대어휘 연속음성 인식 시스템의 성능평가

김 주 곤, 정 현 열
영남대학교 정보통신공학과

Performance Evaluation of Large Vocabulary Continuous Speech Recognition System.

Joo-Gon Kim, Hyun-Yeol Chung
Department of Information & Communication Eng., Yeungnam University
{kjk,chy}@speech.yu.ac.kr

요 약

본 논문에서는 한국어 대어휘 연속음성 인식 시스템의 성능향상을 위하여 Multi-Pass 탐색 방법을 도입하고, 그 유효성을 확인하고자 한다. 연속음성 인식실험을 위하여, 최근 실험용으로 널리 사용되고 있는 HTK와 Multi-Pass 탐색 방법을 이용한 음성인식 시스템의 비교 실험을 수행한다. 대어휘 연속음성 인식 시스템에 사용한 언어 모델은 ARPA 표준 형식의 단어 N-gram 언어모델로, 1-pass에서는 2-gram 언어모델을, 2-pass에서는 역방향 3-gram 언어모델을 이용하여 Multi-Pass 탐색 방법으로 인식을 수행한다. 본 논문에서는 Multi-Pass 탐색 방법을 한국어 연속음성인식에 적합하게 구성한 후, 다양한 한국어 음성 데이터 베이스를 이용하여 인식실험을 수행하였다. 그 결과, 전화망을 통하여 수집된 잡음이 포함된 증권거래용 연속음성 데이터 베이스를 이용한 연속음성 인식실험에서 HTK가 59.50%, Multi-Pass 탐색 방법을 이용한 시스템은 73.31%의 인식성능을 나타내어 HTK를 이용한 연속음성 인식률 보다 약 13%의 인식률 향상을 나타내었다.

I. 서 론

최근 음성인식의 응용분야가 확대됨에 따라 국내에서는 인식 시스템이 다양하게 개발되어 인식 성능 향상을 위한 연구가 활발하게 진행중이다. 음성인식 시스템 구현에 관한 연구의 예를 살펴보면, 대표적인 자연스러운 발성을 대상으로 하는 자연어(Spontaneous Speech) 인식시스템의 예로서는 미국의 경우, CMU (Carnegie Mellon University)의 SPHINX-II[1], JANUS-III[2], 스탠포드대학의 DRAGON Dictate, MIT의 SUMMIT, GALAXY등이 있다.

일본의 대어휘 연속음성인식에 관한 연구의 경우, NTT의 일본경제 신문의 기사를 이용한 연구에 의해 일본어 LVCSR(Large Vocabulary Continuous Speech Recognition) 시스템에 관한 연구가 시작되어 일본어 방송 뉴스 음성을 대상으로 평가 실험을 수행하였다. 또한 1997년 정보처리 진흥 사업협회(IPA)가 일본음향학회에서 구축한 신문기사 낭독체 음성 데이터베이스(ASJ-JNAS)를 이용하여 대어휘 연속음성인식 프로젝트[5,6]가 추진되면서부터 일본어 LVCSR에 관한 연구가 본격적으로 시작되었다[3]. 이 프로젝트의 결과물로서 JULIUS 음성인식 시스템이 사용되고 있다.

국내에서도 1980년도에 접어들면서부터 본격적인 음성인식에 관한 연구가 이루어져 오고 있으며, 현재 한국전자통신연구원(ETRI), 한국과학기술원(KAIST), 한국통신(KT), 그리고 주요 대학 등에서 자체적으로 작성한 연속음성 데이터베이스를 이용하여 연속음성인식에 관한 연구를 수행 중에 있다.

이러한 음성인식 시스템들은 고립단어인식에 대해서는 약간의 잡음이 있는 사무실 환경 하에서도 95%

이상의 인식 성능을 가지며, 한정된 태스크 범주내의 연속음성인식에서도 90% 이상의 높은 인식률을 가진 시스템도 속속 개발되고 있으며, 인식 태스크를 확장하기 위한 여러 가지 연구들이 진행되고 있다[4].

본 논문에서는 한국어 대어휘 연속음성 인식실험을 위하여 실험용으로 널리 사용되고 있는 HTK[7]와 Multi-Pass 탐색 방법[5,6]을 이용한 음성인식 시스템의 비교 실험을 수행한다. 이를 위하여 두 시스템을 한국어 연속음성인식에 적합하게 구성한 뒤, 다양한 한국어 음성 데이터 베이스를 이용하여 인식실험을 수행하였다.

