1999년도 한국음향학회 학술발표대회 논문집 제18권 제1(s)호

음향학적 파라메터를 이용한 한국어 연결숫자인식의 성능개선

김승희*, 김형순**
*한국전자통신연구원, **부산대학교 전자공학과

Performance Improvement of Korean Connected Digit Recognition Based on Acoustic Parameters

Seunghi Kim*, Hyung Soon Kim**
*Electronics & Telecommunications Research Institute
**Dept. of Electronics Eng., Pusan National University
E-mail : seunghi@etri.re.kr, kimhs@hyowon.cc.pusan.ac.kr

요 약

본 논문에서는 한국어 연결숫자인식에 있어서 모델간의 변별력 향상을 통해 인식률을 높이기 위하여 음향학적 파라메터(Acoustic Parameter)를 사용하는 것을 제안한다. 제안된 방법은 음성학적 지식에 근거하여 적절한 주파수 대역별 에너지의 비의 로그값을 추가적인 특징 파라메터로 사용한다. 실험결과, 제안된 방법을 사용함으로써 기본 인식시스템에 비해 오류율이 최고 46% 정도 감소됨을 확인할 수 있었다. 그리고 채널보상 기술을 함께 적용함으로써 69% 정도의 오류율 감소를 얻었다.

1. 서 론

음성인식의 한 분야인 음성입력에 의한 숫자인식은 고립숫자인식과 연결숫자인식의 두 부류로 나눌 수 있으며, 사용자의 편의성을 고려할 때 연결숫자인식의 필요성이 나날이 증대되고 있다.

그 동안 연결숫자인식에 관한 많은 연구가 이루어져 왔으며, 한국어 숫자의 경우에도 최근 들어 활발한 연구들이 이루어지고 있다[1][2]. 그러나 한국어 숫자들은 우선 모두 단음절이고, 게다가 숫자들간의 혼동가능성도 커서 영어 등 타언어권의 숫자인식에 비해 난이도가 높다. 한국어 숫자인식에서 오인식되는 경우들을 살펴보면, 몇몇 숫자쌍들로 오류분포가 집중되는 것을 알 수 있다. 따라서 이런 숫자들간의 변별력을 높일 수 있다면 인식시스템의 성능은 크게 개선될 수 있을 것이다. 그러나, 현재까지 이들 오인식되는 숫자들간의 변별력을 향상시키기 위한 연구들은 별로 이루어지지 못한 실정이다.

이 문제를 해결하기 위한 접근 방향은 특징 파라메터 추출과정에서의 접근방법과 숫자모델 훈련과정에서의 접근방법의 두 가지로 나누어 볼 수 있다.

본 논문에서는 특징 파라메터 추출과정에서 각 숫자들에 대한 사전 지식을 기반으로 하여 이들간의 변별력을 크게 할 수 있는 음향학적인 특징 파라메터를 추가적으로 추출하여 사용하는 방법에 의해 인식성능개선을 도모하였다.

본 논문의 구성은 다음과 같다. 서론에 이어 2 장에서는 본 논문에서 제안하고 있는 음향학적 특징 파라메터에 대해 기술하며, 3 장에서는 채널 특성 및 화자 특성의 보상에 대해 간략히 설명하고, 4 장에서는 인식실험

과정 및 결과를 다룬 후, 5 장에서 결론을 맺는다.

2. 음향학적 특징 파라메터

2.1 오인식되는 숫자쌍들에 대한 특징 분석

인식 실험 후의 결과를 토대로 오인식되는 양상들을 살펴보면 자주 오인식되는 몇 가지 숫자쌍들을 발견할 수 있다. '구⇔오', '오⇔공'의 예도 그 중 하나인데 이들은 모음 스펙트럼이 상당히 유사하며, 모음이 단독으로 있을 경우에는 구별이 힘들다. 따라서 이들의 경우에서는 자음인 /ㄱ/이 인식에서 차지하는 비중을 높일 경우 구분이 보다 용이해 질 것으로 생각할 수 있다(그림 1 참조). '일⇔이'도 스펙트로그램 상으로 구분이 힘들며 화자에 상관없이 고루 발생하는 오인식 쌍이다.

