특집

잡음에 강인한 실감형 미디어 음성 인식 기술

홍 훈 섭[1], 이 상 운[2], 이 연 철[3], 한 문 성[4]

목 차

1. 서 론

컴퓨터 관련기술과 인터넷의 비약적인 발전을 통해 우리는 정보화 사회의 혜택을 일상생활 곳곳에서 느낄 수 있다. 또한 이러한 기술의 발전은 기존에 유통되던 정보의 양과 품질을 비약적으로 향상시켰다. 그러나 종래의 기술 발전은 그 기술을 사용하기 위한 방법을 인간에게 강요하는 특성을 갖는다. 즉, 발전된 기술과 광대한 정보를 이용하기 위한 방법을 익히기 위해서는 추가의 교육 또는 학습을 필요로 하는 문제점을 갖는다. 이러한 문제점들은 기술 소비계층을 기술에 적응할 수 있는 사람으로 국한시키게 되며, 기술에 적응하기 힘든 노약자등의 계층에 기술 소외를 불러오게 된다. 따라서 특별한 학습이나 교육을 요구하지 않으면서도 풍부하고 다양한 정보에 다가갈 수 있는 친숙한 형태의 기술은 반드시 연구가 진행되어야 하는 분야이다. 그러나 이러한 기술분야는 폭발적인 연구가 이루어진 다른 기술분야에 비해 비교적 소홀하게 다루어져 온 것이 사실이며, 이러한 문제점을 극복하기 위한 실감형 미디어 기술을 기반으로 한 기술의 인간화에 대한 연구는 정보화 사회의 보편화를 위해 반드시 이루어야 할 연구이다.

이를 위해서 인간에게 친숙할 수 있는 여러 인터페이스의 형태 중 사람들 사이에서 보통 사용하는 의사소통의 수단인 말, 즉 음성을 매개로 하는 인터페이스는 사용자에게 가장 자연스럽게 다가갈 수 있는 기계와 인간, 정보와 인간 간의 의사소통의 중요한 수단 중의 하나가 될 것이다. 이러한 음성인터페이스 기술은 실감형 미디어분야에 국한되는 것이 아닌, 정보 기술 산업의 모든 분야에 접목되어 다양한 파생시장을 형성할 수 있는 기술이며, 인간과 컴퓨터와의 인터페이스를 편리하게 개선시켜줄 핵심기술 중의 하나로 주목받고 있다. 이러한 음성 인터페이스 관련 기술은 최근 마이크로프로세서의 성능향상과 지능형 정보단말기, 휴대폰 등의 보급과 맞물려 확장일로에 있으며, 현재 음성 인터페이스를 위한 음성 인식 성능 향상과 그 응용분야에 대한 연구 또한 국내외의 여러 연구소, 기업 등에서 진행되고 있다.

음성인터페이스 기술은 기존 기술의 한계를 한

1) 한국전자통신연구원 컴퓨터소프트웨어연구소 지식처리팀 연구원
2) ETRI 컴퓨터 소프트웨어 연구소 지식처리팀 연구원
3) 휴먼미디어테크 연구원
4) ETRI 컴퓨터 소프트웨어 연구소 지식처리팀 책임 연구원

차례 뛰어넘어 보다 인간에게 가까이 갈 수 있는 기술로 각광받고 있으며, 본 논문에서는 이러한 음성인터페이스 기술에 대한 동향과 기존 음성 인식 기술의 문제점, 그리고 멀티 모달 정보를 이용한 음성 인터페이스 기술과 넥마이크를 통한 기존 한계의 극복방안에 대해 알아보도록 한다. 또, 이에 따르는 음성 인터페이스 기술의 전반에 대한 간략한 기술전망을 통해 음성 인터페이스 기술에 대한 전반적인 고찰을 하도록 한다.

2. 음성 인식 기술 분석

음성을 매개로 하는 인터페이스에 대한 연구는 음성을 매개로 한 인간의 의사소통 능력을 기계로 이식하고자하는 인간의 욕구에서 시작되었으며, 그 연구의 역사 또한 길다. 이러한 음성 인터페이스를 위한 음성 인식 기술은 1970년대 미국 국방성의 프로젝트를 통해 본격화되었다[1]. 우리나라의 경우 80년대 중반 이후로 한국전자통신연구원 등을 중심으로 이러한 연구가 진행되었다.