Ⅱ. 연속음성 인식 시스템

본 논문에서는 한국어 대어휘 연속음성 인식실험을 위하여 공개되어있는 HTK(http://htk.eng.cam.ac.uk/)와 JULIUS 시스템(http://winnie.kuis.kyoto-u.ac.jp/)을 한국어 연속음성인식에 적합하게 구성한 뒤, 다양한 한국어 음성 데이터 베이스를 이용하여 인식실험을 수행하였다. 이들 두 시스템의 가장 큰 차이는 인식 알고리즘이다. HTK에서는 토큰 전파방식의 Viterbi 디코딩 알고리듬을 이용한 1-pass로 이루어진다. 하지만 JULIUS 시스템에서는 1-pass 탐색에서는 단어 2-gram을, 2-pass 탐색에서는 역방향 단어 3-gram을 이용하는 Multi-Pass 탐색 방법을 이용한다. 그림 1에서 JULIUS 인식 시스템의 전체 구성도를 나타낸다.

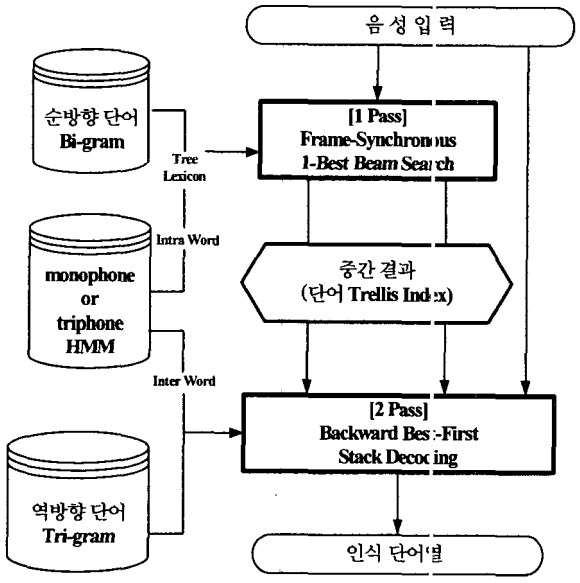

그림 1. JULIUS 시스템의 전체 구성도

2.1 Multi-Pass 탐색 방법

대어휘 연속음성인식은 탐색공간이 매우 크기 때문에, 처음부터 복잡한 언어모델을 사용하면 탐색처리가 복잡하기 때문에 처리량이 증가하게 된다. 따라서, 간단한 음향모델이나 언어모델을 사용하여 일정 개수의 후보들을 구해서 탐색공간을 줄이고, 이 후보들을 이용하여 복잡한 언어모델을 적용하는 방법이 효과적이다. 이를 위해 탐색과정을 여러 개의 pass로 분할하여 간단한 모델에서 정밀한 모델까지 순서대로 적용하는 단계적 탐색 방법을 Multi-Pass 탐색 방법이라고 한다. 이 방법에서는 첫 번째 단계인 1-pass 탐색에서는 단어 2-gram을 이용하여 시간 동기형 빔 탐색을 수행하여, 단어 그래프를 생성한다. 그 다음 단계인 2-pass 탐색에서는 단어 3-gram을 이용하여 단어 그래프로부터 최적의 인식 결과를 얻게 된다. 단어 그래프의 예를 아래 그림에 나타낸다.

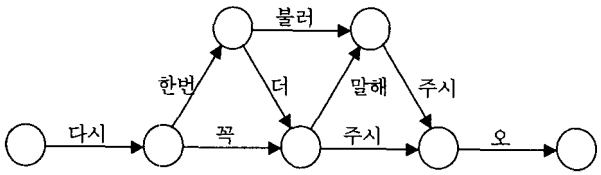

그림 2. 한국어 단어 그래프의 예

JULIUS 인식 시스템에서는 1-pass 탐색에서는 단어 2-gram을, 2-pass 탐색에서는 역방향 단어 3-gram을 이용하는 Multi-Pass 탐색 알고리즘을 이용한다. 1-pass에서는 시스템의 고속화를 위해 프레임 동기형 빔 탐색 알고리즘을 이용하여 목구조 형태의 사전을 생성하고, 각 상태에 2-gram 확률을 동적으로 분할하여 지정한 후 사전에 있는 모든 단어에 대해서 탐색을 수행한다. 2-pass에서는 단어단위의 best-first의 스택 디코딩 탐색을 수행한다. 언어모델은 3-gram을 이용하고 단어단위의 탐색을 이용하는 것은 단어 레벨에서의 제약조건을 다루기 쉽고 정밀한 모델을 적용하는데 용이하기 때문이다. 1-pass에서의 중간 결과로부터 얻은 예측정보를 이용하기 위해 2-pass에서는 1-pass와 반대방향으로 탐색을 수행한다.

