음성 인식을 이용한 증권 정보 검색 시스템의 개발
(Development of a Stock Information Retrieval System using Speech Recognition)

박 성 준 † 구 명 완 †† 전 주 식 †††
(Sung-Joon Park)(Myoung-Wan Koo)(Chu-Shik Jhon)

요 약 본 논문에서는 음성 인식을 이용한 증권 정보 검색 시스템의 개발에 대하여 기술하고 시스템의 주요 특징을 설명한다. 이 시스템은 DHMM (discrete hidden Markov model)에 기반을 두고, 유사 음소를 기본 인식 단위로 사용하였다. 끝점 검출과 반향 제거 기능을 포함시켜 사용자의 음성 입력이 편리하도록 만들었으며, 한 번의 음성 입력이 하나만의 단어가 아닌 여러 개의 단어가 될 수 있도록 연속 음성 인식기를 구현하였다. 상용화 이후의 몇 개월에 걸친 데이터를 이용하여 운용 결과를 분석하였다.

Abstract In this paper, the development of a stock information retrieval system using speech recognition and its features are described. The system is based on DHMM (discrete hidden Markov model) and PLUs (phonelike units) are used as the basic unit for recognition. End-point detection and echo cancellation are included to facilitate speech input. Continuous speech recognizer is implemented to allow multi-word speech. Data collected over several months are analyzed.

1. 서 론

전화망을 통해서 서비스를 제공하는 시스템에 있어서 가장 자연스럽고 편리한 방법은 음성을 이용하는 것이다. 이러한 시스템을 구성하는 핵심적인 부분은 음성 인식기로서 자동 음성 인식에 대한 많은 연구가 진행되어 왔다.

그러나 전화망을 통한 음성 인식은 마이크 입력을 통한 음성 인식에 비해 어려운 편이다. 채널의 왜곡과 배경 잡음으로 인해 음성 인식률이 떨어지며, 일반인의 사용을 위해서 화자 독립적으로 구현되어야 하기 때문이다.

이러한 이유로 전화 음성 인식기를 대어휘 인식에 적용하여 실용적인 시스템을 개발하기는 어려운 편이다. 그러나 작은 크기나 중간 크기의 어휘를 가지는 시스템에는 효과적으로 사용될 수 있으며, 실제로 지난 몇 년간 쓸모 있는 시스템들이 개발되어 왔다. AT&T에서는 신용 카드 정보를 접근할 수 있는 연속 숫자음 인식기를 개발하였는데, 이 시스템의 경우 11개의 단어만을 가지고 있다[1]. IBM에서 개발한 GALAXY 시스템에서는 여러 도시들의 날씨와 같은 정보를 알려준다[2]. 이와 같은 예들이 보여주는 것은 몇 천 단어 정도를 인식하는 음성 인식기를 내장한 시스템들의 경우, 음성 인식 기능이 특정 분야에 효과적으로 적용될 수 있다는 점이다.

한국통신에서도 지난 몇 년간에 걸쳐 전화망을 통한 음성 인식 시스템을 개발하고, 성능 개선을 위한 작업을 수행해 왔는데, 1991년에 시작된 음성 언어에 대한 연구 결과의 하나로서 1994년도에 음성인식 증권 정보 시스템인 KT-STOCK을 개발하였으며[3,4,5], 시험 운용과 시범 서비스를 거치면서 기능 및 성능상의 문제점을 보완하였다[6,7,8,9]. 이러한 과정을 거쳐 1999년 3월에는 상용 시스템이 나오게 되었으며, 지속적으로 성능 향상을 위한 연구를 하고 있는 중이다. 1994년에 처음 개발된 KT-STOCK은 PC상에서 구현된 프로토타입 시스템으로서 DSP (Digital Signal Processor) 보드를 사용하였으며, 인식 대상은 712개의 주식 명칭이었다[3,4]. 이후 시스템에 잡음 모델을 추가하여 잘못된 입력에 대

