□ 기술해설 □

이동전화 음성인식

삼성종합기술원 김상룡

1. 서 론

이미 우리곁을 맴돌며 출현의 서막을 예고한 정보화 사회는 몇몇 기술의 구현을 전제로 모습을 드러내고 있다. 그 예언적 기술중 하나가 무선통신의 세계 일체화와 단말기의 지능화일 것이며 대략의 모습은 그림 1과 같지 않을까 생각된다[1][2].

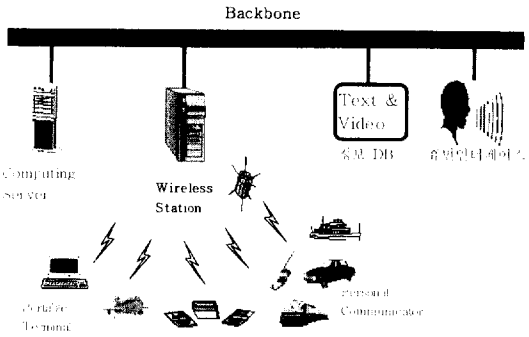

그림 1 정보화 사회의 통신인프라

정보화 사회의 진행 정도는 정보를 수집, 처리, 분배하는 기간망의 질적 기능과 정보 이용자의 망 인터페이스 방식에 따라 결정되는데, 본 글에서는 많은 기간망 중 Cellular망을 대상으로한 음성인식에 관하여 기술코자 한다.

이동통신망을 대표하는 Cellular망은 1983년 미국에서 민간에 상용화 서비스가 시작된 이래 세계적으로 년 40%의 고속 성장을 지속, 2001년에는 5억 9천만의 가입자를 예상하고 있다. 이같은 폭발적인 성장은 이동성을 보장해준 무선통신의 장점뿐 아니라 인공위성을 통한 세계통신의 일체화에 대한 기대감과 단말의 소형화, 지능화에 그 원인이 있을 것이다.

통신망의 효율적 이용을 위하여 인간은 가장 친숙한 정보 교환 수단 즉, 음성을 입출력의 도구로 사용하여 기계를 제어하고 정보를 얻고자 한다. 음성입력이란 음성인식을 통한 입력을 말하며 동시에 단말의 지능화를 의미한다.

음성인식 기술 발전에 있어서 1997년은 방향전환의 중요한 기점이 되는 해로 지금 까지의 대용량 단어 인식을 목표로 전개되던 개발 추세가 인간(사용자)의 자연성과 편리성을 고려한 기술 개발 추세로 전환되기 시작하였다. 즉 사용이 까다로운 대용량 고립어 인식보다는 용량이 작더라도 자연스러운 발성을 인식 대상으로 하는 기술과, 사용자의 환경이 실험실이 아닌 실제 환경에 적응시키는 기술로 연구의 초점이 이동하기 시작하였다.

이동통신망에서 음성인식은 단말에서 행하는가 본체에서 행하는가에 따라, 또 통신망이 아날로그냐 디지탈이냐에 따라 인식방법이 달라지는데, 궁극적으로는 아날로그 음성으로부터 인식에 필요한 특징점을 추출하여 인식하느냐 또는 표준 CODEC(Coder&Decoder)을 거쳐 얻어진 frame단위의 파라메타를 음성 특징점으로 하여 인식하느냐의 문제로 정리할 수 있다. 후자의 경우는 망의 특성에 따라 음성인식의 전처리부분(신호를 입력 받아서 특징벡터를 구하는 부분)이 일반적인 음성인식과는 판이하게 달라진다.

이에 따라 이동통신망의 현황과 표준 CODEC을 다음 단원에서 알아보고 통신망에서의 인식방법과 효과에 관한 장단점과 발전방향에 대해 기술코자 한다.