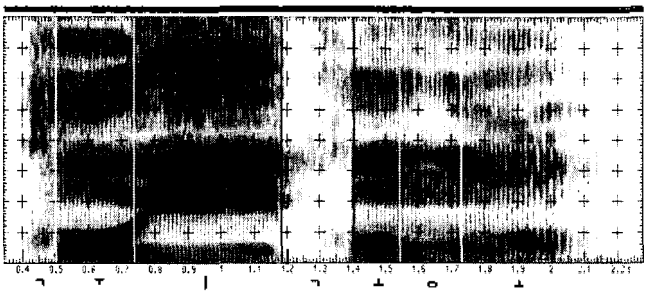

그림 1. 숫자열 /구이공오/에 대한 스펙트로그램의 예

2.2 음향학적 특징 파라메터 (Acoustic Parameter)

음소들 그룹간의 구별에 있어서는 음운학적 자질론에서 출발한 음향학적 파라메터(이하 AP)를 사용하는 것이 켑스트럼에 비교하여 화자들간의 차이에 보다 둔감한 것으로 보고되었다[3]. Hansen[3]은 TIMIT 데이터베이스에서 각 음소들의 그룹들을 구분하기 위해, 주파수 밴드별 에너지에 기반을 둔 AP 와, 유성음일 확률(voicing probability), 그리고 제 1, 제 2, 제 3 포먼트(formant) 주파수 및 그 대역폭을 사용하였다. Hansen 의 논문에서 에너지에 기반한 대부분의 파라메터들은 입력 음성 중 가장 큰 값으로 정규화되었다. 그러나, 이 방법은 일단 음성이 다 입력된 다음에야 정규화의 기준이 되는 가장 큰 값을 구할 수 있으므로 실시간 처리에 문제가 있다. 그래서 본 논문에서는 정규화된 에너지를 사용하지 않는다. 대신에 다음과 같이 주파수 밴드별 에너지의 비에 기반을 둔 파라메터나 혹은 그들의 이동평균(moving average)을 통해서 앞에서 언급한 에너지 분포 패턴에 관한 정보를 구한다.

$$E(t)_{AP} = \log\left(\frac{E_{f1-f2}(t)}{E_{f3-f4}(t)}\right)$$

$$E(t)_{AP\_MA} = \frac{1}{5}\sum_{i=-2}^{2} E(t-i)_{AP}$$

여기서 $E_{f1-f2}(t)$는 t 번째 프레임에서 주파수 $f1$ 에서 $f2$ 까지의 에너지를 말한다.

숫자음에서 모음을 포함한 유성음의 경우 /아/를 제외하고는 대부분 400Hz 근처에서 에너지값이 peak 를 이룬다. 그리고 500Hz 를 경계로 에너지가 급속히 감소하는 경향이 있으며, 4kHz 이상에서는 전체적으로 에너지가 작은 것을 관찰할 수 있다. /아/의 경우에는 0~1kHz 대역에 강한 에너지가 나타난다. 무성음의 경우, 특히 /ㅅ/이나 /ㅈ/과 같은 마찰, 파찰자음들은 전반적으로 백색잡음과 유사한 형태이나, 4kHz 이하보다는 그 이상의 대역에서 높은 에너지를 가지는 경우가 많다. 본 논문에서는 이들 500Hz, 1kHz 와 4kHz 를 주파수 대역을 나누는 경계로 사용하였다.

3. 채널 특성 및 화자 특성의 보상

HMM 을 기반으로 하는 인식 시스템의 성능저하에 영향을 미치는 중요한 요소 중의 하나는 훈련환경과 실제 인식과정에서의 환경의 차이이다. 따라서 인식 시스템을 배경잡음이 존재하는 곳에서 사용하기 위해서는 배경잡음에 의한 영향을 보상해 주어야 한다. 배경잡음 중 특히 채널잡음에 있어서는 CMS(Cepstral Mean Subtraction)나 RASTA 등이 간단하면서도 성능이 우수한 것으로 알려져 있으며 본 논문에서는 global CMS 와 local CMS, RASTA 의 방법을 사용하여 채널왜곡 보상을 시도하였다.

음성신호의 장구간 스펙트럼이 화자의 특성에 영향을

받는다는 것은 잘 알려진 사실이며, 음성신호의 장구간 스펙트럼 특성은 시간에 대해 독립적이거나 느리게 변한다. TIDIGIT 데이터베이스에 대해 실험한 결과, 시간 영역에서의 스펙트럼 파라메터열을 주파수영역으로 변환했을 때 0 에서 1Hz 까지의 분산은 숫자 그룹의 차이보다는 화자의 목소리에 훨씬 많은 영향을 받는 것으로 나타났다[4]. 결론적으로, 이 파라메터열의 저주파 영역의 요소들을 필터를 통해 감쇄시킴으로서 인식시스템은 화자에 보다 더 독립적인 시스템이 될 것이다. 대부분의 채널 왜곡 보상방법들은 위의 파라메터열에서 낮은 주파수 성분들을 제거하는 것이며, 따라서 clean speech 에 대해서도 채널왜곡보상기법을 적용할 경우 화자특성 보상에 의한 얼마간의 성능향상을 기대할 수 있다.

