2000년도 한국음향학회 학술발표대회 논문집 제19권 제1(s)호

음향학적 파라미터의 변화 및 반복학습으로 작성한 언어모델에 대한 고찰

°오세진, 황철준*, 김범국*, 정호열, 정현열
영남대학교 정보통신공학과
*대구과학대학 정보전자통신계열

Language Models constructed by Iterative Learning and Variation of the Acoustical Parameters

°Se-Jin Oh, Cheol-Jun Hwang*, Bum-Koog Kim*, Ho-Youl Jung, Hyun-Yeol Chung
Dept. of Information & Communication Eng., Yeungnam Univ.
*Informational Electronics & Communication Div., Taegu Science College
E-mail : *{osj, hoyoul, chy}@speech.yeungnam.ac.kr*
{hcj, kbk}@electron.taegu-c.ac.kr

요 약

본 연구에서는 연속음성인식 시스템의 성능 향상을 위한 기초 연구로서 시스템에 적합한 음향모델과 언어모델을 작성하고 항공편 예약 태스크를 대상으로 인식실험을 실시한 결과 그 유효성을 확인하였다. 이를 위하여 먼저 HMM의 출력확률분포의 mixture와 파라미터의 차원에 대한 정확한 분석을 통한 음향모델을 작성하였다. 또한 반복학습법으로 특정 태스크를 대상으로 N-gram 언어모델을 적용하여 인식 시스템에 적합한 모델을 작성하였다. 인식실험에 있어서는 3인의 화자가 발성한 200문장에 대해 파라미터 차원 및 mixture의 변화에 따른 음향모델과 반복학습에 의해 작성한 언어모델에 대해 multi-pass 탐색 알고리즘을 이용하였다. 그 결과, 25차원에 대한 mixture 수가 9인 음향모델과 10회 반복학습한 언어모델을 이용한 경우 평균 81.0%의 인식률을 얻었으며, 38차원에 대한 mixture 수가 9인 음향모델과 10회 반복학습한 언어모델을 이용한 경우 평균 90.2%의 인식률을 보여 인식률 제고를 위해서는 38차원에 대한 mixture 수가 9인 음향모델과 10회 반복학습으로 작성한 언어모델을 이용한 경우가 매우 효과적임을 알 수 있었다.

1. 서 론

최근 고도 정보화 사회를 맞이하여 첨단 통신기술의 발전과 더불어 음성인식기술에 대한 연구가 더욱 활발하게 진행되고 있으며, 다양한 분야에 실용화된 음성인식기가 출현하고 있다. 특히 소규모 음성인식기를 비롯한 대량의 음성 데이터를 사용하여 통계적 모델을 작성하는 기술 확립에 의해서 다수화자에 대한 음성인식 성능향상과 대어휘를 대상으로 한 특정화자 인식기술이 실용화 단계에 있으며, 대어휘를 대상으로 한 연속음성인식에 관한 연구도 활발히 진행되고 있다. 이러한 연구의 예를 살펴보면, 증권 · 철도 · 항공안내 시스템, 인터넷 검색 및 가전제품 등이 있으며 이들 시스템의 경우 적용범위가 좁고 특정 태스크에서 불특정화자가 사용할 수 있도록 구축되고 있는 실정이다[2-3].

그러나 이들 시스템의 완전한 상용화를 위해서는 한국어 음성의 정확한 특징 분석과 화자의 개인성, 발성의 종류, 어휘수, 언어의 복잡성, 환경 요인, 인식의 단위 등과 같은 요인에 따른 문제점들을 정확하게 분석하고 검토하는 기초 연구가 요구되고 있으나 아직 이에 관한 연구는 미흡한 실정이라 할 수 있다.

특히, 실용화를 위한 연속음성인식 시스템의 인식률 제고를 위해서는 이러한 기초 연구를 통한 정확한 분석과 검토가 더욱 강조되고 있으며, 음향모델 및 언어 모델에 대한 연구도 선행되어야 하지만 정확한 검토 없이 사용되어 지고 있다.