2. 이동통신망과 CODEC 표준화 [3][4][5][6]

이동통신을 대표하는 Cellular망은 휴대와 이동의 편리성으로 급속한 시장확대와 기술발전이 이루어졌으며, 송수신 신호에 따라 아날로그와 디지탈 방식으로 나누어지고 디지탈 셀룰라의 경우는 음성부호화 방식, 송수신 주파수, 채널의 access방식(CDMA, TDMA)과 서비스 지역에 따라 GSM(Global System for Mobile communications), PCS(Personal Communication System), PDC(Personal Digital Cellular)등으로 불리워 진다. 그림 2는 현재 서비스 중이거나 준비중인 이동통신망의 이동성과 전송률의 관계를 요약한 것이다.

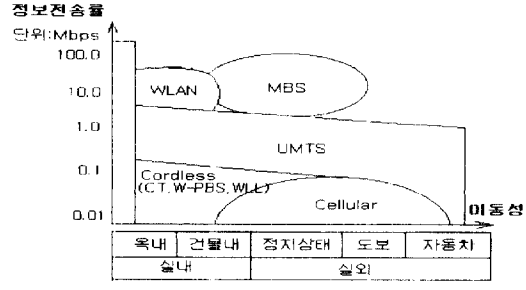

그림 2 단말의 이동성과 전송률

차세대 통신망 IMT-2000에 대해 Lucent, Motorola 등을 중심으로 한 WB-cdmaOne 그룹과 일본 및 유럽을 중심으로한 ARIB그룹간에 표준화 제정을 위한 연구 경쟁도 한창이다. 이외에도 UMTS(Universal Mobile Telecommunications System)과 MBS(Mobile Broad band Systems) 등이 유럽을 중심으로 연구되고 있다.

현재 운영중인 디지탈망의 표준 CODEC으로 TIA(Telecommu. Industry Association)에서는 이동통신의 표준화를 위하여 8kbit/s에서 CDMA(Code Division Multiple Access)를 위한 IS-96 Q-CELP (Qualcomm-CELP)와 TDMA(Time Division Multiple Access)를 위한 IS-54 VSELP (Vector Sum Excited Linear Prediction)를 발표하였다. ETSI(European Telecommunication Standards Institute)에서는 유럽의 다양한 언어와 민족으로 인하여 다양한 발성 방식의 사회적 환경때문에 13kbit/s에서 표준을 제정하고 있으며, GSM-FR(Groupe Speciale Mobile-Full Rate), RPE-LTP (Regular Pulse Excited-Long Term Predicion)가 이에 해당한다. 일본 RCR (Research and Development Center for Radio Systems)에서는 JDC-FR(Japan Digital Cellular-Full Rate) 6.7kbit/s VSELP와 이의 half rate인 JDC-HR 3.45kbit/s PSI-CELP(Pitch Synch. Innovation CELP)가 사용되고 있다.

그러나 이동 무선 통신망은 일반적으로 채널의 정체성으로 인하여 망 가입 수요의 무한 해결이 어렵고, 현재의 CODEC 음질성능 또한 그림 3에서 보는 바와 같이 통화 수준(Toll Quality)에 미치지 못하는 이중의 문제점을 안고 있다.

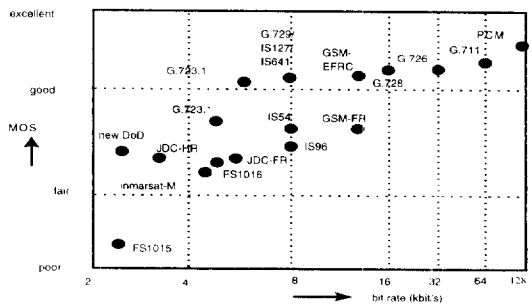

그림 3 국제 표준화 CODEC

이 문제점을 근원적으로 해결하기 위해서는 4kbps 이하의 저전송률로 G.726의 32kbps급 통화 품질을 가지는 압축기술을 확보해야 하나 현재로서는 이 규격을 만족시키는 기술을 가지고 있지 못한 실정이다. 현재는 음질의 문제점을 보인 8kbp의 Q-CELP와 전송률이 높은 유럽형 13kbps RPE-LTP를 각각 8kbps급의 EVRC(CDMA용), EFRC(GSM용)로 대체하는 현실적인 해결안을 구하는 실정이다.