4. 실험 및 결과

4.1 데이터 베이스 및 기본 인식 시스템

본 장에서는 연속 HMM 을 이용하여 연결숫자인식 실험을 하며, 앞 절에서 설명한 방법들을 통해 인식성능의 개선정도를 평가하고자 한다. 실험에 사용하는 음성 특징 파라메터로는 기본적으로 12 차 MFCC, delta MFCC, 그리고 delta 에너지를 사용하여 총 25 차의 파라메터를 사용한다. 국어공학센터의 한국어 4 연숫자 음성 DB 중에서 남성화자들의 음성만을 사용하였다[5]. 방음부스에서 녹음되었으며 16kHz sampling 및 16bits 로 양자화되어 있다. 본 DB 에서 사용된 숫자들은 '공, 일, 이, 삼, 사, 오, 육, 칠, 팔, 구'의 10 개이다. 훈련을 위해 30 명의 화자가 발성한 4130 개의 4 연숫자문장을 사용하며, 인식을 위해서는 위에 속하지 않은 9 명의 화자가 발성한 1260 개의 문장을 사용하였다.

triphone 을 사용하였으며, 음소당 state 3 개와 5 개에 대해 mixture 수를 증가시키면서 실험한 결과 state 5 개의 결과가 우수하여 이후 음소당 state 5 개에 대해서만 실험을 진행하였다.

주파수 밴드는 0~500Hz(E1), 0.5~1kHz(E2), 1~4kHz(E3), 4~8kHz(E4)로 나누었으며, 음향학적 파라메터는 이들 주파수 밴드별 에너지의 비의 로그값을 사용하였다. 표에서 E1 등과 같이 하나의 주파수 대역만 사용했을 경우는 전체 에너지에 대한 비를 나타내며, E1, E2, E3, E4 등과 같이 여러 개의 주파수 대역을 사용한 경우는 가장 마지막 주파수 대역의 에너지에 대한 비를 나타낸다.

실험결과 E1, E2, E3 에 대해 log(E1/E3), log(E2/E3)의 값에 대한 이동평균 및 이들의 차분값을 부가적인 특징 벡터로 추가한 경우의 인식률이 가장 높았다(표 1 참조). 실험은 각각의 경우에 대해 mixture 수를 1 개에서 13 개까지 늘려가면서 수행하였으며, 가장 높은 인식률을 나타낸 mixture 개수에 대한 결과만을 표에 나타내었다.

채널 보상 및 개개 화자 특성의 보상을 위해서 global CMS(GCMS), local CMS(LCMS), RASTA 방법을 적용하여 실험하였다. 300ms 의 윈도우를 사용한 LCMS 방법의 결과가 가장 우수했으며 GCMS, RASTA 에 대해서도 유사한 성능이 얻어졌다(표 2 참조).

다음으로 음향학적 파라메터를 추가한 후에 채널 보상 알고리즘을 적용하여 실험을 하였다. 주파수 밴드 E1, E2, E3 에 대한 음향학적 특징 파라메터를 추가하고 RASTA 방법을 적용하였을 때 98.5%의 가장 좋은 결과를 나타내었다. 표 3 에는 부가적인 특징 벡터와 채널 보상 기술을 함께 적용한 여러 실험들 중에서 우수한 성능을 나타낸 실험들의 결과만을 나타내었다.

기본 인식시스템의 실험결과를 분석해 보면 전체 오류중 '일⇔이'의 오류가 전체 오류의 37%, '구⇔오⇔공

표 1. 음향학적 파라메터(AP)를 사용한 실험의 인식률 (%)

사용한 주파수 대역	최적 mixture 개수	숫자열	개별 숫자
기본 인식 시스템	**3**	**95.2**	**98.6**
E1	3	95.4	98.7
E2	1	95.2	98.7
E3	1	94.7	98.5
E4	3	94.7	98.5
E1, E2, E3	5	97.0	99.2
E1, E2, E3 (이동평균)	5	97.2	99.2
E1, E2, E3(이동평균) + 차분값	**3**	**97.4**	**99.3**

