자동 교환 시스템을 위한 실시간 음성 인식 구현

An Implementation of the Real Time Speech Recognition for the Automatic Switching System

박 익 현*, 이 재 성*, 김 현 아*, 함 정 표*, 유 승 균*, 강 태 익*, 박 성 현*
(Ik-Hyun Park*, Jae-Sung Lee*, Hyun-A Kim*, Jung-Pyo Ham*, Seung-Kyun Ryu*, Tae-Ik Kang*, Sung-Hyun Park*)

요 약

본 논문에서는 음성 인식을 이용한 자동 교환 시스템을 구현하고, 성능을 평가하였다. 이 시스템은 다수의 구성원과 조직 체계를 가지는 관공서나 일반 기업, 학교 등의 교환 서비스를 음성 인식을 통하여 자동으로 제공한다. 본 시스템에 사용된 음성 인식기는 SCHMM(Semi-Continuous Hidden Markov Model) 기반으로 한 전화망에서의 화자 독립 고립 단어 가변 어휘 인식기(Speaker-Independent, Isolated-Word, Flexible-Vocabulary Recognizer)이며, 실시간 구현을 위해 사용한 DSP(Digital Signal Processor)는 Texas Instrument 사의 TMS320C32이다. 자동 교환 서비스를 위하여 음성 인식 기능 외에도 음성 인식 DSP 진단 기능과 인식 대상 어휘의 추가 및 변경을 위한 운용 단말을 구현하여 운용의 편의성을 추구하였다. 본 시스템의 인식 실험은 음성 인식 구내 자동 교환 시스템용 1300여 어휘(부서명, 인명 등)에 대해서 8명의 화자가 유선 전화망에서 수행하였으며 인식률은 91.5%이다.

핵심용어: 음성인식, 자동교환시스템, SCHMM, DSP
투고분야: 음성처리 분야(2,5)

ABSTRACT

This paper describes the implementation and the evaluation of the speech recognition automatic exchange system. The system provides government or public offices, companies, educational institutions that are composed of large number of members and parts with exchange service using speech recognition technology. The recognizer of the system is a Speaker-Independent, Isolated-word, Flexible-Vocabulary recognizer based on SCHMM(Semi-Continuous Hidden Markov Model).

For real-time implementation, DSP TMS320C32 made in Texas Instrument Inc. is used. The system operating terminal including the diagnosis of speech recognition DSP and the alternation of speech recognition candidates makes operation easy.

In this experiment, 8 speakers pronounced words of 1,300 vocabulary related to automatic exchange system over wire telephone network and the recognition system achieved 91.5% of word accuracy.

Key words: Speech recognition, Automatic eychanges system, SCHMM, DSP.

I. 서 론

전화는 일상 생활에서 없어서는 안 될 필수품이고, 이동 전화의 보급으로 전화 사용은 점차 증가하고 있다. 이에 따라 전화를 좀 더 편하게 사용하고자 하는 요구 역시 팽창하고 있는 실정이다. 이러한 측면에서 본 논문에서는 과거 오랜 기간 동안 많은 발전을 해온 음성 인식을 교환기에 적용하여 자동 교환 시스템을 구현하고 평가하였다.

본 논문에서 구현한 음성 인식 시스템은 화자가 전화를 걸어 교환기를 거치면 음성 인식 자동 교환 시스템에서 음성 인식을 하여 시스템에 등록된 전화 번호로 자동 교환을 수행한다. 본 시스템은 관공서, 기업, 혹은 임의의 교환 서비스가 필요한 곳에서 사설 교환기를 통하여 교환원이 문답을 통하여 교환을 하거나 자동 안내 서비스에 접속한 후 서비스 시나리오에 따른 교환 번호를 추가로 사용자가 눌러서 교환되는 경우를 음성 인식을 이용하여 자동으로 구현한다. 서비스를 위해서 구현한 음성 인식 시스템은 화자 독립 고립 단어 인식기로서, 임의의 사용자 모두가 사용할 수 있으며 사용자가 기억하기 용이한

* LG정보통신 디지털 네트워크연구소, 미디어 기기실
접수일자 : 2000년 1월 31일

사람 이름 혹은 부서명을 등록하여 이를 말하면 인식할 수 있도록 하였다.

