특 집

이동단말기용 음성인식 전처리 및 UI 표준화

김 재 인*, 이 수 종**

*KT 음성인식서비스 개발팀, **한국전자통신연구원 음성DB연구팀

I. 서 론

무선통신기기의 보급이 확산되면서 사람들은 장소에 제약을 받지 않고 원하는 사람과 통화를 할 수 있게 되었으며, 무선통신망이나 무선LAN AP(Access Point)를 이용하여 무선으로 인터넷에 접속할 수 있게 되었다. 하지만 이동 단말기 특성상 휴대하기 편리해야 하기 때문에 크기가 작아야만 하고, 그래서 글자를 입력하는 보편적인 수단인 키보드의 장착과 사용이 쉽지 않기 때문에 이를 음성인식기술을 이용하여 해결하고자 노력하고 있다. 하지만 통신이 유선환경에서 무선으로 바뀌면서 이동단말기의 사용환경도 굉장히 다양해졌기 때문에 주변잡음이 음성인식시스템의 성능에 미치는 영향 때문에 사용장소의 제약이 좀처럼 풀리지 않고 있다. 이러한 문제점을 해결하기 위해 음성인식 전처리부에 잡음에 강인한 음성 특징파라미터를 추출하거나, 잡음환경에 적응시키는 방법 등을 사용하여 오랫동안 노력해 왔으나, the European Telecommunication Standards Institute(ETSI) Speech processing, Transmission and Quality aspects (STQ)-Aurora group 이 결성된 후, 이 방면에 대한 연구가 좀더 체계적으로 진행되고 있다. 이 그룹에서는 특히 Distributed Speech Recognition(DSR) 방식 즉 어떤 종류의 통신단말에서나 전처리부만 처리하고, 인식은 중앙처리 방식을 사용하는 시스템을 대상으로 잡음에 강인한 전처리부를 만들기 위해 노력하고 있다. 여기서 전처리란 음성입력의 시작점과 끝점을 찾고, 음성구간을 음성특징파라미터로 변환하는 등 음성인식을 수행하기까지 음성인식시스템에서 수행하는 일을 말한다. 본 고에서는 이동단말기에서의 음성인식 전처리 연구가 활발히 진행되고 있는 Aurora Project의 최근 성과에 대하여 알아보기로 한다. 이어서 음성인식기술이 새로운 유망기술로 부상하고 개인휴대단말기의 보급확대에 따라 음성인터페이스를 적용하게 되면서 사용자 인터페이스를 설계하고 사용자를 대상으로 사용성을 검증하는 일련의 절차에 대해서 살펴보고 음성명령어에 관하여 알아보기로 한다.

II. 음성인식 전처리 연구

서론에서 언급한 대로 ETSI가 DSR에 대한 전처리 표준을 만들기 위하여 작업을 시작했으며 연도별 연구성과 및 음성DB는 다음과 같다.

1. 2000년 2월 : 첫번째 ETSI standard DSR front-end(ES 201 108)

1) Mel-Cepstrum의 DSR 표준을 제안하였다.

2) Aurora 1 음성DB

- TI digits는 미국식 영어를 구사하는 화자들이 연속 숫자를 발성한 것이며 이것을 8kHz로 down sampling 하였다.
- TI digits DB를 DSR 통신환경과 유사한 디지털 필터에 통과시키고 나서, 임의로 만

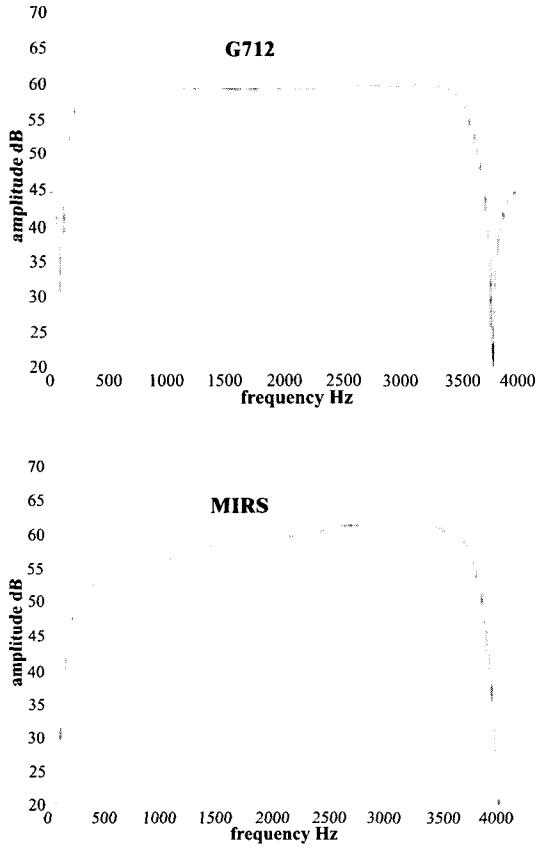
들어진 잡음을 20db에서 −5db까지 첨가시켜서 만들었다.