표 2. 채널 보상 기술을 적용한 실험의 인식률 (%)

적용한 채널 보상기술	최적 mixture 개수	숫자열	개별숫자
-	**3**	**95.2**	**98.6**
GCMS	7	97.4	99.3
LCMS 150ms	5	97.4	99.3
LCMS 300ms	**5**	**97.9**	**99.4**
LCMS 450ms	7	97.8	99.4
RASTA	3	97.5	99.3

표 3. 부가적인 특징 벡터와 채널 보상 기술을 함께 적용한 실험의 인식률 (%)

적용 알고리즘	최적 mixture 개수	숫자열	개 별 숫 자
-	**3**	**95.2**	**98.6**
E1,E2,E3 + RASTA	**3**	**98.5**	**99.6**
E1,E2,E3(이동평균)+LCMS300	9	98.3	99.6
E1,E2,E3(이동평균) +delta+RASTA	3	98.3	99.5
E1,E2,E3,E4 + RASTA	**5**	**98.5**	**99.6**
E1,E2,E3,E4(이동평균) +RASTA	5	98.4	99.6
E1,E2,E3,E4 + delta + RASTA	5	97.9	99.5

의 오류가 전체의 48%를 차지하였다. 음향학적 파라메터를 도입해서 인식성능이 가장 우수한 경우 '일⇔이'의 오류는 개선정도가 미미하나, '구⇔오⇔공'의 오류는 68% 개선된 결과를 보이며, 채널 보상 알고리즘만을 적용했을 경우에는 '일⇔이'의 경우 41% 개선된 결과를 나타내었다. 음향학적 특징 파라메터를 추가한 뒤 채널 보상 알고리즘을 적용한 경우에는 기본 인식시스템의 결과에 대해 '일⇔이'가 55%, '구⇔오⇔공'은 86% 개선된 결과를 나타내었다. 즉, 음향학적 파라메터를 도입함으로써 '구⇔오⇔공'의 오류는 상당히 개선됨을 알 수 있다. 참고로 전화채널환경의 데이터 베이스에 대해서는 음향학적 파라메터를 도입하였을 때 인식률의 향상을 가져오지 못했다.

5. 결 론

본 논문에서는 오인식이 빈번한 숫자쌍 모델들간의 변별력을 향상시킴으로써 연결숫자인식 시스템의 성능을 개선시키기 위해 부가적인 음향학적 특징 파라메터를 추출하여 사용하였다. triphone 을 기본 모델링 단위로 사용하였으며, 숫자간의 변별력 향상을 위해 부가적인 음향학적 파라메터로서 주파수 밴드별 에너지의 비의 로그값을 사용하였다. 실험결과 인식성능이 가장 우수한 경우 46% 정도 오류율이 감소됨을 확인할 수 있었다. 그리고 채널 및 화자특성의 차이를 보상하기 위해 몇 가지 채널 보상 기술을 추가적으로 적용함으로써 최고 69% 정도의 오류율 감소를 볼 수 있었다.

향후 한국어 음소별 특징에 관한 체계적이고 상세한 연구가 뒷받침된다면 보다 더 우수한 성능을 기대할 수 있을 것으로 생각된다. 그리고, 본 논문에서는 구현하지 않았으나, 서론에서 언급한 바와 같이 숫자간의 변별력 향상을 위해 훈련과정에서 MMIE 나 MCE 기법을 적용하게 되면 추가적인 성능향상이 기대되며, 앞으로 이에 대한 연구가 계속 진행되어야 할 것으로 판단된다.

참 고 문 헌

[1] 양태영 외 8 인, "연결 숫자음 인식에서의 상태 및 단어 유지 확률을 이용한 지속시간 모델링," 제 10 회 신호처리 합동학술대회 논문집, pp. 313-316, 1997.

[2] 김기성, 김승회, 김형순, 지민제, "한국어 연결숫자인식을 위한 숫자 모델링에 관한 연구," 제 15 회 음성통신 및 신호처리 워크샵 논문집, pp.293-297, 1998.

[3] A. V. Hansen, "Acoustic-Phonetic Features used in Automatic Speech Recognition," Ph.D. thesis, Department of Communication Technology, Aalborg University, 1998.

[4] C. Nadeu, P. Paches-Leal and B. H. Juang, "Filtering the time sequences of spectral parameters for speech recognition," Speech Communication, vol.22, pp.315-332, June 1997.

[5] Korean speech data base CD-ROM, 국어공학센터, 1998.