전화망에서 동작하는 자동 교환 시스템을 위해서는 음성 인식 이외에 운용에 필요한 기능들이 있다. 인식 대상 어휘가 학습 시에 고정되는 것이 아니라 변경이 가능한 가변 어휘를 구현하기 위해서는 인식 단위를 음소 단위로 하여 인식 대상 어휘 목록에 따라 음소의 재배열로 인식 대상을 구성하는 기능이 필요하다. 본 논문에서 구현된 음성 인식 시스템은 mono-phone과 tri-phone을 사용하여 가변 어휘 구현 및 성능 향상을 꾀하였다.

서론에 이어 2장에서는 음성 인식 자동 교환 시스템을 전반적으로 소개하고, 3장에서는 실시간 음성 인식 구현에 대한 연구 내용을 서술하며, 4장에서는 현재 하남 시청에서 시험 서비스 중인 음성 인식 자동 교환 시스템을 기반으로 한 성능 평가 및 결과를 소개하고자 한다. 그리고 마지막으로 5장에서는 결론을 맺는다.

II. 음성 인식 자동 교환 시스템

2.1. 음성 인식 자동 교환 시스템의 구성

전화망 부가 서비스 시스템인 VIPS-HM(Value Added Information Processing System Hybrid Medium)은 공중 전화망(PSTN), 이동 전화망(AMPS/DCN/PCS) 또는 사설 교환(PBX)망과 연결되어 가입자에게 실시간으로 원하는 음성이나 FAX 정보를 제공하며 운용 단말인 CMS(Central Management System)을 통해 관리된다. 운용 단말은 VIPS-HM의 동작 상태를 확인하고 제어할 뿐만 아니라 각종 통계 정보 및 장애를 관리한다. 또한 음성 인식 자동 교환 서비스와 관련하여 인식 대상 어휘의 리스트, 전화번호 및 기타 데이터베이스를 관리하고 추가, 변경, 삭제에 따른 데이터베이스의 변경 시에 이에 관련된 데이터 파일을 생성하여 VIPS-HM으로 전송하여 주는 역할을 담당한다.

그림 1은 시스템 망 구성도 이며, 그림 2는 시스템 구성도이다.

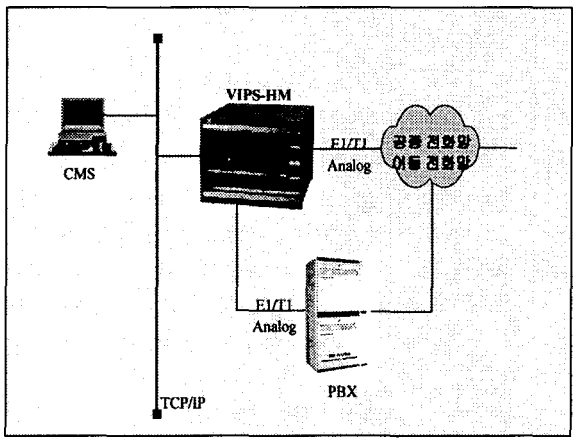

그림 1. 시스템 망 구성도
Fig. 1. Schematic diagram of the system network.

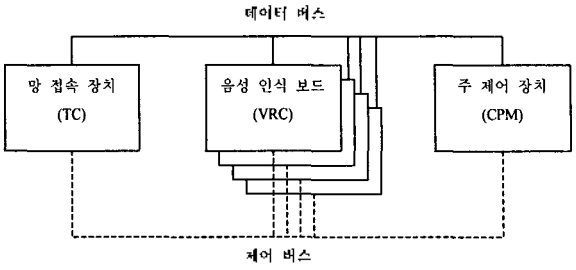

그림 2. 시스템 구성도
Fig. 2. Schematic diagram of the system.