3. 이동통신망에서의 음성인식 [7][8][9]

3.1 음성인식의 발전 방향

과거 음성인식의 주된 이슈는 얼마나 많은 단어를 인식하는냐였지만, 지금은 기계에 대한 사람의 접근을 보다 편리하게 하는 방향으로 흐르고 있다. 사람이 편하게 발성하는 말을 알아듣는 자연발성 음성인식 방법과 제한적인 단어이나 실제 환경에 적용 가능한 환경극복형 음성인식 방법들이 대두되고 있다. 다음의 표 1에 인식 기술의 현재 수준을 정리한다. 실제 음성 워드의 경우 1996년에 발표된 제품은 인식 어휘가 최대 100,000어이나, 사용자에게 분절적이고 명확한 발성을 요구하고 있어, 원래 목적인 사용의 편리성을 살리지 못하고 있는 실정이다. 또한 음성 다이얼링 등 소규모의 인식 기술은 상당히 높은 성능에도 불구하고 실제 환경인 자동차, 사무실, 거리, 공장 등 작업 현장에서 성능 저하가 매우 커서 실용화 전개에 어려움을 겪고 있었다.

표 1 음성인식과 자연어 처리후의 단어 오류율
(Courtesy : John Markhoul, BBN)

CORPUS	TYPE	어휘 크기	WER*
연속 숫자음	Spontaneous	10	0.3%
비행기 예약	Spontaneous	2,500	2.0%
월 스트리트 저널	Read Text	64,000	8.0%
방송(Marketplace)	Mixed	64,000	27%
정해진 대화 (Switchboard)	Conversational Telephone	10,000	38%
전화대화(Call Home)	Conversational Telephone	10,000	50%

WER* : 단어 오류율(Word Error Rate)

이에 따라 1998년의 인식시장은 단어수의 증가 경쟁보다는 잡음 제거/적응, 반향 제거, 발성 변화에 대한 적응 기술과 사용자 친화 기술인 자연스러운 대화체를 인식하는 기술개발의 격전장이 될 것으로 생각된다.

이동통신을 전제로 할 경우 음성이 입력되는 환경에 따라 인식성능과 편리성 확보는 심한 차이를 보인다. 일 예로 자동차 주행시는 도로, 차종, 차상태, 운전습관에 따라 잡음이 다르고 폐쇄 공간에서 발생하는 반향등으로 인한 환경 변화로 음성구간 검출은 물론 음성인식과 통화 자체가 어려운 경우도 발생한다. 그러나 인식의 악조건을 극복할 기술이 개발되면 음성인식을 통한 전화 다이얼링이나 navigation 시스템의 작동등은 자동차 운행중 안전성과 편이성을 획기적으로 증가시키며 이외에도 command and control, customer care 시스템 등에 유용하게 쓰일 것이다.

3.2 이동통신망에서의 음성인식

이동통신망은 아날로그망과 디지탈망이 있는데, 각각의 망에 따라 음성인식에 적용되는 음성신호처리 기술이 달라진다. 통신망에 따라 달라지는 음성인식 기술은 주로 음성신호의 입력에 관련된 신호처리 기술이다. 아날로그망에서의 음성인식은 주로 Filter Bank를 이용한 Band별 power, Cepstrum, LPC Cepstrum, MFCC(Mel frequency cepstrum) 등을 특징벡타로 이용하며, 디지탈망에서는 표준CODEC의 전송 파라메타를 이용한다.

각 이동통신망의 장 단점을 비교하면 아날로그망의 경우 기존에 연구된 음성인식 알고리듬을 그대로 수정없이 적용이 가능하고 필요한 음성정보를 최대로 이용할 수 있다는 장점이 있으나 음성인식을 위하여 추가적인 컴퓨팅 파워(Computing power)가 필요하다. 한편 디지탈망의 경우는 표준CODEC에 의해 압축과 재생에 적합한 파라미터로부터 인식에 필요한 특징을 구하므로 특징추출 등의 신호처리 부분이 필요치 않은 반면 인식에 필요한 충분한 정보를 얻기 힘들다.