또한 특정 분야에 고정도 인식 성능을 가지는 음성인식 시스템을 구현하기 위해서는 한국어만이 가지고 있는 특징분석과 이에 적절한 인식 문법으로 구성된 언어모델에 관한 연구도 중요하다 할 수 있다.

따라서, 본 연구에서는 이러한 기초연구의 일환으로서 실용화를 위한 연속음성인식 시스템의 성능 향상과 인식 시스템에 적합한 음향모델과 언어모델을 작성하고 인식실험을 통하여 그 유효성을 확인하고자 한다. 이를 위하여 음향모델에 있어서는 HMM의 출력확률분포의 mixture와 파

라미터 차원에 대한 분석과 반복학습법으로 특정 태스크에서 사용할 수 있는 N-gram 언어모델을 작성하고 3인이 발성한 항공편 예약 200문장을 대상으로 multi-pass 탐색 알고리즘을 이용하여 인식실험을 수행하고 그 결과를 비교 검토하고자 한다.

2. 음향모델 및 언어모델

2.1 혼합(mixture)계수 및 회귀계수

본 연구에서는 음향학적 모델로서 혼합계수 및 회귀계수에 대해서 고찰하며, 이에 대해 간략히 기술한다.

일반적으로 연속분포 HMM 모델에서 출력확률 $b_j(o)$ 은 식 (1)과 같이 Gaussian mixture 밀도함수[1,4]에 의해 나타낸다. 즉, HMM 모델의 경우 하나의 상태에서 다른 상태로 전이(轉移)할 때 식(1)의 Gaussian 밀도함수에서 각각의 mixture마다 평균과 공분산을 계산하게 된다.

$$b_j(o) = \sum_{k=1}^{M} c_{jk} N(o, \mu_{jk}, U_{jk}) = \sum_{k=1}^{M} c_{jk} b_{jk}(o) \quad (1)$$

$$\sum_{k=1}^{M} c_{jk} = 1 \quad (2)$$

$$c_{jk} \geq 0,\ 1 \leq j \leq S,\ 1 \leq k \leq M$$

식 (1)과 (2)에서 c_{jk}는 상태 j에서 k번째 mixture에 대한 mixture 계수를 나타내고, N은 상태 j에서 k번째 mixture 성분에 대해서 공분산 U_{jk} 와 평균 μ_{jk} 에 의한 Gaussian 밀도함수를 나타낸다. 또한, S는 상태수, M은 mixture의 수를 나타낸다.

또한 음성을 스펙트럼 영역에서 분석할 때 스펙트럼 내에서의 순시적인 변화는 음성의 중요한 정보를 가지고 있으며 스펙트럼 기울기의 변화는 스펙트럼 정보를 읽는데 중요한 단서가 된다. 일반적으로 화자가 달라지면 포만트[1]의 절대적 위치는 화자에 따라 변화하지만 포만트의 기울기는 상대적으로 변화하지 않는다. 이러한 특징은 인식률에 큰 영향을 미치는 데 본 연구에서는 음성으로부터 특징을 추출할 때 스펙트럼 내의 순시적인 변화를 나타내는 동적 특징 파라미터로서 Δ계수 또는 회귀계수를 이용한다[4]. Δ계수의 추출은 음성의 정적 특징 벡터의 각 차원에 대해 식 (3)을 이용하여 계산하고 이를 또 다른 특징 파라미터로 사용한다. Δ계수 d_t는 시간 t를 중심으로 정적 파라미터 $c_{t+\theta}$와 $c_{t-\theta}$에 의해 $2\theta+1$ 폭 만큼의 단위로 식 (3)으로부터 구하여진다.

$$d_t = \frac{\sum_{\theta=1}^{\theta} \theta (c_{t+\theta} - c_{t-\theta})}{2 \sum_{\theta=1}^{\theta} \theta^2} \quad (3)$$

여기서 $c_{t+\theta}$는 t 번째 프레임의 θ번째 정적 파라미터의 계수 값이고 d_t는 여기에 해당하는 Δ계수 값을 의미한다.