주 제어 모듈(CPM : Centralized Processing Module)은 VIPS-HM시스템 전체를 제어하는 장치로서 망 접속 장치(TC : Trunk Circuit)를 통하여 접속된 통화로의 연결, 유지, 복구 등 호 처리 기능을 제공하며 접속된 호에 대한 통계를 수집한다. 또한 시스템 전체의 동작 상태를 감시하고, 장애 발생시 운용 단말 및 원격 경보 장치를 통하여 가시 및 가청 경보를 알려주며 서비스 시나리오를 수행하는 모듈이다.

망 접속 모듈(TC : Trunk Circuit)은 공중 전화 망 및 사설 교환기와의 Digital/Analog Trunk 접속 및 국간 신호 처리 기능을 수행하는 모듈이다.

음성 인식 보드(VRC : Voice Recognition Circuit)는 음성인식 서비스를 담당하는 가입자 보드로서 보드 당 4 채널 또는 8채널의 음성인식, 음성 Play, DTMF/Tone 검출, 디지털 음량조절, ADPCM Encoder/Decoder기능을 제공하며 VIPS-HM에는 4 모듈까지 장착된다. 음성 인식 프로그램은 음성 인식 보드의 DSP에 탑재되며 주 제어 모듈(CPM)로부터 음성 인식 시작 명령을 받아 음성 인식 결과를 보고하는 기능을 수행한다

망 접속 장치와 음성 인식 보드, 그리고 주 제어 장치는 음성 및 데이터의 경로인 데이터 버스와 상호 제어 및 상태 관련 경로인 제어 버스와 연결되어 있다.

2.2. 제어 버스와 데이터 버스

VRC와 CPM 그리고 TC와의 메시지는 제어 버스를 통하여 구현된다. TC는 호 관련 메시지를 CPM과 교환하며, CPM과 VRC는 음성 인식 관련 메시지를 교환한다. 음성 인식 모듈에 있어서, CPM으로부터 VRC로 보내는 메시지는 인식 시작, 인식 멈춤, 진단 명령 등 이 있으며 VRC가 CPM으로 보내는 메시지는 시작점, 끝점, 인식 결과, 진단 결과 보고 등이 있다. 정의된 메시지는 하드웨어 타이머 인터럽트를 이용한 메시지 폴링 방식을 사용하여 교환한다.

음성 인식 자동 교환 서비스 진행을 위한 안내 음성을 CPM이 VRC의 안내 음성 메모리의 안내 음성을 전송을 받거나, 안내 음성 메모리에 안내 음성을 전송 하는 동작과 VRC의 프로그램과 기타 인식 관련 데이터를 CPM이 전송하는 기능은 데이터 버스를 통하여 구현된다.

III. 실시간 음성 인식 구현

3.1. 학습 데이터 베이스 구축

자동 교환 시스템을 위한 가변 어휘 화자 독립 음성 인식기 구현용 학습 음성 데이터 베이스 구축을 위하여 수집 대상 어휘의 단순한 발성의 나열이 아닌 인식 서비스를 실시간으로 수행하여 음성 데이터 베이스 구축 운용 시스템을 구현하였다. 음성 발성자가 실제 음성 인식 시스템 환경에서 수집 대상 어휘를 발성하므로 실제 서비스 상의 입력 음성과 학습용 음성 데이터 베이스의 차이를 극복하였다.

3.2. 특징 추출

전화망을 통해서 들어온 음성을 8KHz로 표본하여 pre-emphasis(1-0.953z^{-1})한 후 10ms 간격으로 20ms의 해밍창을 적용한다. 특징 벡터는 자기상관 계수(autocorrelation) 방법을 사용하여 12차 LPC 켑스트럼 계수를 구하고 24차의 차분 켑트스럼(40ms 간격 12차 + 80ms 간격 12차), 12차의 이차 차분 켑스트럼, 그리고 에너지, 일차 차분 에너지, 이차 차분 에너지로 구성된다.학습 환경과 인식 수행 환경에서의 채널 잡음을 제거하기 위하여 cepstral mean subtraction을 사용하였다. 채널 잡음은 캡스트럼 특징 벡터에 평균 이동의 영향을 끼치므로 입력 특징 벡터의 평균을 제거하는 cepstral mean subtraction을 사용하여 전화망에서 강인한 음성 인식 시스템을 구현하였다.