† 정 회 원 : 한국통신 멀티미디어연구소 연구원
sjpak@kt.co.kr
†† 비 회 원 : 한국통신 멀티미디어연구소 연구원
mwkoo@kt.co.kr
††† 종신회원 : 서울대학교 컴퓨터공학부 교수
csjhon@riact.snu.ac.kr
논문접수 : 2000년 1월 11일
심사완료 : 2000년 5월 9일

해서는 잡음으로 처리할 수 있도록 하였다[5]. 그리고 프로토타입 시스템을 이용한 시험 서비스를 통해 실제 환경에서 다양한 사용자들의 음성을 수집하였으며, 이를 끝점 검출기의 성능 평가 및 개선에 이용하였다[6]. 프로토타입 시스템을 통해 어느 정도의 성능을 파악한 다음, 대용량 시스템으로의 확장 및 상용화 작업에 착수하였으며, 이 과정에서 시스템의 개발비용 절감을 위한 시스템의 구조 개선 연구가 있었다[7,8]. 초기에 개발된 대용량 시스템은 음성의 특징을 추출하는 프로세서들과 인식을 위한 비터비 검색을 수행하는 프로세서들이 분리되어 특징 추출을 하는 동안에 비터비 검색을 수행할 수 있도록 설계되었으며, 이 두 프로세서 개수들간의 비율을 조절함으로써 성능을 감소시키지 않고서도 전체 프로세서의 개수를 줄이는 작업을 하였다. 하지만 현재 시스템에서는 특징 추출과 비터비 검색을 모두 하나의 프로세서에서 처리하는 방식을 취했는데, 그 이유는 프로세서 하나를 사용하는 경우가 개발의 편의성, 안정성, 유지보수 측면에서 유리했기 때문이었다. 그러나 특징 추출과 비터비 검색을 동시에 수행하는 경우에 비하여 시스템의 응답 속도가 떨어지는 단점이 있으며, 이를 보완하기 위하여 작업이 분리된 프로세서를 사용하는 방식을 다시 고려하고 있는 중이다.

시스템의 인식 단어 개수는 계속 증가하여 현재는 약 1500 여 개로서 이는 현재 상장되어 있는 주식 종목의 수와 서비스 제공에 필요한 몇 단어를 합한 것이다. 사용자가 전화가 걸어 연결이 되면 시스템은 안내 메시지를 내보낸다. 사용법을 잘 모를 경우에는 '사용법 안내'를 말하면 되고, 이미 사용법에 익숙할 경우에는 메시지가 나오는 중에 알고자 하는 주식 종목을 말하면 된다. 그러면 현재 나오는 메시지는 중단이 되고 인식 결과에 따른 정보가 나오게 된다. 현재 서비스가 제공되는 지역은 서울을 포함하여 총 7 곳이며, 서울에 설치된 시스템에는 240 회선이 할당되어 있다.

시스템이 실용적인 것이 되기 위해서는 몇 가지 갖추어야 할 요건이 있는데, 일단 신뢰성이 높고 사용자가 불편함을 느끼지 않을 정도의 음성 인식률을 보장해야 한다는 점이다. 그리고 증권 정보의 특성상 새로운 단어를 추가하거나 사용되지 않는 단어의 삭제가 용이해야 하며, 화자 독립적인 음성 인식 기능을 제공해야 하는데, 본 시스템의 개발은 이러한 점들을 고려하여 이루어졌다. 우선 인식률을 높이기 위한 작업으로서 시뮬레이션을 통해 인식 프로그램을 지속적으로 수정하고 성능을 향상시켜 왔다. 그리고 수정된 프로그램들은 프로토타입 시스템에 적용하여 시험 서비스를 통해 검증하였으며, 시험 결과를 분석하여 인식 프로그램, 서비스 시나리오 및 시스템의 하드웨어 사양에 반영시켰다.

본 논문에서는 음성인식 증권정보 검색시스템에 적용된 음성 인식기에 대하여 기술하고 시스템 개발과 관련된 작업들을 정리하며, 구성은 다음과 같다. 2장에서 음성 인식기에 관한 논의를 하고 3장에서는 시스템의 구현에 대하여 설명한다. 4장에서는 시스템의 초기화 및 관리에 필요한 운용 시나리오를, 5장에서는 서비스 기간 동안의 운용 결과를 분석하고 마지막으로 6장에서 결론을 맺는다.