3) 디지털 필터

- ITU에서 제안한 두 개의 표준 주파수 특성을 사용하였다.[1]
- <그림 1>에서 보면 G712가 MIRS보다 300-3400Hz 구간에서 전달특성이 평탄한 것을 알 수 있다.
- MIRS은 GSM 03.50의 기술요구사항과 일치되는 특성을 갖고 있다.
- 이 두 가지 필터의 특성은 ITU STL96 소프트웨어 패키지로 구현할 수 있다.

4) 성능시험

- HTK(Hidden Markov Model Tool Kit)를 사용한 인식기[2]로 특징 파라미터의 성능만을 검증했다.



<그림 1> G712와 MIRS 필터의 주파수 응답특성

2. 2002년 10월 : 두번째 ETSI standard (ES 202 050)

1) 주변 잡음에 대한 인식성능을 향상시켰다.

2) ES 201 108에 비해 53% 에러가 감소되었다.

3) Aurora 2 음성DB

- TI digits DB를 역시 통신망 환경을 시뮬레이션 할 수 있는 digital filter를 통과시키고, 여기에 8가지의 실제환경에서 녹음한 잡음들을 20db, 15db, 10db, 5db, 0db, −5db로 첨가시켰다.
- 잡음을 첨가 시 SNR(signal-to-noise ratio)을 구하는 것이 선택된 주파수 범위에 의존하기 때문에 음성에너지나 잡음에너지는 ITU 권고안 P.56[3]을 적용한 ITU software를 이용하여 구했다.

4) 잡음이 녹음된 환경의 종류

- Suburban train
- Crowd of people(babble)
- Car
- Exhibition hall
- Restaurant
- Street
- Airport
- Train station

3. Aurora 3 음성DB

- 2002년에 만들어졌다.
- 차의 실제 환경에서 음성DB를 구축하였다.

4. Aurora 4 음성DB

- Aurora 1, 2, 3은 저용량 어휘로 구성되어 있다.
- "Eurospeech 2003"을 위해 만들어졌다.
- 5000 단어 WSJ(Wall Street Journal)에 모의 잡음을 섞었다.

Ⅲ. AURORA WI007 FRONT-END

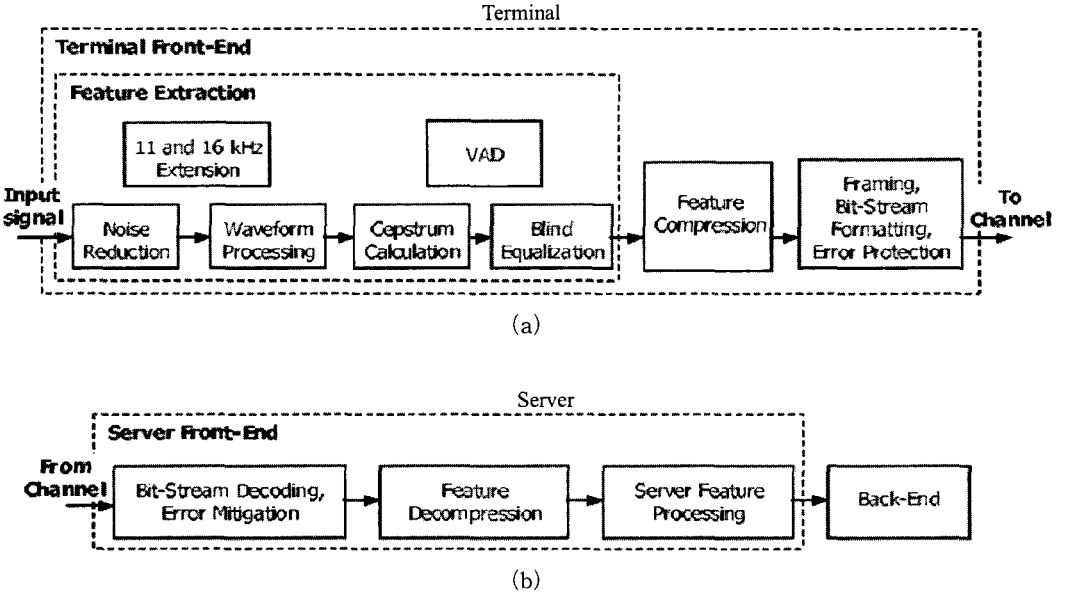
이것은 14개 Mel frequency cepstral coefficients(MFCCs)를 구하는데 자세한 내역은 다음과 같다.