상기 4가지 특징 벡터를 통해 얻어진 코드북은 각 128개의 코드 워드를 갖는다.

3.3. 인식 알고리즘

음성 인식은 음성을 확률 모델로 가정하고 입력된 음성을 학습된 모델 중에서 발생할 확률이 가장 큰 모델을 선택하는 과정으로 정의할 수 있다. 현재 사용되고 있는 음성 모델링 방법으로는 HMM(Hidden Markov Model)이 가장 효과적으로 알려져 있다.

HMM은 음성을 두 계층으로 이루어진 확률 과정으로 모델링 한다. 상태(state)들은 하위 계층의 확률 과정으로서 한 음성 프레임의 관찰 확률을 나타내고, 각 상태들은 마르코프 확률 과정으로 천이할 수 있도록 상위 계층이 구성되어 있다. HMM은 상태의 관찰 확률을 산출하는 방법에 따라, 이산 HMM, 반연속 HMM, 연속 HMM으로 나누어 진다. 본 논문에서는 코드북과 가우시안 확률 분포의 가중합을 함께 사용하는 반연속 HMM을 사용하여 구현하였다.

인식 대상이 가변 독립 단어이므로 인식 대상 집합이 변할 수 있다. 따라서, 음성을 음소 단위로 모델링하여 인식 대상 집합을 새로 구성하였을 때 학습된 음소 모델을 조합하여 단어를 모델링한다. 음소 모델링의 성능을 높이기 위하여 전후 발성된 음소에 따라 다르게 모델링하는 tri-phone과 독립된 음소를 모델링하는 mono-phone을 이용하였다.

인식 과정에서는 인식 시간을 단축하기 위하여 비터비 빔 탐색과 목구조 탐색을 사용하였다. 비터비 빔 탐색은 비터비 탐색에서 최고 확률을 가지는 상태의 확률값보다 적당한 임계값만큼 작은 확률을 가지는 상태를 탐색 과정에서 제외시키는 방법이다. 목구조 탐색은 음소 단위로 학습된 인식 대상 단어들은 중복되어 계산되는 부분이 크다는 특성을 사용한다. 예를 들어, 학교와 학습은 앞의 학에 해당하는 음소를 공통으로 사용하므로 따로 계산할 필요가 없다. 인식 대상을 정렬시키면 이러한 특성을 최대한 사용할 수 있어 인식 시간 개선에 크게 도움이 된다.

인식 모델을 학습하는 방법은 수집된 데이터 베이스 중 일부를 사람이 직접 표시하여 초기화 모델을 만든 후, mono-phone 학습과 tri-phone 학습을 순서대로 수행하였다. 학습은 최대 우도 방법인 Baum-Welch 재추정 알고리듬을 이용하였다. 학습 후 인식 모델의 신뢰성 및 속도 개선을 위해 상태 병합(state tying)을 수행해서 모두 1000개의 상태를 구성하고 다시 학습하였다.

3.4. 실 시간 구현

음성 인식 모듈의 실시간 구현을 위하여 25MIPS-(million instruction per second)의 성능을 갖는 Texas Instrument 사의 TMS320C32-50 DSP를 사용하여 구현하였으며, 메모리 사용의 최적화와 DSP 프로그램의 압축, 그리고 인식 알고리즘의 개선을 통하여 인식 속도 향상을 추구하였다.

그림 3은 음성 인식 모듈의 동작 구성도이며, 표1은 구내 자동 교환 서비스 용 1311개 어휘에 대한 메모리 요구량이다. 화자 독립 가변어휘 인식기는 8Mbyte의 메모리에서 약 3000단어의 인식이 가능하다. 대용량의 어휘를 인식하는 가변 어휘 인식기의 경우 대부분의 메모리는 인식 모델이 차지하며 인식 모델의 크기는 인식 대상 어휘에 따라 유동적이다.