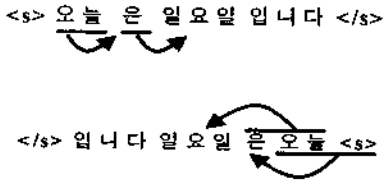

그림 3. 단어 2-gram, 역방향 단어 3-gram의 예

표 2.1 단어 n-gram의 예

2-gram	역방향 3-gram
\2-grams:	\3-grams:
-0.0000 </s> <s>	-0.0001 </s> 같군요 것
-1.9932 <s> 가격	-0.0001 </s> 같습니까 것
-2.9932 <s> 가구	-0.0004 </s> 걸리는군요 더
-2.9932 <s> 가까운	-0.1252 </s> 걸릴까요 얼마나
-2.6921 <s> 가능	-0.6023 </s> 걸릴까요 이
-2.6921 <s> 가능성	-0.0001 </s> 걸립니까 얼마나
-2.2150 <s> 가능한	-0.4774 </s> 걸립니다 은
-2.6921 <s> 가령	-0.1764 </s> 걸립니다 정도
-2.9932 <s> 각	-0.7784 </s> 겁니까 된
...	...

2.2 Unseen Model을 고려한 트라이폰 모델

인식의 기본단위로는 한국어의 음향학적 특징과 대어휘 연속음성인식으로의 확장성을 고려하여 48개의 유사음소단위를 사용하였다. 이것을 트라이폰으로 확장할 경우 학습 데이터 문장에서 생성되는 트라이폰 만을 가지고 음향모델이 만들어진다. 이 경우 인식을 수행할 때 테스트 문장은 학습 문장에 종속된다. 본 논문에서는 학습시 만들어지지 않은 트라이폰 모델을 고려하기 위하여, 48개 유사음소단위에서 생성 가능한 모든 트라이폰 모델에 대하여 학습시 만들어진 모델 중 유사한 것으로 대체하는 방법을 사용하였다.

표 2.2 Unseen 트라이폰 모델 처리의 예

번호	생성가능한 모델	생성된 모델
....		
65550	gz-xx+jz	dz-xx+aa
65551	xi-jj+yv	nf-jj+wv
65552	sz-jj+wa	nf-jj+wv
65553	jz-ss+ee	jj-ss+mf
65554	lf-ss+nz	jj-ss+mf
65555	ch-ss+oo	jj-ss+mf
65556	vv-ch+oo	ss-ch+vv
65557	nf-ch+xx	ss-ch+vv
65558	vv-gf+jj	sil-gf+jj
65559	nf-gf+ss	bf-gf+uu
65560	nz-bz	sil-bz+ya
65561	jj-bz+ng	sil-bz+ya
....		

Ⅲ. 인식 실험 및 고찰

본 논문에서 대어휘 연속음성 인식기의 기본 성능 테스트를 위하여 한국전자통신연구원(ETRI)의 ETRI 445단어 DB를 이용하여 단어 인식 실험을 수행하였다.

ETRI 445 단어 DB는 445종류의 단어로 이루어졌으며 16kHz, 16bit로 저장한 것이다. 화자 독립 단어 인식 실험을 위하여 학습에는 14명의 음성 데이터를 사용하고, 인식에는 학습에 참가하지 않은 5명의 음성 데이터를 사용하였다.

연속음성 인식 실험을 위하여 문장과 단어가 많이 포함된 KAIST 무역상담 DB를 대상으로 인식 실험을 수행하였다. 잡음이 포함되지 않은 KAIST 무역 상담 DB는 남자 100명, 여자 50명이 각각 60 ~ 100 문장을 발성한 것을 16kHz, 16bit로 저장한 것이다. 학습에는 남자 90명의 8790문장(2753단어)을 사용하고, 인식에는 학습에 참가하지 않은 남자 10명의 983문장(1632단어)를 이용하였다.

또한 잡음이 많이 포함된 실 환경의 전화 음성 DB로 증권 거래 음성 DB를 이용하였다. 이 음성 DB는 실제 일반 전화망을 통하여 수집된 Ulaw 방식의 8kHz, 8bit 음성 데이터를 PCM형식의 8kHz, 16bit로 변환하여 저장한 것이다. 학습에 남자 10명(500문장)을 사용하고, 인식에는 학습에 참가하지 않은 남자 5명(250문장)을 대상으로 인식실험을 수행하였다.

음성데이터는 $1-0.97z^{-1}$의 전달함수로 프리엠퍼시스 하였으며, 25ms의 해밍 윈도우를 곱하여 10ms씩 이동하면서 분석하였다. 이를 통해 음성 특징 파라미터는 12차 MFCC와 12차의 delta MFCC, 그리고 정규화된 대수 에너지 성분을 포함하여 총 25차의 특징 파라미터를 구하였다.

