자바를 이용한 음성인식 시스템에 관한 연구

Study of Speech Recognition System Using the Java

최 광 국*, 김 철*, 최 승 호*, 김 진 영**
(Kwang-Kook Choi*, Cheol Kim*, Seung-Ho Choi*, Jin-Young Kim**)

* 동신대학교 정보통신공학과, ** 전남대학교 전자공학과
(접수일자: 2000년 5월 15일; 채택일자: 2000년 7월 18일)

본 논문에서는 자바를 사용하여 연속분포 HMM 알고리즘과 Browser-embedded 모델로 음성인식시스템을 구현하였다. 이 시스템은 웹상에서 음성분석, 처리, 인식과정을 실행할 수 있도록 설계되었으며, 클라이언트에서는 자바애플릿을 이용하여 음성의 끝점검출과 MFCC와 에너지 그리고 델타계수들을 추출하여 소켓을 통해 서버로 전송하고, 서버는 HMM 인식기와 학습DB를 이용하여 인식을 수행하고 인식된 결과는 클라이언트에 전송되어 문자로 출력되어진다. 또한 이 시스템은 플랫폼에 독립적인 시스템으로 네트웍상에서 구축되었기 때문에 높은 에러율을 갖고 있지만 멀티미디어 분야에 접목시켰다는 의의와 향후에 새로운 정보통신 서비스가 될 가능성이 있음을 알 수 있었다.
핵심용어: 연속분포 HMM, 브라우저 삽입 모델, 자바애플릿, 멀티미디어
투고분야: 음성처리 분야(2.5)

In this paper, we implement the speech recognition system based on the continuous distribution HMM and Browser-embedded model using the Java. That is developed for the speech analysis, processing and recognition on the Web. Client sends server through the socket to the speech informations that extracting of end-point detection, MFCC, energy and delta coefficients using the Java Applet. The server consists of the HMM recognizer and trained DB which recognizes the speech and display the recognized text back to the client. Because of speech recognition system using the java is high error rate, the platform is independent of system on the network. But the meaning of implemented system is merged into multi-media parts and shows new information and communication service possibility in the future.
Key words: Continuous HMM, Browser-embedded model, Java Applet, Multi-Media

I. 서 론

음성인식이란 기계나 컴퓨터가 사람의 말을 알아듣게 하는 기술을 말한다. 이 기술에 대한 본격적인 연구는 컴퓨터의 성능이 향상되기 시작한 1960년 이후부터였다. 그 당시의 연구 목표는 특정한 화자를 대상으로 10개의 숫자음을 인식하는 정도였지만, 현재는 회화체를 인식할 수 있는 수준까지 도달된 상태이다.

최근 음성처리 분야에서는 사운드와 비전을 함께 다룰 수 있는 멀티미디어 컴퓨터 네트워킹 기술이 새롭게 등장하고 있으며, 인터넷의 활성화에 따라 네트웍상에서 음성인식을 수행할 수 있는 프로그램인 Voicepower for internet browsing, Incube for IE 4.0 등의 제품들이 출현되고 있다. 특히, 미국의 SUN Microsystems에서 자바를 이용한 음성인식/합성/오디오 제어기능을 갖는 JSAPI(Java Speech Application Programmer's Interface)가 '98년에 발표됨으로써 인터넷상에서 음성인식시스템이 구현되기 시작하는 원동력이 되었다.

자바의 음성처리분야에 대한 접목은 M. Sokolov의 "Speaker Verification on the world wide web"이 1997년도 Eurospeech에 발표되었으며, V. Digalakis, L. Neumeyer, Perakakis의 "Quantization of Cepstral Parameter for Speech Recognition Over the World Wide Web"이라는 제목으로 1998년도 ICASSP에 발표되었다[1][2].

II. 음성분석 방법

2.1. 실시간 끝점검출 알고리즘

음성에 포함되어 있는 정보를 나타내기 위하여 우선 입력신호를 LPF를 거쳐 고주파 잡음을 제거하고 8KHz로 샘플링한 후 16bit양자화과정을 거친 뒤 실시간 끝점검출을 한다.