가. 분석프레임 길이는 25ms이고 10ms씩 이동한다.
나. 0차부터 13차의 MFCC를 포함하므로 14개를 사용한다.
다. DC를 제거하는 모듈이 포함되어 있다.
라. 프리 엠파시스 계수는 0.97이다.
마. FFT 기반 Mel filterbank는 64Hz에서 샘플링 주파수의 절반까지 23개 주파수 밴드를 갖는다.
바. 분석된 파라미터들은 4800 Bit/s가 되도록 코딩되며, 14개의 음성특징 파라미터는 코딩되어 프레임별로 44 Bit로 만들어진다. [4]

〈그림 2〉는 Aurora 2에 대한 final draft ETSI ES 202 050[5]의 블록 다이어그램[5]인데 이것은 다음의 3개의 부분으로 나누어져 있다.

1. Noise Reduction 부분
2. Waveform processing
3. Cepstrum calculation

표준안으로 제안된 모듈의 성능은 참고문헌[6]에 나와 있듯이 잡음환경에서 Aurora I에 비해 평균 65.14% 향상되고 있는 것으로 나와 있으며, 국내에서 발표된 논문[7] 결과 역시 참고문헌[6]보다는 못하지만 50.67%인식률이 향상된 것으로 보고 되고 있다. Final draft에 대한 시험들은 잡음이 섞인 신호가 이동 단말기에서 처리되어 음성 특징파라미터로 변환되어 전송되는 경우이므로 전송로의 잡음이 섞이지 않는 경우에 대한 것이다. 만일 단말기에서 특징파라미터 변환이 없이 중앙처리 시스템에서 전송되어 온 음성을 특징파라미터 분석부터 인식까지 수행하는 경우의 결과는 전자의 경우와 조금 차이가 날 것으로 생각된다.



〈그림 2〉 Block scheme of the proposed front-end. (a) shows blocks implemented at the terminal side and (b) shows blocks implemented at the server side

Ⅳ. 음성인터페이스의 설계

다음은 음성인터페이스의 설계에서 꼭 고려되어야 할 사항들에 대하여 알아보기로 한다. 그 첫 번째는 사용자들은 GUI(Graphic User Interface)에서와는 달리 음성언어 응용에서의 모든 오류를 시스템 오류로 간주한다는 것이다. 따라서, 이들의 오류를 어떻게 효율적으로 관리하느냐가 중요하다. 둘째, 사용자가 필요로 하는 업무, 수행업무의 환경이 너무 다양하고 변화가 많으므로 반복적으로 시험하는 것이 중요하며, 이를 통하지 않고서는 좋은 설계를 기대할 수 없다. 셋째, 사용자의 정신적 및 신체적 한계가 있다는 것을 인식하여 사용하기 쉽도록 하고, 정보접속의 용이성을 추구해야 한다. 넷째, 사용자들이 기본적으로 갖고 있는 멘털모델과 조화되어야 하고 호감이 가도록 편의성을 제공해야 하며, 적합한 속도의 유지와 반응(response), 음성의 일시성에 따른 보완책 강구 및 필요한 정보의 피드백이 이루어져야 한다. 다섯째는 사용자와 컴퓨터 시스템 간의 상호작용으로서 대화방식이 필요하며, 다중모드 대화방식이 에러를 취급할 때 더 효과적이며 자연스런 방법이 될 수 있다.