2.2 통계적 언어모델

본 연구에서는 소량의 텍스트 데이터만을 이용하여 특정 태스크에서 효율적으로 음성인식에 필요한 언어모델을 작성할 수 있는 반복학습법[8]에 의한 N-gram 언어모델[5]을 이용하였다.

반복학습법은 기존의 sparseness를 다루는 언어모델 작성 방법과 유사하지만 본 연구에서는 sparseness를 다루는 언어모델을 작성하기 위한 텍스트 데이터에 비하여 매우 적은 텍스트로 언어모델을 작성하는 방법으로 단어의 발생확률을 강건하게 하기 위해서 문장을 인위적으로 확장하였으며, 확장된 텍스트를 반복하여 언어모델의 학습에 이용하였다. 즉, 일반적인 N-gram 언어모델의 경우 대량의 텍스트 데이터가 필요하지만 본 연구에서는 소량의 데이터만으로 효율적인 언어모델을 작성하기 위한 방법으로 인식될 수 있는 가능한 문장으로부터 제한된 태스크에서 임의로 만들 수 있는 문장으로 확장하고 제한된 태스크에서의 문장만으로 언어모델을 만들 경우 발생할 수 있는 언어모델의 출현확률을 좀더 보강하였다. 또한, 통상 이용되는 출현빈도가 높은 단어를 선택하지 않고 1번 이상 나타나는 모든 단어를 언어모델의 출현확률 추정에 이용하였다.

3. 인식 방법

본 연구에서는 JULIUS[1] 연속음성인식 시스템을 사용하고 있으며 인식방법으로서는 1-pass 탐색에서는 단어 2-gram을, 2-pass 탐색에서는 단어 3-gram을 도입하는 multi-pass 탐색 알고리즘을 이용하고 있다[7].

1-pass에서는 시스템의 고속화를 위해 전향으로 프레임 동기형 빔 탐색 알고리즘을 이용하여 목구조 형태의 사전의 각 상태에 단어 2-gram 확률을 동적으로 분할하여 지정한 후 모든 단어에 대해서 탐색을 수행한 후 결과로 단어 그래프(word graph)를 출력한다. 2-pass에서는 1-pass의 중간결과로 단어 그래프를 입력으로 단어단위의 best-first의 스택 디코딩 탐색을 수행한다. 언어모델은 단어 3-gram을 이용하고 단어단위의 탐색을 이용하는 것은 단어 레벨에서의 제약조건을 다루기 쉽고 단어단위의 정밀한 언어모델을 적용하는데 용이하기 때문이다. 1-pass에서 출력된 단어 그래프 형태의 중간결과로부터 얻은 정보를 이용하기 위해 2-pass에서는 1-pass와 반대방향인 후향으로 탐색을 수행한다.

1) distributed by Information-technology Promotion Agency, Japan, 1999.

4. 인식실험 및 고찰

4.1 시스템 구성

전체 시스템의 구성도를 그림 1에 나타내었다. 기본 HMM 음향모델은 HTK[4]를 이용하여 연속분포를 가지는 유사음소단위(PLUs) HMM 모델을 구성하고, 언어모델은 텍스트 데이터로부터 CMU-SLM toolkit[5]을 이용하여 2-gram과 3-gram을 작성하였다. 발음사전은 한국어 처리시스템(KPS)[6]을 이용하여 한국어 음운규칙을 적용하였다. 이렇게 작성한 음향모델과 언어모델을 이용하여 입력된 음성으로부터 추출된 특징 파라미터에 대해 프레임 동기형 Viterbi 빔 탐색과 A* 스택 디코딩을 수행하여 인식결과를 출력하였다.

