 높은 언어 모델 점수를 가지는 섹션으로 할당한다.

4. 검색

앞의 과정을 모두 거치고 나면, 음성 전사 색인 테이블, 컨텐츠 테이블, 매칭 테이블 이렇게 세 가지 테이블이 생성된다. 이 테이블을 이용하여 검색 과정을 수행한다. 검색 엔진은 사용자 질의 $Q = q_1 \cdots q_i \cdots q_M$와 음성 문서 D가 주어졌을 때 다음의 두 가지 점수를 이용한다.

첫째는 음성 내용 인덱스 테이블을 이용해서 계산한 점수로 다음과 같이 계산한다.

$$S^T(D, q_i \ldots q_{i+N-1}) = log\left[1 + \sum_{lattice\ t} \sum_{position\ k} \prod_{l=0}^{N-1} P(word_{k+l}(t) = q_{i+l} | D)\right]$$

$$S^T_{N-gram}(D, Q) = \sum_{i=1}^{M-N+1} S^T(D, q_i \ldots q_{i+N-1})$$

$$S^T(D, Q) = \sum_N w_N \cdot S^T_{N-gram}(D, Q)$$

이 점수에서는 사용자 질의에서 모든 가능한 N-gram에 대해 PSPL에서 나온 확률값을 이용하여 계산한다. W_N은 각 N-gram에 대한 가중치를 의미한다.

둘째는 컨텐츠 테이블과 매칭 테이블을 이용해서 계산한 점수로 다음과 같이 계산한다.

$$S^C(D, Q) = log\left[1 + \sum_{i=1}^{M} \sum_{\substack{contents\ c, \\ scope(c) \subset scope(D)}} w_{level(c)} \cdot \delta(word(c), q_i)\right]$$

이 점수에서는 사용자 질의의 단어들과 일치하는 컨텐츠 테이블 엔트리 중 주어진 음성 문서 D의 영역에 포함되는 것만 추출한 뒤 level에 따라 가중치 $W_{level(c)}$를 곱하여 계산한다.

키워드와 문서의 최종 연관 점수는 아래와 같이 계산된다. 여기서의 가중치는 음성 전사로부터 얻은 점수와 보조 자료로부터 얻은 점수를 조정하는 역할을 한다.

$$S(D, Q) = S^T(D, Q) + \lambda \cdot S^C(D, Q)$$

이 점수를 바탕으로 하여 강의 오디오 검색 시스템은 최종적으로 사용자가 검색을 요청한 키워드에 대해서 강의 오디오 모음으로부터 가장 연관된 강의 오디오를 제공해준다.

5. 실험

본 연구에서는 온라인 교육 사이트로부터 추출한 고등학교 과정의 수학 강의 동영상을 추출하여 음성 문서 검색 실험에 사용하였다. 실험 과정에서 사용한 보조 자료 역시 교육 사이트에서 제공한 강의 노트를 이용하였다. 수학 강의 동영상은 총 40개로 20시간 정도의 분량을 가지고 있다. 사용자에게 보다 편리한 형태로 제공하기 위해 강의 동영상은 최종적으로 511개로 분할되었다. 음성 인식기의 단어 오인식률은 1-best에서 38.7%였다.

표 1은 음성 문서의 표현 형태에 따라 필요한 저장 공간의 크기를 나타낸다. 음성 신호는 음성 인식기를 거쳐 1-best와 lattice로 표현되고, lattice로부터 PSPL을 최종적으로 생성했다. PSPL은 1-best에 비해 많은 공간을 차지했으나, lattice에 비해서는 보다 적은 공간을 차지했다.

표 1 음성 문서의 표현 형태에 따른 저장 공간의 크기

	저장 공간
음성 오디오 파일	2.17GB
1-best	821KB
lattice	262MB
PSPL	147MB

음성 문서 검색 성능 실험을 위해 총 50개의 텍스트 질의를 사용했다. 자주 발생하는 n-gram 후보군 중에서 인위적으로 50개를 선택했다. 각 텍스트 질의는 최소 1개의 단어에서 최대 3개의 단어를 포함하고 있다. 모든 텍스트 질의는 음성 인식기의 어휘에 포함된 단어들로 구성되었고, 평균 길이는 1.27이였다.

표 2 음성 문서 검색 실험 결과

	MAP	R-Precision
음성 인식 결과 (PSPL)	0.8657	0.7183
+ 보조 자료	0.8811	0.7526

표 2는 음성 전사만을 이용했을 때와 보조 자료를 추가적으로 이용했을 때의 음성 문서 검색 성능을 나타낸다. 검색 성능의 평가 방법으로는 MAP(Mean Average

Precision)과 R-Precision을 사용했다. 실험 결과를 보면, 보조 자료를 추가적으로 활용했을 경우에 검색 성능도 조금은 향상 되는 것을 확인할 수 있다.

6. 결론 및 향후 연구 계획

본 논문에서는 음성 전사와 보조 자료를 사용하여 강의 검색 시스템을 구현하였다. 일부 온라인 강의 시스템에서도 동영상 검색을 제공하는 데, 이는 대부분 수작업을 통해 색인한 것을 이용한다. 여기서는 음성 전사와 보조 자료를 통해 자동으로 음성 전사 색인 테이블, 컨텐츠 테이블, 매칭 테이블을 생성했다. 세 가지 테이블을 이용해 검색 과정에서 연관 점수를 구했다. 실험 결과에서 보조 자료를 활용하는 것이 음성 문서 검색 성능 향상에 도움이 되는 것을 보였다.

하지만, 본 연구에서는 강인한 음성 문서 검색 시스템에서 필수적인 미등록어(Out-of-Vocabulary;OOV) 처리 부분이 고려되지 않았다. 앞으로는 사용자 질의에 미등록어가 포함되어 있을 경우에도 검색을 효과적으로 수행할 수 있는 방법을 연구할 계획이다. 추가적으로 지금보다 큰 규모의 데이터에 대해서 연구를 확장할 계획이다.

7. 참고 문헌

[1] Garofolo, J., Auzanne, C., Voorhees, E. The TREC Spoken Document Retrieval Track: A Success Story. In *Proceedings of the 8th Text Retrieval Conference*. 1999.

[2] Glass, J., Hazen, T., Cypers, S., Malioutov, I., Huynh, D. and Barzilay R. Recent progress in the MIT Spoken Lecture Processing Project. In *Proceedings of Interspeech*, Antwerp, Belgium. 2007.

[3] Young, S., Evermann, G., Hain, T., Kerwhaw, D., Moore, G., Odell, J., Dave Ollason, D.P., Valtchev, V. and Woodland, P. The HTK Book. Cambridge University Engineering Department, Cambridge, England. 2002.

[4] Jinsik Lee, Seungwon Kim, Gary Geunbae Lee. Grapheme-to-Phoneme Conversion Using Automatically Extracted Associative Rules for Korean TTS System. In *Proceedings of Interspeech-ICSLP*. 2006.

[5] Stolcke, A. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, Denver. 2002.

[6] http://www.fon.hum.uva.nl/praat/

[7] Brin, S. and Page, L. The anatomy of a large-scal hypertextual Web search engine. *Computer Networks and ISDN Systems*, Vol. 30, pp. 107-117. 1998.

[8] Chelba, C. and Acero, A. Indexing uncertainty for spoken document search. In *Proceedings of Eurospeech*, Lisbon, Protugal. 2005.

[9] L. Mangu, E. Brill, A. Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech and Language* 14(4), 373-400. 2000.