따라서, 통신망의 특성에 따라 성능도 차이가 나는 데, 대체로 입력된 신호로부터 제약없이 정보를 얻을 수 있는 아날로그 신호에서 인식률이 좋다. 지금까지 발표된 자료에 의하면 자동차내에서 인식어가 40단어 정도이고 화자종속인 경우 아날로그망에서는 99%, 디지탈망에서는 97%정도로 보고되고 있다. 그러나 이동통신망의 발전방향이 아날로그에서 디지탈로 옮겨지고 있기 때문에 음성인식은 향후 디지탈망을 중심으로 한 기술이 발전할 것이다. 따라서 표준CODEC의 전송 파라미터로부터 음성인식을 위한 특징 등을 얻어내는 연구가 중요한 문제가 된다.

디지탈 이동통신망에서의 인식 기술은 적용대상에 따라 cellular, PCS 및 PDA 등의 단말기 적용 기술과 기지국의 교환기 및 각종 정보

서비스의 서버에 구현되는 시스템 적용 기술로 분류할 수 있다.

A. 단말기 적용 기술

음성으로 전화를 걸거나 단말기를 동작 시키는 음성 다이얼링기술이 현재 대표적인데, 보통 화자종속의 인식 기술이 채용된다. 즉, 사용자가 필요한 명령어와 발신하려는 곳의 이름을 사전에 음성으로 저장해 두고 이를 참조패턴으로 하여 인식을 수행하는 것을 의미한다. 이러한 단말기 적용 기술에서는, 기본적인 인식률 이외에도 계산량과 필요 메모리 양에 의해서 그 성능이 결정되는데, 이는 결국 단말기 생산의 경제성을 결정하게된다.

현재 적용되고 있는 단말기 적용 인식 기술은 인식 대상의 단어 수가 수십 단어 수준이며, 각 단말기 내부에서 이미 쓰여지고 있는 DSP 칩이나 별도의 인식 전용 칩에 구현되고 있다. 전자의 경우는 표준 CODEC용으로 사용되고 있는 DSP 칩을 이용하여 전치리부를 대신하여 음성인식 기능의 추가에 따른 비용을 최소화할 수 있는 방법이나, 왜곡된 정보를 이용하므로 인식률의 저하가 발생한다. 따라서 표준 CODEC용 파라미터로부터 효과적인 특징벡터를 얻어내는 방법이 성능을 좌우하게 된다. 인식 전용칩을 사용하는 경우는 정보의 유실이 없기 때문에 인식 성능에는 장점이 있으나 두 개의 DSP 칩(CODEC용과 인식용)을 사용하게 되므로 비용이 증가한다.

최근, 각 국에서 운전중 휴대폰 사용을 금지하는 법안의 제정이 확산 일로에 있음으로 하여, 기존 단말기에 hands-free 기능을 탑재하는 기술이 크게 대두되고 있다. 이에는 일반적으로 사용되는 음성 다이얼링 기술 이외에도 자동차 내부에서의 잡음 및 반향에 대한 처리와 원거리 마이크의 입력 신호 보상기술이 필수적으로 요구된다. 한편, 단말기의 도난 및 분실 시의 위험을 방지하기 위하여, 화자식별(speaker verification)에 대한 요구가 높아지고 있는데, 이에는 각 화자의 특성을 극명하게 반영할 수 있는 동적인 특징 추출이 요구된다.