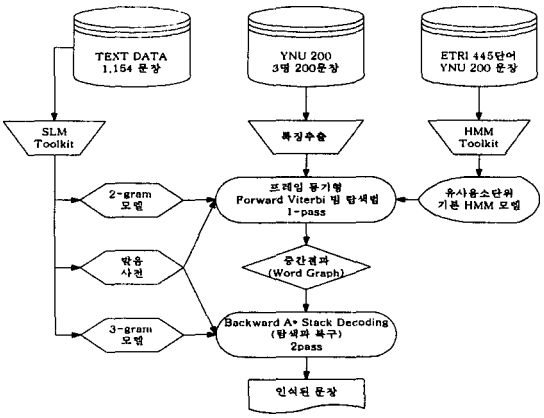

그림 1. 시스템 전체구성도

4.2 음성 데이터

표 1에 나타낸 것과 같이 기본 HMM 모델을 작성하기 위한 음성 데이터는 한국전자통신연구원(ETRI)에서 작성한 PBW 445단어 음성 데이터베이스 중 19명 2회 발성 중 1회분 총 8,455 단어와 영남대학교(YNU)에서 작성한 연속음성 데이터베이스 중 8명의 200문장을 사용하였다. 수작업에 의해 이루어진 ETRI 445 단어의 유사음소단위 레이블 정보와 자동으로 구한 YNU 200 연속음성의 레이블 정보를 이용하여 HTK에 의해 기본 HMM 모델을 작성하였다. 또한 평가용 데이터로는 YNU 200 문장 중 모델 작성에 참여하지 않은 3명의 200문장을 사용하였다.

본 연구에서 사용된 모든 음성 데이터는 16KHz, 16bits로 샘플링하고 $1-0.97z^{-1}$의 필터로 전처리한 후, 입력 음성의 각 프레임에 25msec의 Hamming windows를 곱하여 10msec 마다 분석하였으며, HTK에 의한 HMM 모델의 출력확률에서 mixture의 수를 1, 3, 5, 7, 9로 증가시켰으며, 분석을 통해 추출한 정적 파라미터인 12차의 MFCC(Mel-Frequency Cepstral Coefficient)와 동적 특징 파라미터인 12차의 회귀(Δ)계수와 Δpower 성분을 포함하여 25차 및 그 차분 성분인 $\Delta\Delta$계수와 $\Delta\Delta$power를 포함하여 38차인 특징 파라미터를 각 mixture 수에 따라 각각 구성하였다. 또한 모든 음성 채널을 고려하여 켑스트럼 평균 정규화(CMN)를 $\Delta\Delta$power에 대해 적용하였다.

표 1. 음성 데이터 베이스

음성 데이터			
발성형태	단어	연속음성	연속음성
화자	남성 19	남성 8	남성 3
사용단계	모델학습		인식
단어수/문장수	445	200	200
발성횟수	1	1	1
발성환경	방음부스		

4.3 인식실험

인식실험에서 사용한 음향모델은 4.2절의 표 1에 나타낸 것과 같이 ETRI 445 단어 19명과 YNU 13명의 음성 데이터 중 발성 상태가 양호한 8명의 연속음성에 대해 HTK에 의해 mixture 수를 1, 3, 5, 7, 9로 증가시키면서 각각 25차원과 38차원으로 작성하였다.

또한 언어모델에 있어서는 항공편 예약의 질문과 대답에 관련된 13명의 제한된 텍스트 데이터를 이용하여 항공편 예약 태스크 범위 내에서 인위적으로 확장한 후 반복학습으로 N-gram 언어모델을 작성하였다. 기존의 실험에 의해 최적의 언어모델은 동일 문장을 10회 정도 반복 학습한 것의 성능이 우수하여 이 언어모델을 적용하였다[8].

이상의 음향모델과 반복학습한 N-gram 언어모델을 이용하여 3인의 화자가 발성한 200문장을 대상으로 인식실험을 수행하였다. 그 결과를 표 2와 그림 2, 표 3과 그림 3에 각각 나타내었다.