2. 음성 인식기

본 시스템에 구현된 음성 인식기는 DHMM (discrete hidden Markov model)을 사용하였으며, 연속 음성 인식을 지원한다. 이 장에서는 음성 인식기에 구현된 여러 가지 기법들에 대하여 살펴본다.

2.1 끝점 검출

음성 인식을 위해서는 전화선을 통해 연속적으로 들어오는 신호의 어느 부분이 실제 음성인지를 먼저 파악해야 한다. 이를 위해서 끝점 검출기가 필요하며, 끝점 검출기가 음성의 시작점과 끝점을 적절히 찾아냄으로써 인식률을 높일 수 있다. 끝점 검출 방법은 두 가지로 나눌 수 있는데, 검출 대상 단어의 음성 신호가 다 들어온 후에 구간을 찾기 시작하는 방법과 연속적으로 들어오는 입력에 대하여 특정 시점 이전까지 입력된 데이타만을 이용하여 음성 구간을 찾는 실시간적인 방법이 있다. 첫째 방법을 사용하기 위해서는 시스템이 사용자가 말할 수 있는 시간을 주고, 그 주어진 시간 동안에 들어온 입력에 대해서만 처리를 한다. 이 방법은 미리 정한 일정 시간 내에 발성해야 하므로 주어진 시간이 너무 짧으면 사용자가 말한 음성의 일부만을 받아들이게 된다. 그리고 주어진 시간이 너무 길면 사용자가 말한 뒤에 불필요한 시간을 기다려야 하고, 또한 남은 시간 동안 불필요한 잡음이 포함될 가능성도 있다. 따라서 사용자와의 자연스러운 인터페이스를 위해서는 두 번째 방법이 사용되어야 한다.

실시간 처리를 위한 끝점 검출은 음성 입력과 동시에 시작점 검출 작업을 수행한다. 본 시스템에서는 끝점 검출을 위해 에너지, 영교차율(*ZCR, zero crossing rate*)을 사용하였다. 초기 과정에서 처음 입력된 일정 시간의 데이타를 묵음 구간의 잡음으로 가정하여 여기서 얻은 에너지와 영교차율을 기준값으로 정하게 되는데, 실시간적인 동작의 성격상 묵음 구간을 미리 예상할 수는 없기 때문에, 묵음 구간이 일정 시간 계속되면 최근 묵음

구간의 데이타를 새로운 기준값으로 정하여 변화된 잡음 환경에 적응하여 끝점을 검출할 수 있게 한다. 끝점 검출의 전체적인 알고리즘을 그림 1에 나타내었다[6].

끝점 검출기가 시작할 때의 상태는 START로 설정되어 있으며, 따라서 처음 수행하는 작업은 기준값의 초기화이다. 이 작업이 끝나면 상태는 NOISE로 바뀐다. NOISE 상태에서부터는 음성의 시작점을 찾는 일을 하는데, 음성의 시작점은 에너지가 Thresh_EL보다 클 때인 경우를 기준으로 하여 음성의 일부 앞부분에서 찾는다. 그 다음에는 SP_E 상태로 바뀌며, 끝점 검출 단계로 넘어간다. 끝점 검출은 에너지가 Thresh_EH보다 작은지를 검사하는 데서부터 시작하여 묵음 구간이 설정된 시간을 넘는 데까지 이루어진다. Thresh_EL과 Thresh_EH는 테스트를 통해 조정된 값을 사용하였다.

2.2 특징 추출

음성의 시작과 끝이 정해졌으면 이제는 음성을 프로그램이 다룰 수 있는 형태로 변환하는 작업이 필요하다. 이를 위해서 음성을 디지털 형태로 바꾸고 스펙트럼(spectrum) 분석을 통해 음성의 특징을 나타내는 파라미터를 추출한다. 본 시스템에서는 LPC 켑스트럼 계수(linear predictive coding cepstral coefficient)를 사용하였다[10] .