B. 기지국 시스템 적용 기술

이는 매우 광범위한 용도로 사용되고 있는데, 이들은 현재 제공되는 서비스의 비용을 절감하기 위한 목적으로 사용되거나 새로운 서비스의 창출하기 위한 목적으로 쓰인다. 먼저, 서비스의 비용을 절감하기 위하여 적용된 예를 살펴보면, 자동교환 서비스(automatic operator service), 자동 전화번호 안내(automatic directory assistance), 음성 다이얼링 등을 들 수 있다. 자동교환 서비스는 수신자 부담 통화, 교환원을 통한 통화, 신용카드 통화, 개인 대 개인 통화와 같은 특수 통화 서비스를 수행하는데, AT&T의 Voice Recognition Call Processing(VRCP)와 Nortel의 Automated Alternate Billing System (AABS)이 현재 사용되고 있는 서비스의 예로 교환원의 업무를 절감하는 기능을 수행한다. 자동 전화번호 안내는 현재 NYNEX와 Nortel 등에서 제공되고 있는데, 안내원이 찾아주는 전화번호의 탐색대상을 줄여주는 역할을 수행한다. 예를 들면, 지역번호 및 한정된 수의 상호에 대한 인식 정보가 그것이다. 음성 다이얼링 기능은 단말기에서의 역할과 비슷하지만, 그 기능이 시스템에 구현되어 개개의 단말기에 적용된 인식기능을 대신하는 장점이 있으나 선로의 신호 왜곡에 의해 성능 저하가 불가피하다. 현재, AT&T, NYNEX, Bell Atlantic 등에서는 이름을 통한 다이얼링 서비스(voice dialing by name)가 실시 중이며 AT&T SDN/NRA에서는 번호를 통한 다이얼링 서비스(voice dialing by number)가 제공되고 있다. 이 분야에서도, 화자식별에 대한 요구가 높은데, 현재 통신 도용으로 인한 손실이 엄청나기 때문이다.

C. 서비스 분야

한편, 새로운 서비스를 창출한 부분을 살펴보면, 음성 banking 서비스, voice prompter, directory assistant call completion, reverse directory assistance, 정보검색 서비스 등을 들 수있다. 음성 banking 서비스는 전화 음성인식을 통하여 계좌를 조회하고 각종 입출금 업무를 수행하는데, NTT에서는 이미 십년 전에 이러한 서비스를 제공하는 시스템을 개발하였고, 현재는 세계 여러 은행들이 채용을 준비하고 있다. Voice prompter는 현재 전화기에서 제공 중인 touch tone 입력을 대치하는 역할을 수행하는데, 이는 touch tone 제공 전화기의

보급이 상대적으로 적은 스페인 등에서 먼저 실시되었다. Directory assistant call completion은 전화 번호 안내가 끝난 후에 자동으로 안내된 번호에 접속하도록 하는 서비스로 일반적으로 TTS(text-to-speech)를 통하여 출력된 음성을 인식하게 되며, AT&T와 NYNEX 등에서 실시중이다. Reverse directory assistance는 전화 번호 안내와는 반대로 전화 번호를 인식하여 주소, 인명 또는 상호 정보를 제공하는 것으로 NYNEX, Bellcore, Ameritech 등에서 개발된 바 있다.

마지막으로, 정보검색 서비스는 매우 널리 사용되고 있는 응용처로, 현재는 주식조회, 날씨 안내, 교통 정보, 표 예매 등에 적용되고 있는데, 이들 응용을 위해서는 주어진 음성중에서 원하는 부분만을 추출하는 word spotting 기술이 필수적이다.

4. 전 망

지금까지는, 현재 개발되거나 실시 중인 응용 분야를 살펴보았다. 이를 토대로 미래에 출현할 기술 또는 응용 분야를 살펴보면, 단말기 적용 기술에서는 지능형 단말기의 출현이 예상된다. PDA와 같이, 통신 단말기들은 음성통신 기능뿐만 아니라 각종정보의 교환에 적합한 형태로 진화할 것이다. 이러한 진화에는 음성인식이 필수적이다. 따라서, 단말기의 OS와 연동되는 음성인식 기술 및 대화처리가 가능해야 한다. 이러한 기술은 각 단말기에 표준적으로 채택될 OS를 기반으로 개발하여야 하며 다른 정보기기와의 호환에도 신경을 써야 할 것이다.