입력의 초기 5프레임(125msec)을 묵음구간으로 가정하여 에너지와 영교차율의 통계적 특징을 구간검출과 끝점검출의 기준값으로 사용한다. 이러한 실시간 끝점검출을 구하는 알고리즘은 그림 1과 같다.

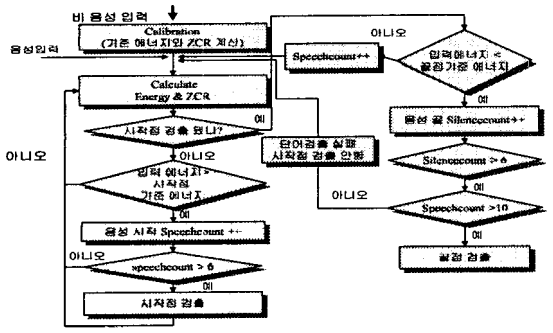

그림 1. 실시간 끝점검출알고리즘
Fig. 1. Real-time end-point detection algorithm.

2.2. 특징파라미터의 추출

인간의 발성기관을 모델링하여 필터계수를 구하고 중복되는 음성정보를 제거한다. 계수의 추출방법으로는 필터뱅크, BPF를 이용해 각 주파수대역의 에너지를 특징벡터로 사용하는 선형예측계수, 인간의 귀의 특성을 고려하여 주파수영역내의 가중합수를 구하는 멜-스케일 분석 등을 들 수 있다.

본 논문에서는 FFT를 이용하여 주파수 영역에서 멜-스케일로 이루어진 각 필터뱅크의 출력값들을 로그변환 한 후 역퓨리에 변환하는 MFCC(Mel-Frequency Cepstral Coefficient)를 사용하여, 12차 켑스트럼, 1차 정규화된 로그에너지, 12차 델타-켑스트럼, 1차 델타에너지를 특징파라미터를 구하고, 26차 리프터링을 통해 가중치를 부여하여 HMM 파라미터의 초기화입력으로 사용한다. 그림 2는 특징파라미터 추출과정을 나타낸 것이다[3].

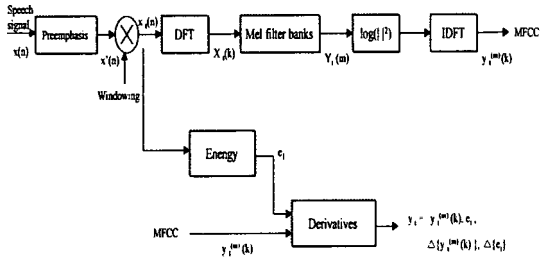

그림 2. MFCC 추출과정의 흐름도
Fig. 2. Flow chart of MFCC processing.

2.3. 필터뱅크의 처리

음성신호의 주파수영역에서 직접 필터뱅크를 사용하여 멜-켑스트럼계수를 추출하기위해 그림 3과 같은 멜-스케일을 사용한다. 이때 응답과 중심주파수는 DFT로 직접 필터링되고, 이 응답은 삼각형 모양으로 워핑되어 식 (1)과 같다.

$$U_{\triangle m}(k)\begin{cases} |k|<\triangle_m \rightarrow 1-|k|/\triangle_m \\ |k|\geq\triangle_m \rightarrow 0 \end{cases} \quad (1)$$

여기에서 k는 DFT영역의 인덱스이고, m번째 필터의 대역폭은 $2\triangle_m$이다.

이때, m번째 필터의 출력은 식 (2)와 같다.

$$Y_i(m)=\sum_{k=b_m}^{b_m+\triangle_m} X_i(k)U_{\triangle_m}(k+b_m) \quad (2)$$

여기에서 중심주파수 $b_m=b_{m-1}+\triangle_m$이고, 중심주파수가 1KHz미만이면 $\triangle_m=1.0$으로 1KHz이상이면 $\triangle_m=1.2\times\triangle_{m-1}$로 정의한다.