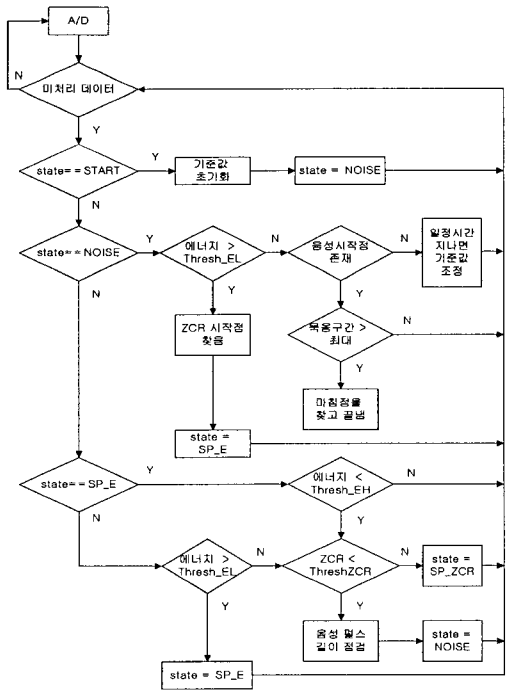


그림 1 끝점 검출 알고리즘

현재 전화망에서 사용하는 음성의 대역폭은 8kHz이므로 본 시스템에서도 동일한 주파수를 사용하여 음성을 샘플링(sampling)한다. 샘플링된 음성은 필터(filter)를 통해 프리엠퍼시스(pre-emphasis) 과정을 거친다. 이후, 음성은 20msec 길이의 프레임(frame)으로 분할되는데, 각 프레임들은 앞뒤 프레임과 10 msec 씩 중첩된다. 각 프레임에 대해 LPC (linear predictive coding) 분석을 하여 14차의 LPC 계수를 구하고 이를 다시 변환하여 12차의 켑스트럼 계수를 얻는다. 사용되는 음성 특징 데이타에는 켑스트럼 계수를 비롯하여 이 계수의 1차, 2차 차이값 및 로그 파워(loged power)의 1차, 2차 차이값 등도 포함된다. 이 값들은 종류별로 벡터 양자화를 거쳐 코드북 인덱스(codebook index)로 나오는데, 사용되는 코드북은 모두 4개이다. 로그 파워의 경우에는 1,2 차 차이값을 함께 표현한 64*2의 크기를 가지는 배열로 되어 있으며, 나머지는 256*12 크기의 배열이다. 코드북을 만드는 데 사용한 알고리즘은 LBG (Linde-Buzo-Gray) 방식에 기반을 두었다[11].

2.3 반향 제거

전화선로에서는 단방향 4선을 양방향 2선 신호로 변환해 주는 하이브리드(hybrid) 회로와 전화선간의 임피던스 불일치에 의해 원단(far end) 신호의 반향 성분이 중첩되어 근단(near end) 신호의 음질을 저하시키게 된다. 대화 방식으로 음성의 입출력이 양방향으로 동시에 진행되는 음성 인식 서비스 시스템에서 반향 신호에 의해 음질이 나빠지는 현상을 줄이고 음질 저하로 인한 인식률 저하를 막기 위한 방법으로서, 출력 음성, 즉 안내 메시지의 반향 성분을 자동적으로 제거하기 위한 적응 디지털 필터(adaptive digital filter)를 이용한 것이 있다[12]. 본 시스템에서는 NLMS (normalized least mean square) 알고리즘의 빠른 수렴 특성을 살리면서 더블톡(double talk) 구간에서의 왜곡 현상을 줄이는 새로운 반향 제거기가 적용되었는데, 개선된 적응 필터 알고리즘은 시스템 입력 신호(사용자 음성 + 반향 신호)와 출력 신호(안내 방송)가 더블톡 구간에서는 서로 상관성이 떨어지는 점을 이용하여 시스템 입/출력 신호의 상호 상관계수를 구하여 NLMS 적응 필터의 스텝 크기에 곱해 줌으로써 더블톡 구간에서의 왜곡 현상을 막도록 하였다[9].