또한 시스템 적용기술에 있어서는, agent technology, customer care, computer-telephony integration, voice dictation 등에 대한 관심이 고조될 것으로 전망된다. 우선, agent technology는 사용자와의 대화를 통하여 각종 통화 서비스, 인터넷에서의 정보검색, 특정 정보 제공, 사용법 공지 등의 업무를 지능적으로 수행하는 응용 분야이다.

Customer care는 기존의 ARS를 대화를 통한 쉬운 interface로 대체하여 사용자의 요구를 더욱 쉽고 편안하게 유도하는 것으로, AT&T에서는 How May I Help You(HMIHY)와 같은 시스템을 개발한 적이 있다. Computer-telephony integration은, 앞으로는 기존의 전화망과 패킷망의 결합이 점점 가중될 것이라는 전망에 기초를 두는 것으로, 전화를 통하여 각종 서비스를 제공하는 컴퓨터에 접속하여, 예약, 등록, 시스템 정보 제공 등의 업무를 가능하게 함을 의미한다. Voice dictation은 현재에는 PC를 비롯한 각종 컴퓨터에서의 사용을 위하여 판매 혹은 개발되고 있지만, 앞으로는 통신 시스템에도 적용되어 e-mail을 비롯한 각종 message와 정보 검색을 위해 사용될 것이다.

단말의 경우는 유무선 통신이 가능하고 휴대가 간편한 wearable형의 출현이 기대되며 입출력 방식도 키패드등을 대신하여 인간의 언어와 감성을 직접 입력받고, 언어와 소리 등으로 오감에 직접 전달하는 기술이 크게 부각될 것으로 예측된다.

참고문헌

[1] B. Barringer, T. Burd, et. al., "Infopad : A System Design for Protable Multimedia Access", 14pages, http://*infopad.EECS.Berkeley.edu/..infopad-system-design. wireless 94/*, 1994.

[2] M. Madfors, K. Wallstedt, S. Magnusson, H. Olofsson, P. Backman, and S. Engstrom, "High Capacity with Limited Spectrum in Cellular Systems", *IEEE Comm. Mag.* Vol. 35, No. 8, pp. 38~45, Aug. 1997.

[3] J.S. Dasilva, B. Arroyo, B. Barani, and D. Ikonomou, "European Third-Generation Mobile Systems", *IEEE. Comm. Mag.* Vol. 34, No. 10, pp. 68~83, Oct., 1996

[4] Falconer, F. Adachi, and B. Gudmundson, "Time Division Multiple Access Methods for Wireless Personal Communications", *IEEE Comm. Mag.* Vol. 33, No. 1, Jan. 1995.

[5] 김상룡, "음성처리 기술의 현황과 미래", 한국음향학회 학술발표대회 논문집, 연세대학교, vol. 16, No. 2, 11월, 1997.
[6] Richard V. Cox, "Three New Speech Coders from the ITU Cover a Range of Applications", *IEEE Comm. Mag.*, vol. 35, No. 9, pp. 40~47, Sep. 1997.
[7] Lawrence R. Rabiner, "Applications of Speech Recognition in the Area of Telecommunications", *IEEE Workshop on Automatic Speech Recognition and Understanding Proc.* CA. Dec. 1997.
[8] C. A. Kamm, M. Walker, and L. R. Rabiner, "The Role of Speech Processing in Human-Computer Intelligent Communication", *Proc. HCI Workshop*, Washington, DC, Feb. 1997.
[9] L. R. Rabiner, 'Applications of Voice Processing to Telecommunications", *Proc. IEEE*, Vol. 82, No. 4, pp. 199~228, Feb. 1994.

김 상 룡

1980 한국항공대학 전자공학과 학사
1980~현재 삼성전자(종합기술원) 연구위원
1982 한국과학기술원 전기 및 전자공학과 석사
1989 한국과학기술원 전기 및 전자공학과 박사
관심분야 : 음성처리(인식, 합성, 압축), 언어이해, 영상이해, 음성/데이타통신