〈표 1〉 음성 인식 시스템의 구분

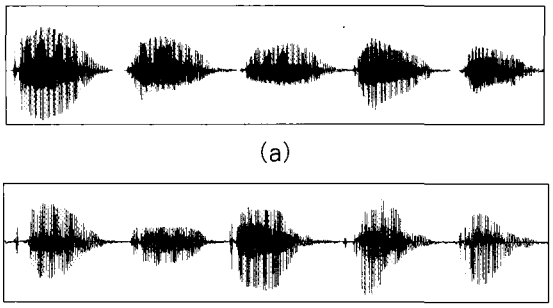
	용 어	내 용
화자에 따른 구분	화 자 독 립	화자에 따른 별도의 훈련과정이 필요없음
	화 자 종 속	화자에 따른 별도의 훈련과정이 필요함
인식 방법에 따른 구분	고립단어 인식	다른 단어와 연결되지 않은 단어를 인식함
	연결단어 인식	연결 숫자음등 개개의 단어를 별도로 인식함
	연속어 인식	정확한 발음으로 읽는 발성을 인식함
	대화체 인식	일상대화에서 주고받는 발성을 인식함
	핵심어 인식	입력되는 다양한 음성신호중 의미있는 부분만을 검출 인식함.

음성인식 기술은 시스템에 따라 〈표 1〉과 같이 정리할 수 있으며 이러한 음성 인식 기술 중 현재 가장 관심을 끄는 것은 화자 독립 형태의 연속어 음성 인식 기술로 화자 종속과 고립단어 인식보다 비교적 인간에 가까운 인터페이스를 제공할 수 있는 기술이다. 즉, 실감형 미디어 기술에 기반한 음성인터페이스는 자연스런 인식을 위한 화자 독립과 대화체, 핵심어 기술을 기반으로 한 음성 인식을 필연적으로 필요로 하며, 이를 위해 최근 은닉 마코프(Hidden Markov Model) 기법과 TDNN(Time Delayed Neural Network) 등의 기법을 적용하여 보다 인간 친화적 음성인터페이스 시스템의 구축을 서두르고 있다.

이러한 기존의 음성 인식 기술은 훈련단계에 사용된 음성데이터 베이스와 동일한 환경과 잡음이 없는 환경에서의 입력 음성에 대해서는 상당한 인식 성능을 보인다. 하지만 입력마이크의 종류나 실제환경에서 잡음이 존재하는 경우, 그 인식성능이 현저히 저하되는 것이 일반적인 현상이다. 이러한 상황을 극복하기 위한 다양한 방법들이 연구되고 있다. 즉, 전처리 과정에서 잡음을 제거하는 기법으로 음질향상법[2]과 신호원분리법, 강인한 특징벡터 추출[3], 채널잡음 제거법[4] 등이 있고, 모델 차원에서 문제를 해결하는 방법으로는 PMC[5], 모델 파라메터 적응법[6] 등이 있다. 또한, 새로운 입력원으로 넥마이크를 사용하여 외부 잡음원을 차단하거나[7] 입술 신호를 입력 신호로 병행하여 잡음에 강인한 음성 인식 시스템[8]을 만들어내는 방법이 있다.

3. 넥마이크를 이용한 음성 인식 시스템

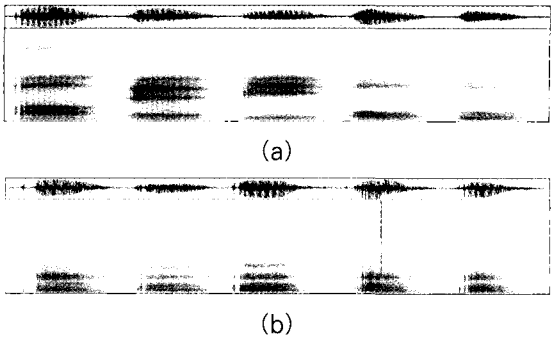
넥마이크는 목부분에 접촉 장착되어 발성이 이루어질 때 목부분의 피부조직의 울림을 센싱하여 음성파형을 얻는 장치로, 넥마이크에 의한 음성파형은 일반 마이크에 의해 공기의 울림을 센싱한 음성파형과 유사한 형태를 나타내고 있으나 재생할 때 음색은 다르게 나타나며 스펙트로그램상의 대역별 에너지 및 포먼트의 위치들이 다르게 나타난다. 넥마이크를 통해 입력받은 음성 파형을 (그림 1)의 (b)에 나타내었다.