이러한 값들은 청각시스템에서 인간의 특정주파수를 감지해 내기 위해 사용한다[4].

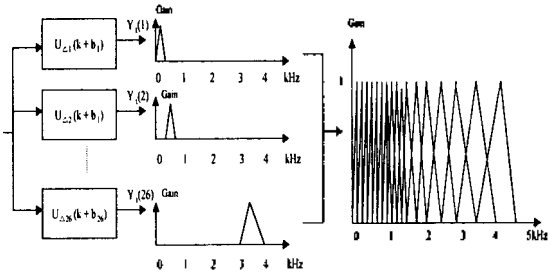

그림 3. 멜스케일에 따른 필터뱅크
Fig. 3. Filter-bank scaled according to Mel scale.

III. 연속분포 HMM

인간의 음성인식 매카니즘과 기계의 HMM 인식 시스템을 표 1과 같이 비교하였다.

표 1. 인간과 HMM 인식시스템의 비교
Table 1. Comparison of human between HMM recognition system.

항목	인간	HMM인식시스템
모 델	커다란 신경망	확률밀도함수를 갖는 한정된 HMM
파라미터	신경세포	특징파라미터
초기화	사전지식이 없음	초기화값은 제로
인식시작	훈련으로 언어습득	인간과 동일
훈 련	소리의 의미를 언어와 관련지음	소리의 의미를 단어의 스트링으로 표현
훈련의 효과	학습에 의해 신경망의 연결이 강화됨	학습에의해 파라미터의 통계적 추정
음성분석	귀는 주파수밴드로 처리되어 소리 인지.	멜스케일에 의한필터뱅크 구성
인식과정	HMM인식시스템과 유사함	관측확률의 최대값을 인식된 단어로 함

HMM의 파라미터 초기화를 통해 얻어진 값들은 음성신호에서 추출한 특징 자체를 나타내며 연속분포HMM(Continous HMM: CHMM)의 관측열 분포인 $B=\{b_j(O)\ |1\le j\le N\}$이다. 연속분포 HMM의 천이확률과 출력확률을 추정하기 위해 $b_j(O)$는 여러개의 가우시안 밀도에 가중치를 주고 합친 가우시안 혼합밀도함수(Gaussian Mixture Model : GMM)를 사용한다.

GMM함수 $b_j(O)$는 식 (3)과 같이 표현 될 수 있다.

$$b_j(O)=\sum_{K=1}^{M} C_{jk}\, N(O,\ \mu_{jk}, U_{jk}) \qquad (3)$$

여기에서 $N(O,\mu,U)$는 평균벡터 μ와 분산행렬 U를 갖는 특정벡터 P차수의 정규분포함수를 나타내고 가지(mixture) 가중치 c_{jk}는 식 (4)를 만족해야한다.

$$\sum_{K=1}^{M} C_{jk}=1, \quad 1\le j< N$$
$$C_{jk}\ge 0, \quad 1\le j< N,\ 1\le K< M \qquad (4)$$

CHMM에 GMM함수를 적용한 모델이 음성인식에서 많이 사용되는 이유는 첫째, 가우시안 분포와 음성신호의 확률분포사이의 오차를 최소화할 수 있고 둘째, 많은 수의 독립적인 확률 변수들을 합친 분포는 가우시안 분포로 근사화 될 수 있으며 셋째, 가우시안 분포는 주어진 분산에서 다른 분포보다 더 큰 엔트로피를 갖기 때문이다[5].

GMM으로 관측확률분포를 산출한 뒤 전·후향 알고리즘을 사용하여 각 단어의 확률적 모델링을 통해 모델파라미터를 재추정하고 재추정된 파라미터를 가지고 수렴할 때까지 확률값을 다시 구한다.