표 2. 25차 특징 파라미터의 mixture 변화에 따른 인식률 변화

언어 모델	25차 mixture	화 자 (200문장) CJR	JYC	KHN	평균(%)
10회 반복 학습 언어 모델	1	47.5(39.0)	61.5(55.5)	60.0(52.0)	56.3(48.3)
	3	65.5(53.0)	75.5(68.5)	75.5(67.5)	72.2(63.0)
	5	71.5(60.0)	75.0(68.0)	82.5(71.5)	76.3(66.5)
	7	75.0(63.5)	76.0(71.5)	82.5(73.0)	77.8(69.3)
	9	76.5(68.0)	81.0(79.5)	85.5(79.5)	81.0(74.2)

인식률(%): 2-pass with 3-gram(1-pass with 2-gram)

그림 2. 25차 특징 파라미터의 mixture 변화에 따른 인식률 변화

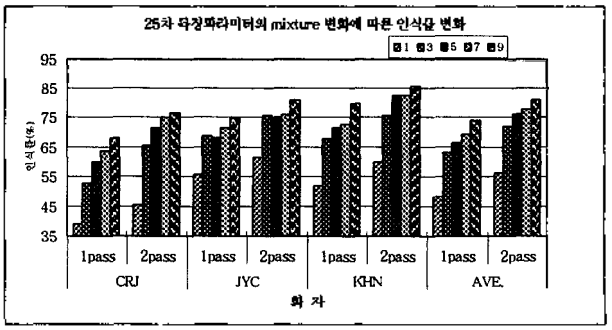

표 3. 38차 특징 파라미터의 mixture 변화에 따른 인식률 변화

언어 모델	38차 mixture	화 자 (200문장) CJR	JYC	KHN	평균(%)
10회 반복 학습 언어 모델	1	55.0(43.5)	75.5(65.5)	68.0(58.0)	66.2(55.7)
	3	80.5(69.0)	89.0(81.5)	85.5(73.0)	85.0(74.5)
	5	85.0(73.5)	90.0(84.5)	88.5(78.5)	87.8(78.8)
	7	86.0(73.5)	91.5(88.5)	89.0(79.5)	88.8(80.5)
	9	88.0(76.0)	91.0(87.0)	91.5(76.0)	90.2(79.7)

인식률(%): 2-pass with 3-gram(1-pass with 2-gram)

그림 3. 38차 특징 파라미터의 mixture 변화에 따른 인식률 변화

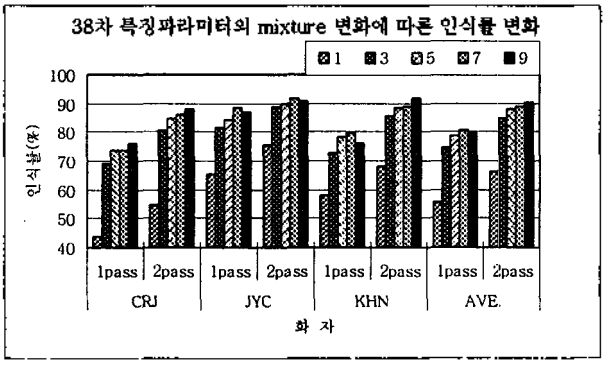

표 2와 그림 2로부터 25차의 특징 파라미터를 가지는 음향모델과 10회 반복학습한 언어모델에 대해서 mixture 수가 1에서 3으로 증가시켰을 경우 인식률의 변화가 평균 15%로서 높은 인식률 향상을 보였으나, mixture 수가 3이상으로 증가시키면서 실험한 경우에 있어서는 평균 3% 정도의 인식률 증가를 보였다. 전체적으로 볼 때 25차원에 대한 mixture 수가 9인 음향모델과 10회 반복학습한 언어모델을 이용한 경우가 평균 81.0%로 가장 높은 인식률을 보임을 알 수 있었다.