음성인식기술이 사용자에게 주는 매력은 최종적으로는 음성인식기와의 대화방식 인터페이스에 좌우된다. 대화가 매끄럽지 않거나 모호하여 어떻게 반응해야 할지 확신할 수 없도록 한다면 중대한 문제가 된다. 유연한 대화흐름과 간결 명료하고 적절한 속도의 유지는 사용자와의 신뢰 구축과 함께 흥미로운 경험을 갖게 한다. 음성인터페이스 연구와 설계 및 사용자 검증에 관하여 세부 절차를 차례로 알아본다.

제1단계는 음성인터페이스를 위한 연구 과정이다. 먼저, 사용자가 음성서비스로부터 무엇을 원하는 지에 관하여 알아야 한다. 사용자에게 원하는 바를 정확하게 도와주도록 설계될 때, 음성서비스 제공자와 사용자 모두의 필요를 충족시킬 수 있다. 둘째는 사용자들이 해결하고자 하는 문제를 해결할 수 있는 것으로서 어떤 것들이 존재하는 지를 파악하여 비교 분석해야 한다. 유사한 제품, 기술 혹은 시스템을 찾아보고 비슷한 점과 차이점을 서로 비교해야 한다. 이 분석의 결과는 개발과정에 있는 음성서비스를 위해서 최선의 사용자 인터페이스를 만드는데 도움을 줄 수 있다. 셋째는 사용자의 의견을 묻는 것이다. 인터뷰와 질문이나 설문형식을 통하여 음성시스템이 그들을 위해서 무엇을 해 주기를 원하는지를 결정하는데 도움을 받을 수 있다. 이러한 과정을 거친 후 네번째 과정은 설계의 윤곽을 그리고 현실화해야 한다. 인터페이스 명세서 작성 및 서비스 시기, 제약사항, 옵션의 설정, 정보흐름, 설계 스케줄 등을 마련해야 한다. 다섯번째, 서비스에 내재될 수 있는 위험분석을 해야 하고, 설계의 우선순위를 부여해야 한다. 복구가능 에러와 복구불가능 에러를 구분해야 하며, 복구 불가능 에러가 발생하지 않도록 설계되어야 한다.

제2단계는 실제 설계과정에 관한 것이다. 첫째, 독창적인 설계를 이끌어 내야 한다. 지향하고자 하는 음성서비스에 접근하기 위한 다른 모든 방법을 분석해야 하고, 해결되어야 할 문제를 해결할 수 있도록 지혜를 동원해야 한다. 둘째는 개념적인 설계로부터 구체적이고 완전한 설계로 좁혀가는 과정이다. 일련의 시나리오를 만들고 피드백을 위해서 스크립트를 작성하고 정제한다. 셋째는 시제품을 설계한다. 모든 설계를 문서화하며, 프로그래머가 작업할 수 있도록 하고, 음성서비스의 프로토타입을 제작하고 구축한다.

제3단계는 사용자를 대상으로 하는 마지막 검증 과정이다. 음성서비스를 통하여 사용자 인터페이스 설계결과를 시험하여 그 동작 상태를 점검해야 한다. 사용자들이 음성서비스를 효율적으로 사용하고 있는지 확인하기 위하여 시험하는 것이다. 소규모의 인원을 그룹으로 구성하여 유용성 시험을 통하여 그 결과를 반영하여 개량하고, 문제가 있다면 재설계하고 다시 시험하는 절차를 반복함으로써 최상의 결과를 도출해가야 한다.

V. 음성인터페이스 명령어

사용자는 궁극적으로 음성명령어를 통하여 음성서비스를 이용하게 된다. 유럽에서는 ETSI를 중심으로 Task Force 182를 구성하여 전화망 음성서비스 상에서 네비게이션, 정보검색, 콜핸들링, 사용자 설정기능을 위한 유럽 5개 국어(영국, 프랑스, 독일, 이탈리아, 스페인)로 음성명령, 제어, 편집용의 음성사용자와 정보통신 디바이스 및 서비스간의 인터페이스 표준을 위한 음성명령어 표준화 작업을 진행하여 왔다. 국내에서도 개인휴대단말기로서 관심을 모으고 있는 PDA에 적용하기 위한 음성명령어 선정 작업이 진행 중에 있다. 명령어를 선정하기 위한 기준은 선정의 주체와 필요에 따라 여러 가지로 분류할 수 있으나 첫째, 사용자에게 가장 익숙한 단어, 둘째, 길이가 짧은 단어, 셋째 한글선호에 따른 외래어 배제, 넷째, 단어간 변별력 유지를 주안점으로 할 수 있다. 또한, 명령형이나 청유형의 문체를 선호한다는 점을 고려하여야 한다. 앞에서 논의된 사용자 인터페이스 설계절차와 연관하여 PDA를 중심으로 한국어 음성명령어 선정을 위한 노력이 음성처리산업협의회를 중심으로 2002년 시작되었다. 여기서는 공통명령어, 특정명령어를 제안하고 있다. 공통명령어에서는 세부적으로 핵심명령어, 확장명령어, 엔진 제어용 명령어로 구분한다. 특정명령어에는 네비게이션용, 미디어제어용, 편집용 등이 있다. PDA를 중심으로 한국어 음성명령어 선정을 위한 주요항목을 살펴보면 다음과 같다.