주어진 모델에 대해 관측열의 확률을 최대화하기 위해 모델 파라미터를 조정하는 방법으로서 해석적인 방법은 아직 알려져 있지 않고 Baum-Welch 방법과 같이 반복에 의해 국부적으로 최소화하는 방법을 사용한다.

IV. 자바를 이용한 웹상에서의 음성인식 시스템 구현

4.1. 자바의 구성요소

자바는 운영체제, 플랫폼API, 응용프로그램 등으로 구분된다.

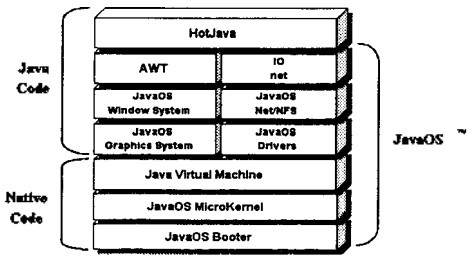

그림 4. 자바운영체제의 계층구조
Fig. 4. Hierarchy of Java OS.

1) 운영체제

Java O/S는 계층적 구조로써 플랫폼에 종속되어 그림 4와 같이 native code와 java code로 구성된다.

2) 플랫폼 API

플랫폼 API는 응용프로그램과 애플릿을 개발하는데 필요한 표준 인터페이스 집합으로서 JavaSoft 및 제 3의 업체들(Third party)에 의해 제공되며, 대부분 Java Base API와 Java Standard Extensive API의 2가지로 구성된다. 전자를 자바애플릿 API라고 하며 애플릿과 응용프로그램을 실행시키는 API 집합이며, 후자는 Java Base API를 확장시킨 것으로 그림 5와 같다.

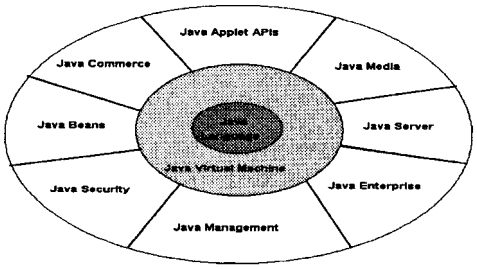

그림 5. 자바플랫폼 API모델
Fig. 5. API model of Java platform.

3) 응용프로그램

자바플랫폼에는 애플릿과 응용프로그램 형태의 두 프로그램이 있으며, 전자는 실행시에 브라우저가 필요하고, 후자는 로컬에서 호출되어 실행된다. 애플릿은 안정성과 보안을 위해 세가지의 제약점을 갖고 있다.

첫째, 클라이언트 시스템의 파일에 대해 read/write를 금지하며 둘째, 클라이언트 시스템의 프로그램을 동작시키거나 새로운 프로세스를 생성시킬수 없고 셋째, 접속한 서버를 제외한 다른 호스트와 새로운 연결을 만들 수 없다[6].

4.2. 자바 미디어 API

미디어 API는 오디오, 비디오, 3D 그래픽, 애니메이션, 공동작업, 전화통신, 음성 등을 포함한 다양한 범위의 대화형 매체를 처리하도록 멀티미디어 클래스를 제공하고 통신 API를 지원하여 완전성과 이식성을 높인다. JDK1.2에서 제공된 플랫폼은 그림 6과 같다

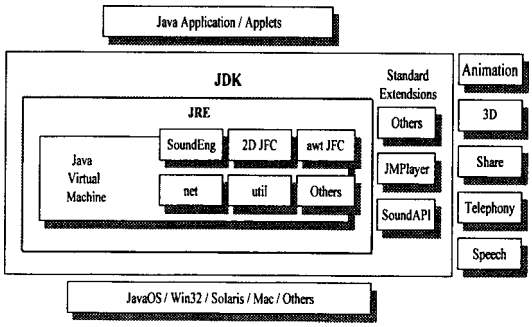

그림 6. 자바미디어의 플랫폼
Fig. 6. Platform of Java media.