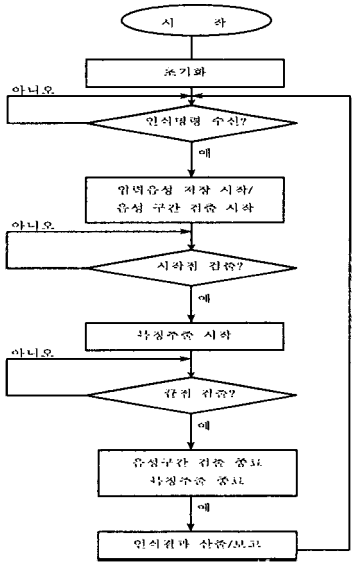

그림 3. 음성 인식 모듈의 순서도
Fig. 3. Flowchart of the speech recognition module.

표 1. 음성 인식 프로그램 메모리 요구량
Table 1. Memory requirement for the speech recognition program.

구분	메모리 크기(Kbyte)
소스코드	30
상수 및 초기값	4
광역 변수	227
Heap 및 Stack	720
Codebook	200
인식 모델	3312
계	4493

3.5. 자체 진단 기능

음성 인식 자동 교환 시스템 구현에서 시스템의 신뢰성은 매우 중요하다. 시스템의 신뢰성 향상을 위한 조치로 VIPS-HM의 주 제어 장치는 Active/Stand-By방식을 이용한 이중화로 구성되며, 그 외 각 부분들은 주기적인 진단을 통해 장애 발생시 장애를 복구하도록 되어 있다.

음성 인식 모듈의 경우 주기적으로 운용 단말로부터 내려오는 진단 명령을 받아 자체적인 진단을 수행하여 오류가 있거나 진단 결과를 보고할 수 없을 정도의 장애가 있을 시 장애의 복구를 수행한다. 음성 인식 모듈의 자체 진단 방법으로는 소스 코드와 초기화 영역인 상수, 코드북 및 인식 모델 등 인식 데이터 메모리 영역과 제어 버스의 메시지 전달 기능에 대한 오류 검사를 수행한다.

인식 모듈 메모리 영역의 오류 검사 수행을 위해 인식 프로그램 및 인식 데이터와 관련된 로딩 파일은 생성시 자체 파일 구조 정보를 가지고 있도록 하였다.

3.6. 인식 대상 어휘 변경 기능

음성 인식 구내 자동 교환 시스템은 특정 부서나 사람의 전화 번호를 외우지 않고서도 부서명이나 인명을 발성함으로써 해당 전화 번호로 연결하므로 사용자는 편리하지만 부서나 인원의 추가, 삭제, 변경에 따른 인식 대상 어휘의 변경 발생시 인식 대상 어휘의 변경이 뒤따라야 한다는 측면에서 운용자는 인식 모듈 자체를 교체하거나 인식 모델을 재생성하는 어려움이 있다. 그러나, 이러한 작업을 운용단말 GUI(graphic User Interface)를 통해 이루어지게 함으로써 부서나 인원의 추가, 삭제, 변경 사항만 입력하면 자동으로 인식 모델을 생성할 뿐만 아니라 운용자의 실수로 인하여 생길 수 있는 오류를 최소화 하도록 하였다. 즉, VIPS-HM의 운용단말에서 부서나 인원의 리스트 및 전화번호 등 필요한 정보를 데이터 베이스로 구축하여 관리하고 변동 사항이 생길 시 운용자는 GUI를 통해 편집하고 적용 버튼을 누르면 자동으로 새로운 인식 모델이 생성된 후, 시스템으로 전송되어 각 인식 모듈의 해당 메모리의 위치로 전송되어 인식 대상 어휘의 변경이 실시간 적용된다.

그림 4는 인식 대상 어휘가 변경되어 전송되는 과정을 나타내며, 그림 5는 CMS의 실제 화면의 예이다.