2.4 음소 모델

음성의 모델링에 사용되는 기본 단위는 여러 가지가 있다. 본 시스템에서 사용한 단위는 유사 음소(phonelike units)로서 각각의 음소는 동일한 위상(topology)을 가지며, SPHINX 시스템에서 사용된 것과

같다[13]. 그림 2에 음소 모델의 위상을 나타내었다.

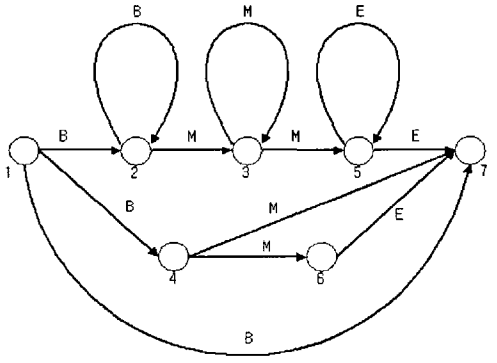


그림 2 음소 위상 모델

이 모델은 7개의 상태(state)와 12개의 전이(transition)로 이루어지는데, 각각의 전이는 세 그룹 B, M, E 중의 하나에 속해 있다. 같은 그룹에 속한 전이는 동일한 출력 확률을 가진다. 이 모델에서는 각 음소에 대하여 최대 세 개의 안정한 상태가 있는 것으로 가정한다. 음소 모델의 개수는 문맥 독립(context-independent) 유사 음소의 경우 약 60 개이며, 이를 확장한 문맥 종속(context- dependent) 유사 음소의 수는 약 500 개 정도이다. 문맥 종속 유사 음소는 인식 단위 축소 법칙(unit reduction rule)을 적용하여 생성하였다[14].

2.5 언어 모델

현재 상장된 주식 중에는 부가주를 가진 종목들도 있다. 따라서 예를 들어 만약 사용자가 어떤 종목 '갑'의 제1우선주에 대해서 알고 싶다면 '갑 제1우선주'라고 말해도 인식할 수 있도록 해 주는 것이 필요하다. 이 경우에는 인식기가 독립 단어 인식이 아닌 연속 음성 인식을 수행해야 한다. 즉, '갑'이라는 단어와 '제1우선주'라는 단어로 구성된 하나의 문장을 인식해야 하는 것이다. 시스템이 연속 음성을 인식하기 위해서는 이것에 맞는 언어 모델이 필요한데, 한 예를 그림 3에 나타내었다[15]. 그림에서 'wd0_*'로 표시된 것은 부가주가 없는 단어들, 'wd1_*'과 'wd2_*'는 부가주가 있는 단어들, 'share1_*'과 'share2_*'는 부가주를 나타내는 단어들이다. 'noise'는 잡음을 나타내며, 'q1', 'q2', 'q31' 'q32'는 묵음을 나타낸다. 이 그림에 나타낸 것은 하나의 간단한 예로서 부가주로 사용되는 단어가 세 가지인 경우만 보여주지만, 실제로 시스템에 적용된 것은 모든 부가주에 대해서 인식할 수 있는 모델이다. 그리고 묵음의 명칭이 다른 이유는, 연속 음성 인식기가 바이그램(bigram)을 사용하기 때문에 묵음이 인식과정에서 혼돈을 일으키지 않도록 하기 위한 것이다. 비록 명칭은 다르지만 이 시스템에서는 동일한 내용의 모델을 사용하였다. 예를 들어, 그림 3에 의하면 'wd1_1 share2_1' 형태의 문장을 만들 수 없으나, 동일한 이름의 묵음을 사용할 경우 'share2_1'을 인식하는 단계에서 바이그램 정보로서 확률 p(묵음 | share2_1)를 사용하기 때문에 묵음의 앞의 단어가 'wd1_*'인지 'wd2_*'인지를 알지 못한다. 따라서 바이그램 정보만을 사용할 경우, 옳지 않은 형태인 'wd1_* share2_*'를 인식 결과로 내놓을 수가 있으나, 이 문제는 묵음의 명칭을 서로 다르게 줌으로써 피할 수 있는 것이다.