4.3. 시스템의 구조

시스템의 구조는 그림 7과 같이 C/S의Browser-embedded 모델을 적용하였으며, 음성인식 서버를 액세스하는 웹브라우저에서 자바애플릿으로 음성을 record/play하고 그 결과를 클라이언트에 나타내도록 설계하였다.

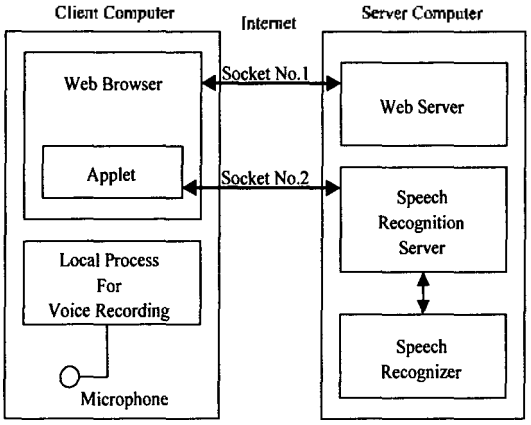

그림 7. Browser-embedded 모델의 구조
Fig. 7. Structure of Browser-embedded model.

클라이언트는 웹브라우저를 통해 서버로 접속하여 서버의 애플릿을 전송받아 GUI을 생성하고 음성인식 서버와 데이터 송/수신을 위해 소켓2를 연결한다. 이때 서버는 웹을 액세스할 수 있는 웹서버, 음성녹음과 전처리을 위한 애플릿, 소켓2의 인터페이스인 음성인식서버, CHMM으로 설계된 음성인식기로 구성된다.

4.4. 자바애플릿

1) record와 play의 설계

애플릿은 최종 사용자의 하드웨어 및 소프트웨어를 사용할 수 없는 제한점이 있기 때문에 제3의업체인 Scrawl 사의 SoundBite 클래스를 사용하여, 그림 8과 같이 사운드 카드를 제어하였다[7].

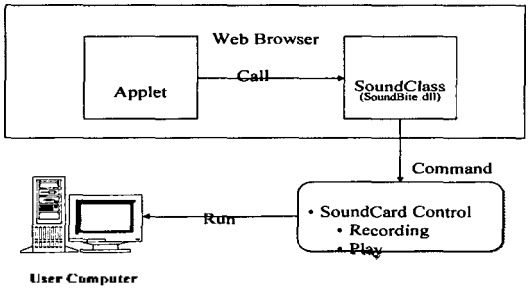

그림 8. 음성의 record/play 구성도
Fig. 8. Block diagram of the speech record/play.

2) 소켓의 생성과 종료

소켓은 컴퓨팅 시스템에서 두 응용 요소간에 통신을 지원하는데 이용되며 스트림 소켓, 데이터그램 소켓, RAW 소켓으로 구분된다. 그 중 스트림 소켓을 사용하여 그림 9와 같이 그 과정을 나타내었다.

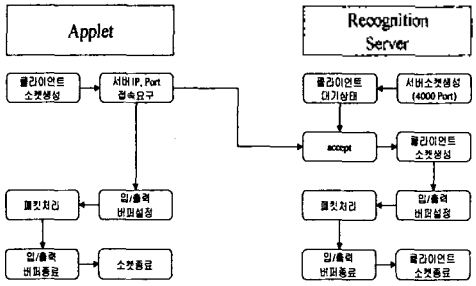

그림 9. 소켓의 생성과 종료과정
Fig. 9. The create and exit processing of stream socket.

4.5. 음성인식기

음성인식기는 애플릿에서 전처리된 특징파라미터와 학습된 DB와의 최대 확률을 추정하여 인식된 결과를 애플릿에 문자로 전송하며 CHMM을 이용하여 단어인식을 수행한다.