3.1 ETRI 445 단어 DB를 이용한 단어 인식 실험

먼저, 본 논문에서 사용된 연속음성인식기의 기본 성능 테스트를 위하여 ETRI 445단어에 대하여 인식 실험을 수행하였다. 음향 모델로는 모노폰과 트라이폰을 이용하였다. 그 결과, 단어 단위의 인식기에서는 N-gram 언어 모델의 효과가 나타나지 않으므로 HTK와 JULIUS 인식 시스템 모두 비슷한 인식 성능을 나타내었다.

표 3.1 ETRI 445 단어 인식률

	HTK	JULIUS
모노폰	90.68	87.28
트라이폰	94.82	94.97

3.2 무역 상담 DB를 이용한 연속인식 실험

대어휘 연속음성 인식 실험을 위하여 문장과 단어가

많이 포함된 KAIST 무역 DB를 대상으로 인식 실험을 수행하였다. HTK에서는 단어 2-gram을 이용하여 1-pass 탐색을 수행하지만 JULIUS 음성인식기는 1-pass 탐색 뿐만 아니라 2-pass 탐색을 수행하므로 연속음성 인식실험에서는 표 3.2에서 나타난 결과에서 HTK를 이용한 경우 보다 JULIUS을 이용한 경우, 더 좋은 인식 성능을 나타내었다.

표 3.2 무역 상담 DB를 이용한 연속음성 인식률

	HTK	JULIUS
단어	80.16	79.92
문장	42.01	49.60

3.3 증권 거래 DB를 이용한 연속인식 실험

마지막으로, 잡음이 많이 포함된 실 환경의 전화 음성 DB인 증권 거래 DB를 이용한 연속음성 인식실험을 수행하였다.

표 3.3 증권 거래 DB를 이용한 연속음성 인식률

	HTK	JULIUS
단어	80.90	84.35
문장	59.50	73.31

그 결과, 표 3.3에서와 같이 JULIUS 음성인식 시스템을 이용한 경우, 단어인식률 84.35%, 문장인식률 73.31%로 비교대상인 HTK를 이용한 문장인식률 보다 약 13%의 인식률 향상을 나타내었다.

Ⅳ. 결론

본 논문에서는 한국어 대어휘 연속음성 인식실험을 위하여 공개되어있는 HTK와 JULIUS 음성인식 시스템을 한국어 연속음성인식에 적합하게 구성한 후, 다양한 한국어 음성 데이터 베이스를 이용하여 인식실험을 수행하였다. 이들 두 시스템의 가장 큰 차이는 인식 알고리즘이다. HTK에서는 토큰 전파방식의 Viterbi 디코딩 알고리듬을 이용한 1-pass로 이루어진다. 하지만 JULIUS시스템에서는 1-pass 탐색에서는 단어 2-gram을, 2-pass 탐색에서는 역방향 단어 3-gram을 이용하는 Multi-Pass 탐색 방법을 이용한다.

그 결과, 전화망을 통하여 수집된 잡음이 포함된 증권거래용 연속음성 데이터 베이스를 이용한 연속음성 인식실험에서 HTK가 59.50%, JULIUS 시스템은 73.31%의 인식성능을 나타내어 HTK를 이용한 문장인식률 보다 JULIUS 시스템을 이용한 경우가 약 13%의 인식률 향상을 나타내었다. 따라서 한국어 대어휘 연속음성 인식 실험을 위하여 JULIUS 인식 시스템의 유효함을 확인하였다.

감사의 글

본 연구는 한국과학재단 목적기초연구(R01-2000-000-00276-0)지원으로 수행되었으며, 실험에 사용된 "KAIST 무역상담 연속어 DB"는 과기원에서 제공받았습니다.

참고 문헌

[1] Alleva, F., et al., "Applying SPHINX-II to the DARPA Wall Street Jounal CSR task," Proc. Speech and Natural Language Workshop, pp. 393-398, 1992.

[2] Alon Lavie, et al, "JANUS-III: Speech-to-speech translation in multiple languages," Proc. IEEE ICASSP-97, Vol.1, pp. 99-102, 1997.

[3] 오세진, "의사 N-gram 모델과 반복학습법에 의한 한국어 대어휘 연속음성인식의 언어처리", 영남대학교 박사논문, 2001

[4] 정현열, "음성인식 연구의 국내외 현황과 전망," 제15회 음성통신 및 신호처리 워크샵 논문집, pp. 23-30, 1998.

[5] A. Lee, T. Kawahara, and S. Doshita, "Large vocabulary continuous speech recognition based on multi-pass search using word trellis index," *Trans. of IEICE*, Vol. J82-D-II, No. 1, pp. 1-9, 1999.

[6] T. Kawahara et. al., "Shareable software repository for Japanese large vocabulary continuous speech recognition," *Proc. of ICSLP'98*, pp. 3257-3260, 1998.

[7] Steve Young, The HTK Book (for HTK Version 3.0, 2000