그림 3에 나타난 바와 같이 인식 결과가 잡음이 될 수도 있는데, 이것은 시스템의 인식 대상 단어 리스트에 잡음을 추가하였기 때문이며, 잡음으로 인식할 경우에는 사용자에게 다시 한번 말씀해 달라는 안내 음성을 내보내게 되어 있다. 하지만 잘못된 입력이 들어왔을 때 잡음으로 처리하는 경우가 많지는 않으며, 이는 시스템의 개선 사항 중의 하나이다.

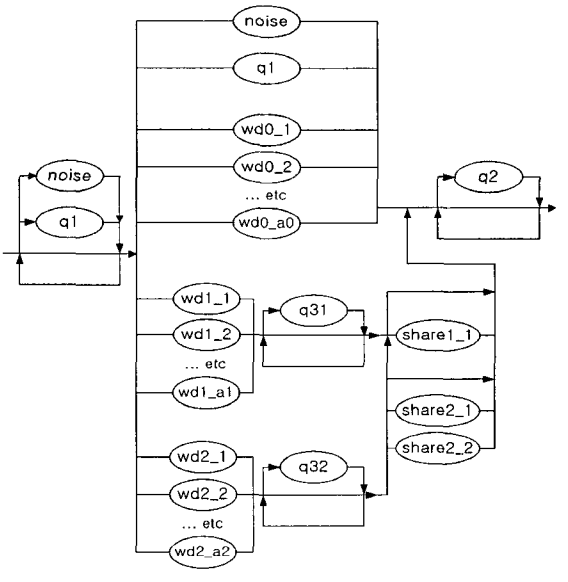


그림 3 언어 모델

2.6 훈련 및 인식 실험

본 시스템은 일반인이 사용하기 때문에 HMM 모델의 훈련에 필요한 음성 데이타도 남녀별과 연령별에 따라 일정 비율로 수집하였다. 그리고 시스템이 전화를 통하여 입력된 음성을 인식하기 때문에 음성 데이타 구축에도 전화를 이용하였다. 모델을 훈련하는 데에는

74,971개의 음성 데이타가 이용되었는데, 우선 독립 음소를 훈련한 다음, 이를 확장하여 종속 음소를 생성시키고 다시 훈련하였다. Baum-Welch 알고리즘을 사용하였으며[10], 독립 음소와 종속 음소에 대하여 각각 5회씩 반복 훈련시켰다.

인식 테스트는 두 종류로 나누어서 수행했는데, 첫 번째는 고립 단어만을 테스트하였고 두 번째는 부가주가 포함된 두 단어 문장을 테스트하였다. 고립 단어의 경우 5,467 개의 음성 데이타를 사용하였으며, 85.96%의 인식률을 얻었다. 두 번째 테스트에서는 870 개의 문장을 사용하였으며 87.13%의 인식률을 보여 주었다[15].

3. 시스템의 구현

시스템은 크게 두 부분으로 구성되어 있는데, 전단계 신호 처리 및 음성 인식 알고리즘이 구현된 음성 인식 모듈(Speech Recognition Module)과 인식된 결과에 따라 데이타를 수집하고 사용자에게 들려 줄 안내 메시지를 작성하는 관리 모듈(Management Module)로 나누어진다. 그림 4는 하드웨어의 논리적 구성을 나타낸 것이다.