훈련DB는 전처리과정에서 추출된 26개의 특징파라미터를 HMM의 초기화 입력으로 사용해서 K-Means 알고리즘에 의해 8개의 상태로 등분할된다. 등분할된 각 상태는 3개의 가지로 나뉘어 평균과 분산을 구함과 동시에 각 상태의 가중치를 얻는다.

또한, HMM의 토폴로지는 left-to-right 모델을 사용하였으며, HMM 초기화 출력값들은 가우시안 혼합밀도함수, 전·후향 알고리즘, Baum-welch 재추정 알고리즘에 의해 문턱값이 5×10^{-3}이 될 때까지 수렴하여 파라미터를 갱신한다. 이 데이터가 훈련DB에 저장되어 사용된다.

4.6. JNI(Java Native Interface)의 설계

자바로 표현 불가능한 부분을 해결하기 위해 만들어진 JNI는 다른 언어로 작성되어진 라이브러리나 응용프로그램을 JVM(Java Virtual Machine)에서 자바 코드와 같이 실행될 수 있도록 설계된 것이다. 본 논문에서는 인식서버와 인식기 사이의 연결에 사용되었으며 그 구성도는 그림 10과 같다.

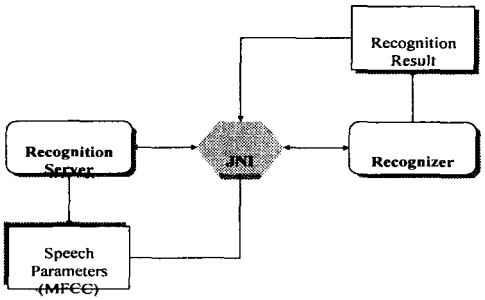

그림 10. 인식서버와 인식기의 JNI 구성도
Fig. 10. JNI of recognition server between recognizer.

4.7. 웹상에서 구현된 사용자인터페이스

웹상에서 구현된 사용자 인터페이스는 그림 11과 같이 애플릿에 2D Graphics Class를 사용하여 구현하였으며 첫째, Recording버튼이 클릭되면 음성을 3초동안 녹음하여 애플릿 하단에 출력되도록 하였고 둘째, 인식된 결과는 애플릿에 문자열로 출력되고 출력된 단어결과의 버튼의 색을 변화시켜 사용자가 쉽게 확인 할 수 있도록 하였다[8].

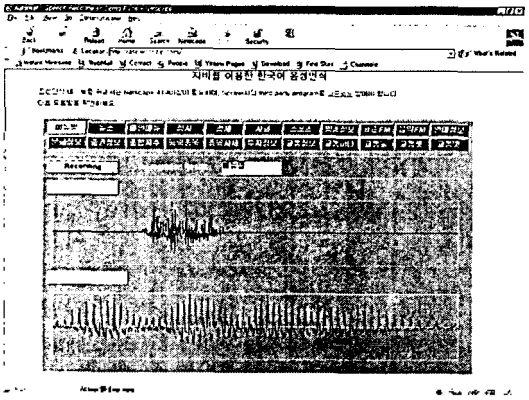

그림 11. 음성인식시스템의 GUI
Fig. 11. GUI of the speech recognition system.

V. 실험환경 및 결과

5.1 실험환경

실험을 위해서는 1대의 서버컴퓨터 시스템과 최소 1대 이상의 클라이언트 컴퓨터가 필요하고, 인터넷에 연결되어 있어야 한다. 본 논문에서는 10Mbps LAN, 서버는 Window NT Server 4.0을 탑재하였고, 클라이언트는 2대를 사용하였다. 표 2와 3은 H/W와 S/W의 실험실환경을 나타낸것이다.

표 2. 실험에 사용된 H/W
Table 2. Environment of H/W.

컴퓨터 / 종류	서버	클라이언트
Processor	450MHz	150MHz
Main Memory	72MB	32MB
Network	10Mbps LAN	10Mbps LAN
Sound	-	16Bit Support
Microphone	-	Dynamic 60Ω

표 3. 실험에 사용된 S/W
Table 3. Environment of S/W.