또한 표 3과 그림 3으로부터 38차의 특징 파라미터를 가지는 음향모델과 10회 반복학습한 언어모델에 있어서도 mixture 수가 1에서 3으로 증가시킨 경우 인식률의 변화가 평균 19%로서 높은 인식률 향상을 보였으며, mixture 수가 3이상의 경우에서는 mixture 수가 증가함에 따라 평균 1.7%의 인식률 향상을 보였다. 특히 38차원에 대한 mixture 수가 9인 음향모델과 10회 반복학습한 언어모델을 이용한 경우 평균 90.2%의 인식률을 보여 25차원에 대한 mixture 수가 9인 음향모델과 10회 반복학습한 언어모델을 이용한 경우보다 평균 9.2% 향상된 인식률을 얻었다.

이상의 인식실험 결과로부터, 인식률 제고를 위해서는 38차원에 대한 mixture 수가 9인 음향모델과 10회 반복학습으로 작성한 언어모델을 이용한 경우가 매우 효과적임을 알 수 있었다. 그러나 인식 시스템의 계산량 증가에 따른 시스템의 속도 제고를 위해서는 38차원에 대한 mixture 수가 3인 음향모델과 10회의 반복학습으로 작성한 언어모델을 이용하는 것도 바람직할 것으로 생각된다.

5. 결 론

본 연구에서는 연속음성인식 시스템의 성능 향상을 위한 기초 연구로서 시스템에 적합한 음향모델과 언어모델을 작성하고 항공편 예약 태스크를 대상으로 인식실험을 실시한 결과 그 유효성을 확인하였다.

이를 위하여 HMM의 출력확률분포의 mixture 수와 파라미터의 차원에 따른 음향모델을 작성하였으며, 반복학습법을 이용하여 특정 태스크를 대상으로 N-gram 언어모델을 작성하고, 화자 3인의 200문장에 대해 인식실험을 수행하였다. 그 결과, 25차원에 대한 mixture 수가 9인 음향모델과 10회 반복학습한 언어모델을 이용한 경우 평균 81.0%의 인식률을 얻었으며, 38차원에 대한 mixture 수가 9인 음향모델과 10회 반복학습한 언어모델을 이용한 경우 평균 90.2%의 인식률을 보여 25차원에 대한 mixture 수가 9인 음향모델과 10회 반복학습한 언어모델을 이용한 경우보다 평균 9.2% 향상된 인식률을 얻었다.

전체적으로 볼 때, 인식률 제고를 위해서는 38차원에 대한 mixture 수가 9인 음향모델과 10회 반복학습으로 작성한 언어모델을 이용한 경우가 매우 효과적임을 알 수 있었다. 그러나 인식 시스템의 계산량 증가에 따른 시스템의 속도 제고를 위해서는 38차원에 대한 mixture 수가 3인 음향모델과 10회의 반복학습으로 작성한 언어모델을 이용하는 것도 바람직할 것으로 생각된다.

참고문헌

[1]L.R. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993.

[2]김범국, 정현열, "가변장 음소모델을 이용한 음소인식," 한국음향학회지, 제16권 제8호, 1997.11.

[3]김득수, 황철준, 정현열, "음성인식 기능을 가진 주소입력 시스템의 개발과 평가," 한국음향학회지, 제18권 제2호, 1999.2.

[4]S. Young, J. Jansen, J. Odell, D.Ollason, and P. Woodland, "The HTK Book," 1995.

[5]P. Clarkson, R.Rosenfeld, Statistical Language Modeling Using the CMU-Cambridge Toolkit, Proc. Eurospeech 97, pp. 2707-2710, Sept. 1997.

[6]이상호, 오영환, 서정연, "한국어 문서 음성변환 시스템을 위한 문서 분석기," 한국음향학회지, 제15권 제3호, 1999.

[7]A. Lee, T. Kawahara and S. Doshita, "Large Vocabulary Continuous Speech Recognition Based on Multi-Pass Search Using Word Trellis Index," In IEICE, Vol. J82-D-II, No. 1, pp. 1-9, 1999.

[8]S.J.Oh, C.J.Hwang, H.Y.Jung and H.Y.Chung, "A study on Statistical Language Models for Large Vocabulary Continuous Speech Recognition," Proc. of ICSP'99, Vol. 1, 1999. 8.