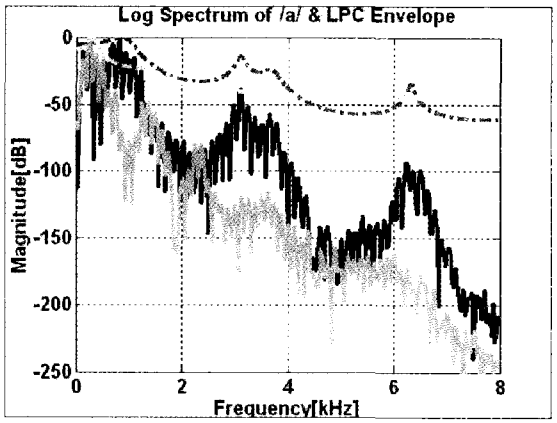
(그림 1) 모음 /a/-/e/-/i/-/o/-/u/를 발성했을 때의 음성 파형 : (a)일반 마이크 (b)넥 마이크

(그림 1)에서 알수 있는 바와 같이 일반마이크의 음성파형(a)와 비슷한 형태를 보여주므로 넥마이크에서 채집된 신호의 이용가능성을 보여준다. 하지만 실제로 음성을 들어보면 넥마이크의 음성은 명료도가 낮아 분별력을 떨어뜨리는 경우도 있다. 여기서 주목할 만한 점은 음성의 발성과정이 생략된 상황인데도 입을 통해서 발성된 음파를 신호원으로 하는 경우와 목에서의 진동을 신호원으로 하는 경우가 비슷한 파형을 보여준다는 것이다. 음성의 발성과정을 살펴보면 모음의 경우 성대의 주기적인 떨림에 의한 여기신호에 성도(vocal track)를 거치면서 혀의 위치나 비강 등의 영향을 받아 특정 주파수 성분이 공진이 되어 음의 분별력을 제공하고 입을 통하여 음파로써 전달된다. 성도를 거치지 않은 신호가 비슷한 파형으로 관측되는 것은 피부조직을 통해서 전달 되는 것으로 여겨지는데 이에 대한 자세한 고찰과 규명이 필요할 것으로 보인다.

주파수 영역에서의 특성을 살펴보기 위해 (그림 2)에 스펙트로그램을 나타내었다. (a)의 일반마이크를 이용하여 입력받은 음성에 대한 스펙트로그램에서 각 모음 음소들의 포먼트의 위차가 더 이상 음소의 분별력에 도움을 주지 못하고 또한 높은 주파수 대역쪽에서 신호가 거의 나타나지 않아 넥마이크의 특성 또는 목부분의 피부조직의 전달 특성에서 대역제한 특성이 있는 것으로 보인다.



(그림 2) 모음 /a/-/e/-/i/-/o/-/u/의 스펙트로그램; (a)일반 마이크 (b)넥 마이크



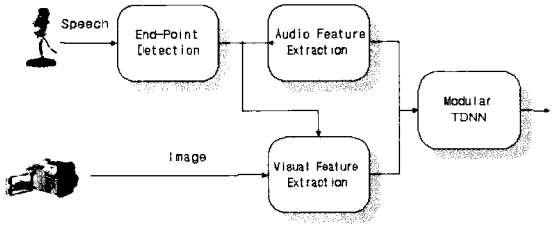
(그림 3) 모음 /a/의 로그 스펙트럼 및 LPC계수의 포락선; 진한 실선: 일반 마이크의 로그스펙트럼, 연한 실선: 넥 마이크의 로그스펙트럼, 일점쇄선: 일반마이크의 LPC포락선, 파선: 넥마이크의 LPC 포락선

모음 /a/에 대한 로그 스펙트럼과 LPC계수의 포락선을 (그림 3)에 나타내었다. 일반 마이크에 의한 신호의 로그 스펙트럼에서는 성도의 특성을 나타내는 포먼트 성분들이 뚜렷이 나타나지만 넥마이크에 의한 신호의 로그 스펙트럼에서는 저주파수 대역에서 포먼트가 발생하며 고주파 대역으로 갈수록 감쇄가 심하게 일어난다. 이러한 현상은 다른 모음 음소에서도 관찰되었으며 연속 음성인식에서 음소 모델의 분별력을 약화시킬 것이다.