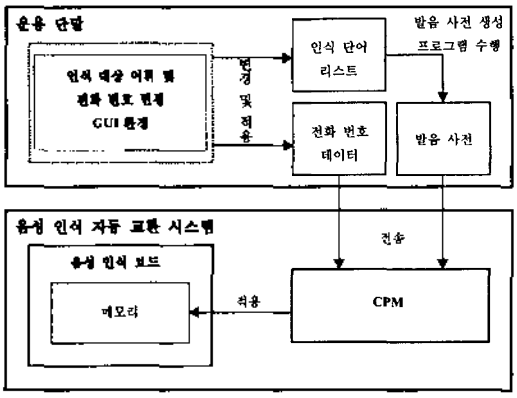

그림 4. 음성 인식 대상 어휘 변경작업 과정
Fig. 4. Operation of the change of the speech recognition candidate word.

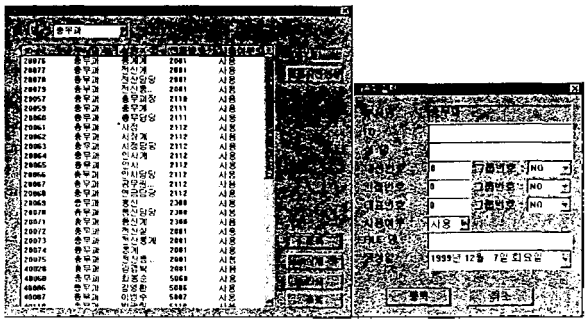

그림 5. CMS의 실제 화면의 예
Fig. 5. Example of the CMS screen.

3.7. 음성 수집 기능

음성 인식기 구현에서 시스템에 입력된 신호를 분석하거나 학습 데이터 베이스로 사용하기 위해서 음성 데이터의 지속적인 수집은 필수적이다. 시스템의 동작 모드에서 음성 수집 모드를 설정하면 끝점 보고된 음성 혹은 호 접속이 일어난 시점부터의 음성이 수집된다. 음성 수집은 데이터 버스를 통하여 시스템의 하드 디스크로 수집 되거나, LAN을 통하여 특정 컴퓨터에 저장된다.

IV. 성능 평가 및 결과

현재 하남 시청에서 시험 서비스 중인 음성 인식 자동 교환 시스템을 기반으로 아래의 성능 평가 및 결과를 산출하였다.

4.1. 인식 대상 어휘

자동 교환 시스템 용 인식 대상 어휘는 부서명과 인명 및 직책명으로 이루어져 있으며, 부서명은 다시 사용자의 편의성을 고려하여 정식 명칭 외에도 통상적으로 사용되는 약식 명칭과 기능에 따른 민원 명칭으로 구성된다. 표 2는 인식 대상 어휘들의 분류표이다.

표 2. 인식 대상 어휘들의 분류
Table 2. Classification of the speech recognition candidate words.

구 분	예
인명	손영채, 박우량, 김창배, 조세환, 금광형,...
직책명	시장, 부시장, 총무, 국장, 개발 국장, 세무 과장,...
부서명(정식명칭)	기획 감사실, 문화 공보실,...
부서명(약식명칭)	기감실, 공보실,...
부서명(민원명칭)	등초본 발급, 여권 신청, 출생 신고, 가스 점검,...

4.2. 인식 실험

인식 실험은 사무실 환경에서 8명의 화자가 구내 자동 교환 서비스용 1311단어에 대해 유선 전화기를 이용, 직접 전화를 걸어 성공여부를 기록하는 방식의 on-line 시험을 수행하였으며 인식결과는 표3과 같다.

화자들의 인식 응답 시간 측정을 통해 평균1~2초의 인식 시간이 소요되었다.

표 3. On-line 인식 실험 결과
Table 3. Recognition result of on-line test.

	시도 횟수	성공 횟수	인식률(%)
화자 1	483	365	83.3
화자 2	219	202	92.2
화자 3	438	421	96.1
화자 4	441	429	97.3
화자 5	215	190	88.4
화자 6	551	523	94.9
화자 7	551	487	88.4
화자 8	551	496	90.0
계	3404	3113	91.5

V. 결 론

본 논문에서는 현재 운용되고 있는 음성 인식 구내 자동 교환 시스템을 전반적으로 살펴보았다.