공통명령어는 사용자가 어떤 음성어휘를 사용

〈음성명령어 선정 시 고려사항〉

◇ 필요한 최소한의 명령어 개수 (아직은 음성인식률이 높지 않음을 고려)
◇ 짧은 음절일수록 인식오류 증가
◇ 기능의 의미를 잘 나타낼 수 있는 용어
◇ 사용자 오류에 대비하여 변별력 있는 명령어 (유사발음 명령어 제외)
◇ 사용빈도를 고려한 익숙하고 친숙한 명령어
◇ 하위메뉴 내에 있는 기능을 직접 불러내어 수행할 명령어
◇ 인터페이스 수단의 조화 및 협동 (소프트키보드, 펜 마우스, 음성)

〈사용자 편의성 평가방법〉

◇ 전문가 평가 : 음성인식시스템의 각 평가 대상들이 사용 편의성 원칙이나 지침에 따라 설계되었는지 여부를 전문가의 판단으로 평가
◇ 벤치마크 평가 : 대표적인 작업을 Benchmark Task로 선정하여 사용자에게 수행해 보면서 문제점 파악
◇ 사용자 평가 : 사용자에게 시스템을 자유롭게 사용하게 함으로써 사용 중 발생하는 문제점 파악

〈음성적용 시스템기능 요구사항〉

◇ 음성명령어 기능실행 여부를 사용자가 버튼조작 등의 손쉬운 방법으로 선택
◇ 음성명령어와 펜 마우스를 상황에 따라 선택적으로 사용
◇ 음성명령이 적용된 아이콘/기능을 시각적으로 병행표시 (스피커, 마이크, 입술모양,...)
◇ 음성명령을 실행대기/선택된 아이콘/기능을 강조(Highlight)하여 표시
◇ 음성명령어 그룹화 (기본명령어, 특정서비스 명령어, 1st 메뉴, 2nd 메뉴, 작업표시줄, 툴바)
◇ 음성명령어 정렬기능 (자모순, 자주/최근 사용한 명령어,...)
◇ 음성명령어 출력확인 및 검색기능, 기능설명
◇ 음성단축 다이얼링 기능 검토

〈음성명령어 선호도 조사분석〉

◇ 목적
- 공통음성명령어를 중심으로 실제 활용 가능성 및 선호하는 명령어 선별
- 기존 핀마우스 기반의 명령어에 음성을 적용함에 따른 음성대체명령어 발굴
- 활용 가능한 기본적인 음성명령어 적용을 바탕으로 음성명령어를 염두에 둔 서비스 개발 유도

◇ 조사대상
- 사용자그룹 : 기기에 대한 사용지식과 함께 실제 사용중인 자를 대상으로, 주어진 명령어를 제시한 후 명령어 상호간 선호도 조사
- 비 사용자그룹 : 기기에 대한 상식은 있으나 활용 경험이 거의 없는 자를 대상으로, 주어진 명령어를 포함하여 다양한 대체어를 제시한 후 선호도 조사

◇ 결과요약
- 기본기능 외에 인터넷, 이동전화, 오디오 등으로 서비스 영역이 확대되고 있으므로, 이들 서비스 각각에 대하여도 공통으로 제어하는데 적합한 명령어가 제시되어야 하겠음.
- 특정서비스 자체에 대하여는 각 개발자 및 사용자가 개별적인 특성을 감안하여 가변적으로 설정하여 활용할 수 있도록 공통명령어에서는 제외
- 선호도 조사결과에 따라 선호도가 높은 대체어를 사용하되, 어학적 및 음성학적 연구는 계속 검토

하여 기기를 제어할 것인가의 기준을 마련하는 것이며, 사용자로 하여금 다양한 상황에서의 음성인터페이스를 활용하도록 지원하는 한편, 각양의 제품과 서비스를 일관성 있고 체계적으로 제어할 수 있도록 편의를 제공하는 동시에 개발자로 하여금 GUI와 음성을 동시에 적용하는데 적합한 서비스 개발을 유도하는데 그 중요성이 있다.