넥마이크를 이용한 음성 인식은 고주파 대역의 정보가 일반마이크와 다르고 모음 음소정보의 경우 성도의 특성을 나타내는 포먼트의 위치 정보가 많이 소실되어 나타나므로 음소단위의 모델링에 문제가 있는 단점이 있다. 그러나 외부 잡음원의 유무에 관계없는 신호를 제공하므로, 추가의 연구가 이루어진다면 음성 인식에서의 잡음 문제를 해결할 수 있는 새로운 대안으로 떠오를 것으로 생각된다.

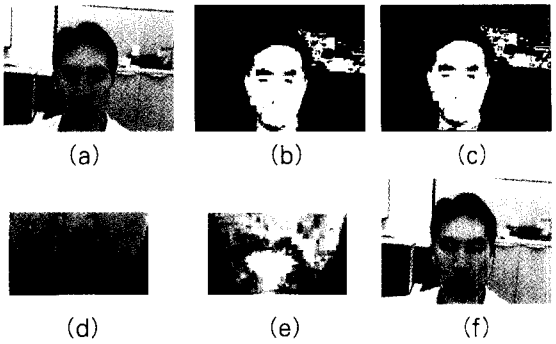
4. 멀티모달을 이용한 음성 인식 시스템

음성 인식 시스템의 한계를 극복하기 위한 여러 시도 중 최근 주목을 받고 있는 시스템은 오디오와 비디오 정보의 융합을 통한 멀티모달 음성 인식 시스템[8]이다. 이것은 인간의 음성 인식 능력 자체가 멀티모달의 특성을 갖고 있다는 점에 착안한 방법[9]이다. 즉, 잡음이 많은 환경에서 효율적으로 사용자의 음성을 인식하기 위해 입술의 움직임과 같은 영상단서 정보를 분석하여 음성 인식 성능을 향상하는데 그 목적이 있다[10].



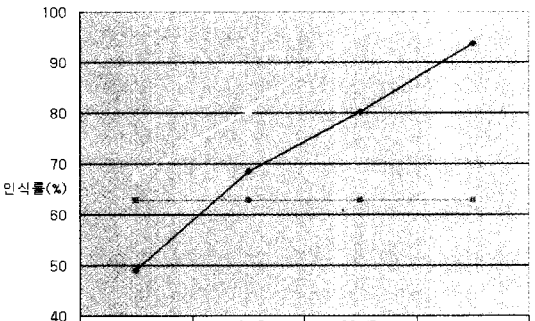
(그림 4) Automatic 멀티 모달 음성 인식 시스템 구조도

음성 인식 성능의 향상을 위해 사용하는 멀티모달 단서는 크게 음성과 영상 특징을 이용하며, 개략적인 시스템은 (그림 4)와 같이, 끝점 검출 모듈, 음성 특징 추출 모듈, 영상 특징추출 모듈과 인식 모듈로 구성되어 있다. 끝점 검출 모듈은 마이크로 입력된 음성 신호에 음성의 시작부분과 끝부분을 검출하는 모듈이며, 음성 및 영상 추출 모듈은 인식에 용이한 각각의 특징을 추출하는 모듈로 구성되어 있다.



(그림 5) 얼굴과 입술영역 검출과정

(그림 5)는 영상의 특징 추출을 위한 한 예로, 입술의 특징 추출은 피부색 모델을 통해 얻어진 얼굴의 이진화 영상을 타원에 근사시키는 방법을 찾아내게 된다. 이렇게 찾아진 얼굴을 기반으로 하여 입술색의 연결성 성분분석(connected component analysis)을 통해 사각형의 입술 영역을 검출하는 방법으로 입술영역을 추출하여, 입술 영역에 대한 고유공간(eigen space)상에서 각 영역의 가중치 벡터를 사용하여 영상 특징 벡터로 사용한다. 음성의 경우 기존에 알려진 다양한 방법으로 특징을 추출하여 적용하는 것이 가능하다.



	5 dB	10 dB	20 dB	clean
음성	49.07	68.44	80.25	93.84
영상	62.84	62.84	62.84	62.84
융합	67.54	80.65	96.08	98.33

(그림 6) 음성 영상 정보 융합 실험 결과

이렇게 추출된 영상과 음성의 특징은 TDNN 등의 기법을 통해 훈련하여 음성 인식을 수행하게 되는데 (그림 6)에서와 같이 음성과 영상 단서를 융합하여 음성인식을 시도하는 경우에 있어 인식률이 20%가량 향상하며 특히 잡음 환경하에서 기존의 단일 단서를 이용하는 음성 인식 시스템보다 비교적 안정적인 인식률을 보이는 것을 알 수 있다.