하나의 시스템이 원활히 역할을 수행하기 위해서는 여러 부분의 노력이 필요하다. 음성 인식 서비스이므로 인식률과 인식 시간이 중요한 요소이지만 그 외에도 사용자 습관을 고려한 약식 명칭, 민원 명칭의 추가 및 등록, 인식 대상 어휘 변경에 따른 용이한 적용, 진단 기능을 이용한 신뢰성 있는 서비스 제공 등 사용자 측면, 운용자 측면, 시스템 신뢰성 측면에서 많은 고려를 통해 구현하였다.

향후 다양한 환경과 화자 특성에 강인한 인식기의 구현, 거절 기능 및 핵심어 인식을 이용한 서비스 이용자의 편이성 수용에 관한 연구와 인식기의 성능과 적용 분야를 고려한 서비스 시나리오의 연구와 개선은 지속적으로 진행되어야 할 것이라 고려되며, 음성 인식 서비스의 대중화 및 효율성 향상을 위하여 고속의 DSP를 기반으로 실시간 다 채널 음성 인식기의 구현이 필요하다.

참 고 문 헌

1. 유하진, 김훈, 최재승, 이종석, "가변어휘 인식모델을 이용한 한국어 방송뉴스 음성의 인식," KSCSP98 15권1호, pp.70-73.
2. Jae-Seung Choi, Jong-Seok Lee, Hee-Youn Lee, "Smoothing and Tying for Korean Flexible Vocabulary Isolated Word Recognition," ICSLP98.
3. M.Y. Hwang, X. Huang, "Shared-Distribution Hidden Markov Models for Speech Recognition," IEEE Trans. Speech and Audio Processing, vol. 1, no. 4, 1993, pp.414-420.
4. "VIPS-HM 설치 및 사용 설명서," LG정보통신㈜, 1999.1.

▲ 박 익 현(Ik Hyun Park)

1997년 2월 : 포항공과대학교 전자계산학과(공학사)
1997년~현재 : LG정보통신 연구원
※ 주관심분야 : DSP, 음성인식

▲ 이 재 성(Jae-Sung Lee)

1992년 2월 : 부산대학교 전자공학과(공학사)
1994년 2월 : 부산대학교 전자공학과(공학석사)
1994년~현재 : LG정보통신 주임연구원
※ 주관심분야 : DSP 응용, 음성인식, 음성신호처리

▲ 김 현 아(Hyun-Ah Kim)

1997년 2월 : 서강대학교 전자계산학과(공학사)
1999년 2월 : 서강대학교 전자계산학과 (공학석사)
1999년 7월~현재 : LG정보통신 연구원
※ 주관심분야 : 음성인식

▲ 함 정 표(Jeong-Pyo Ham)

1997년 2월 : 연세대학교 전자공학과 (공학사)
1999년 2월 : 연세대학교 전자공학과 (공학석사)
1999년 2월 ~ 현재 : LG정보통신 연구원
※ 주관심분야 : DSP, 음성인식

▲ 유 승 균(Seung-Kyun Ryu)

1989년 2월 : 연세대학교 전자공학과 (공학사)
1991년 2월 : 연세대학교 전자공학과 (공학석사)
1991년 2월 ~ 현재 : LG정보통신 선임 연구원
1997년 9월 ~ 현재 : 연세대학교 전자 공학과(박사과정)
※ 주관심분야 : DSP 응용

▲ 강 태 익(Tae-Ik Kang)

1979년 2월 : 연세대학교 전자공학과 (공학사)
1981년 2월 : 연세대학교 전자공학과 (공학석사)
1990년 1월 : Polytechnic University (Ph.D)
1981년 3월 ~ 1984년 6월 : 육군사관 학교 교수요원
1990년 5월 ~ 현재 : LG정보통신 책임연구원
※ 주관심분야 : DSP, 영상신호처리, VOIP, ATM

▲ 박 성 현(Sung-Hyun Park)

1978년 2월 : 한양대학교 전자공학과 (공학사)
1990년 2월 : 한국과학기술원 전기전 자과(공학석사)
1978년 6월 ~ 1987년 6월 : 금성통신 선임연구원
1987년 7월 ~ 현재 : LG정보통신 정보 네트워크연구소 연구위원
※ 주관심분야 : 음성신호처리, VOIP, NGN