VI. 결 론

이제까지 이동단말기 중 가장 널리 보급되어 있는 무선단말기에 대한 전처리 연구와 음성인터페이스 연구에 대한 진행사항에 대하여 알아보았다. 전처리 연구에서는 2000년부터 시작된 ETSI의 Aurora 그룹의 연구를 중심으로 알아 보았다. 국내에서는 음성에 섞인 잡음에 대한 처리를 음성 특징파라미터로 극복하려는 시도가 많았었고, 최근 산업계에서는 잡음처리 성능이 우수한 모듈을 음성인식엔진 앞에 붙여서 실용화하려는 노력이 시도되고 있다. 하지만 Aurora의 연구결과를 좀 더 깊숙이 검토하여 실용성 여부를 객관적으로 판단하고 그 결과에 따른 대응노력이 필요한 시점이라 하겠다. 음성인터페이스 연구에서는 음성인터페이스를 설계하면서 고려해야 할 사항과 절차 그리고 범용으로 적용하기 위한 음성명령어 선정에 관한 주요항목에 대하여 살펴보았다. 음성기술의 개발과 응용과정에서 음성인터페이스 기술이 더욱 체계적으로 연구되고 설계와 구현과정에 반영되어야 하겠다.

참 고 문 헌

〔1〕 ITU recommendation G. 712, "Transmission performance characteristics of pulse code modulation channels", Nov. 1996

〔2〕 S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland, "The HTK book version3.0, Entropic, 2000

〔3〕 ITU recommendation P. 56 "Objective measurement of active speech level", Mar. 1993

〔4〕 ETSI standard document, "Speech Processing, Transmission and Quality

aspects(STQ) ; Distributed speech recognition ; Front-end feature extraction algorithm ; Compression algorithm", ETSI ES 201 108 v1.1.1(2000-02), Feb. 2000

〔5〕 ETSI ES 202 050 V1.1.1, "Speech Processing Transmission and Quality aspects : Distributed speech recognition ; Advanced front-end feature extraction algorithm ; Compression algorithms", 2002

〔6〕 Dusan Macho, Laurent Mauuary, eltal, "Evaluation of a noise-robust DSR front-end on a Aurora Databases, "Proc. ICSLP 2002, volume I, pp.17-20, Sep. 2002

〔7〕 김규홍, 김회린, "전화망 환경에서 한국어 숫자음 인식을 위한 잡음처리, " 대한음성학회 봄 학술대회 발표 논문집, pp.211-214, 2003

〔8〕 Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, "Spoken Language Processing", Prentice-Hall, inc. 2001.

〔9〕 Lawrence Rabiner, Biing-Hwang Juang, "Fundamentals of Speech Recognition", Published by Prentice Hall PTR, 1993.

〔10〕 이수종, 박문환, 김상훈, 이영직, "음성인식 시스템의 사용자 편의성 분석", 한국음향학회 제19회 음성통신 및 신호처리학술대회, 2002. 8.

저 자 소 개

김 재 인

1981년 2월 고려대학교 전자공학과 졸업(학사), 1986년 8월 고려대학교 대학원 졸업(석사), 1996년 2월 고려대학교 대학원 졸업(박사), 1986년 9월~1988년 3월 : 금성(현LG)전기 연구소 근무, 1988년 3월~현재 : 케이티 마케팅기획본부 서비스 개발연구소 음성서비스 개발팀, 〈주관심 분야 : 음성인식, 음성합성, VoiceXML platform〉

이 수 종

1984년 2월 충남대학교 경제학과 졸업(학사), 1990년 8월 건국대학교 경제학과 졸업(석사), 2003년(현재) 한밭대학교 정보통신공학과(박사과정), 1984년 3월~2001년 12월 : 한국전자통신연구원 네트워크연구소, 2002년 1월~현재 : 한국전자통신연구원 컴퓨터소프트웨어연구소 음성DB연구팀 (책임연구원), 〈주관심 분야 : 음성인식, 음성합성, 멀티모달 인터페이스〉