5. 결 론

음성 인식 기술에 기반한 음성 인터페이스 기술은 인터페이스의 혁명을 가져올 기술로, 실감형 미디어 기술의 기반기술이라 할 수 있는 기술이다. 이러한 음성 인식 기술은 오랜 연구가 되어온 분야이나, 잡음 환경에 취약한 면이 있어 아직까지 실질적인 적용을 어렵게 하는 면이 없지 않았다. 그러나 기존의 잡음환경을 개선하기 위한 방법들과 넥마이크와 멀티모달 단서를 이용한 음성 인식 시스템을 통하여 극복이 가능하며, 이러한 분야에 대한 집중적인 연구가 필요할 것으로 보인다. 또한 기존의 음성 인식 기술은 음성 인식 성능에 대한 연구를 위주로 이루어져 있었으나, 앞으로는 음성 인식 기술을 어떤 방식으로 실감형 미디어 서비스에 다양하게 적용할 수 있을 것인가에 대한 연구 또한 병행하는 것이 필요하다고 본다.

참고문헌

[1] Tsuhan Chen "The Past, Present, and Future of Speech Processing", IEEE Signal Processing Magazine MAY1998 24-48

[2] S. F. Boll, "Suppression of acousic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-27, pp. 113-120, Apr. 1979.

[3] H. Hermansky and N. Morgan, "RASTA processing of speech," IEEE Trans. Speech Audio Processing, vol. 2, pp. 578-589, Oct. 1994.

[4] F.-H. Liu, "Environment normalization for robust speech recognition using cepstral comparison," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, vol. 2, pp. 61-64, 1994.

[5] Gales, M.J.F.and Young, S.J., "Robust continuous speech recognition using parallel model combination," IEEE Trans. Speech and Audio Processing, vol. 4 pp. 352-359, Sep. 1996.

[6] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adapdation of continuous density hidden Markov models," Computer Speech and Language, vol. 9, pp. 171-185, 1995.

[7] 이상운, 한문성 "넥마이크로 음성신호의 분석 및 음성인식 적용에 관한 연구", 한국음향학회 추계학술발표대회 논문집, 제21권, 제2호, pp.31-34, 2002.

[8] 이연철, 한문성 "Automatic 멀티모달 음성 인식 시스템", 한국음향학회추계학술발표대회 논문집, 제21권, 제2호, pp.49-53, 2002.

[9] H. McGurk and J. MacDonald, Hearing lips and seeing voices, Nature, pp. 746-748, 1976.

[10]T. Chen, Audiovisual speech processing, IEEE Transactions on Signal Processing Magazine, pp. 9 -21, 2001.

저자약력

홍 훈 섭

2000년 연세대학교 전자공학과 (공학사)
2002년 연세대학교 전기전자공학과 (공학석사)
2002년-현재 한국전자통신연구원 컴퓨터소프트웨어연구소 지식처리팀 연구원
관심분야 : 화자 검증, 칼만 필터, 영상신호처리
이메일 : barney@etri.re.kr

이 연 철

2000년 밀양대학교 컴퓨터공학과 (공학사)
2002년 경북대학교 컴퓨터공학과 (공학석사)
2002년 ETRI 컴퓨터 소프트웨어연구소 지식처리팀 연구원
2002년-현재 휴먼미디어테크 연구원
주관심분야 : 패턴인식, 컴퓨터비전, 음성인식, 영상처리
이메일 : yclee@e-human.co.kr

이 상 운

1997년 경북대학교 전자공학과 (공학사)
1999년 경북대학교 대학원 전자공학과 (공학석사)
2001년 경북대학교 대학원 전자공학과 박사과정 수료
2000년-2002년 포항1대학 컴퓨터응용과 전임강사
2002년-현재 ETRI 컴퓨터 소프트웨어 연구소 지식처리팀 연구원
관심분야 : 음성인식, 신호처리
이메일 : lsu63479@etri.re.kr

한 문 성

1977년 서울대학교 수학과 (이학사)
1977년-1979년 전국 경제인 연합회
1980년-1981년 한국 IBM
1981년-1988년 미국 인디애나 대학교 컴퓨터과학과 박사과정 수료
1989년-현재 ETRI 컴퓨터 소프트웨어 연구소 지식처리팀 책임연구원
주관심분야 : 음성인식, 화자검증, 화자적응
이메일 : msh@etri.re.kr