컴퓨터 / 종류	서버	클라이언트
Operation S/W	NT Server 4.0	Window 98
Web Server	IIS 3.0	-
JDK	1.3Beta Version	1.3Beta Version
C Compiler	Visual Studio 97	-
Audio Class	SoundBite 1.0	SoundBite 1.0
Web Browser	-	Netscape 4.6

5.2. DB

인식기의 학습용 DB는 자동차 항법시스템에서 사용되는 22개 단어를 52명의 20대 남성화자가 조용한 실험실 환경에서 발성한 것을 수집하였다. 수집된 단어리스트는 그림 11의 상단의 목록과 같다.

5.3. 인식결과

인식실험은 훈련에 참가하지 않은 화자 5명이 22단어를 2회 발성한 220개의 시험데이터로 사용하여 9.1%의 오인식율이 나타났다.

VI. 결 론

본 논문에서는 자바를 이용하여 한국어 음성인식 시스템을 인터넷상에서 구현하였다. 대부분의 음성인식 시스템은 시스템 의존적인 자원을 사용한다는 문제점을 갖고 있다. 이를 해결하기위해 자바를 이용하여 각 컴퓨터 O/S 플랫폼에 의존하지 않도록 인식서버와 클라이언트 프로그램을 구분하여 인터넷상에서 실행하여 그 타당성을 입증하였다.

참 고 문 헌

1. M. Sokolov, "Speaker Verification on the world wide web," Proceeding Eurospeech. pp 847-850, 1997.
2. V. Digalakis, L. Neumeyer, Perakakis, "Quantization of Cepstral Parameter for Speech Recognition Over the World Wide Web," Proceeding ICASSP, pp. 989-992, 1998.
3. C. Becchetti, Prina Ricotti, *Speech Recognition*, John Wiley & Sons, 1999.
4. L. R. Rabiner, B. H. Juang, *Fundamentals Of Speech Recognition*, Prentice-Hall International, Inc., 1993.
5. 홍형진, "한국어 음성인식의 성능향상을 위한 이산, 연속 및 준연속 HMM의 모델링 개선에 관한 연구," 한국과학 기술원 석사학위논문, 1994.
6. ZhemlnTu, philips C. Loizou, "Speech Recognition Over the Internet Using JAVA," Proceeding ICASSP, pp. 2267-2370, 1999.
7. Scrawl, Inc. " http://www.scrawl.com/store/".
8. 최광국, 이재왕, 김 철, 최승호, "웹상에서의 HMM을 이용한 한국어 음성인식," 한국음향학회 학술발표대회 논문집, 제18권, 2호, pp. 77-80, 1999.

▲ 최 광 국(Kwang Kook Choi) 1977년 2월 25일생

1999년 2월 : 동신대학교 정보통신공학과(공학사)
1999년 3월~현재 : 동신대학교 정보통신공학과 석사과정재학중
※ 주관심분야 : 음성인식, 통신소프트웨어, 자바플랫폼, 멀티미디어서비스

▲ 김 철(Cheol Kim) 1946년 9월 29일생

1970년 2월 : 조선대학교 전기공학과 (공학사)
1982년 2월 : 한양대학교 산업대학원 (공학석사)
1998년 2월 ~ 현재 : 동신대학교 정보통신공학과 박사과정 재학 중
※ 주관심분야 : 음성인식 및 신호처리, 통신망운용/관리

▲ 최 승 호(Seung Ho Choi) 1955년 8월 24일생

1981년 2월 : 전북대학교 물리학과 (이학사)
1984년 8월 : 명지대학교 전자공학과 (공학석사)
1992년 2월 : 명지대학교 전자공학과 (공학박사)
1992년 3월 ~ 현재 : 동신대학교 정보통신공학과 교수
※ 주관심분야 : 음성인식, 멀티미디어통신, 멀티모달 MMI

▲ 김 진 영(Jin Young Kim)
한국음향학회지 제18권 4호 참조