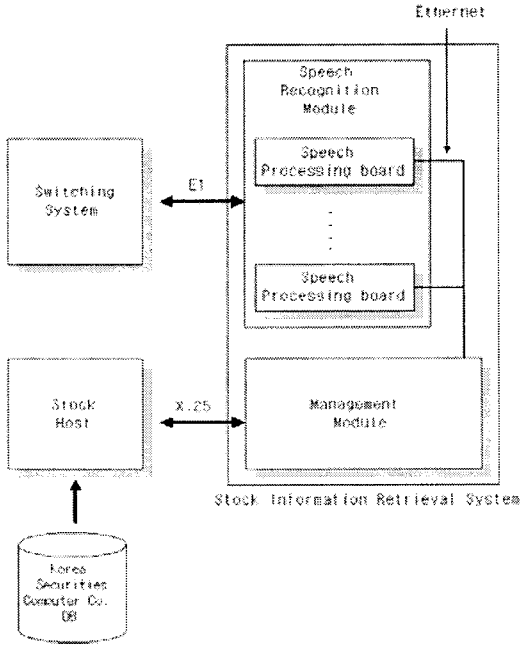


그림 4 시스템의 구성

3.1 음성 인식 모듈

음성 인식 모듈은 TMS320C32 DSP를 네 개씩 탑재한 보드들로 구성되어 있는데, 각각의 DSP가 음성 인식 프로그램을 수행하며, 각 보드당 4 회선을 담당한다. 시스템은 보드를 추가함으로써 확장 가능하도록 구현되었다. 입력은 음성뿐만 아니라 DTMF(dual tone multi-frequency) 신호도 처리가 가능하도록 각 보드당 별도의 DTMF 신호 처리 프로세서가 추가되어 있다.

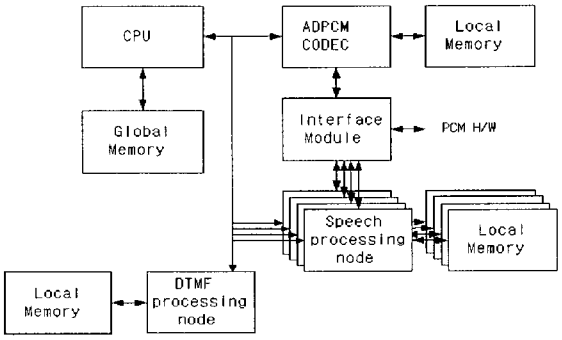


그림 5 음성 인식 모듈의 블록도

전화선을 통해 들어온 음성은 음성 인식 모듈에서의 처리 과정을 통해 특정 단어를 나타내는 인덱스로 결과가 나오고 이 결과는 관리 모듈로 전달된다. 또한 관리 모듈로부터 출력할 음성이 넘어 오면 이를 전화선을 통해 사용자에게 들려준다. 음성 인식 모듈을 그림 5에 나타내었다.

DSP에 할당된 메모리는 각각 8 MB(Megabyte)로서 인식 프로그램과 인식에 필요한 데이타, 즉 코드북, HMM 파라미터, 단어 사전, 클래스 정보 등이 초기화 과정에서 적재된다. CPU는 Intel 계열의 80C186이 사용되었으며, 보드의 초기화 및 관리 기능을 담당한다. CPU 레지스터 및 각종 외부 칩들의 초기화가 모두 끝나면, 무한 루프 내에서 입출력과 관련된 버퍼 제어, 타이머 태스크, DTMF 태스크, 인식 DSP 태스크 등의 작업을 수행하며, 위와 같은 과정을 반복하다가 인터럽트가 들어오면 인터럽트 처리 루틴을 실행한다.

인식 DSP 태스크에는 DSP와 관리 모듈간의 통신 기능, DSP를 체크하는 기능, DSP 프로그램 변수를 제어하는 기능, DSP 에러 발생시 프로그램을 다시 다운로드할 수 있도록 상위 프로세서에 요구하는 기능, 입력 음성의 시작점 검출시 출력 음성을 차단해 주는 기능 등이 있다. DSP와 관리 모듈간의 통신은 메시지를 통해 이루어지는데, 메시지의 종류에는 다음과 같은 것들이 있다.

관리 모듈에서 DSP 쪽으로 보내는 메시지는 인식 시작 명령, 인식 정지 명령, 반향 제거 중지 명령, DSP

자체 진단 명령, 다운로드된 데이타 검증 명령, DSP 변수값 조정 명령, DSP 변수값 확인 명령 등이며, DSP 쪽에서 관리 모듈쪽으로 보내는 메시지에는 인식 결과값 보고, 인식 실패 보고, 음성 시작점 검출 보고, 음성 끝점 검출 보고, DSP 자체 진단 보고, 다운로드된 데이타 검증 보고, DSP 변수값 보고 등이다.

메시지의 전송은 전역 메모리(Global Memory)에 메시지 내용을 쓰고 읽음으로써 이루어지는데, 그 방식은 다음과 같다.

DSP에서 관리 모듈로 메시지를 전송할 때는 서로간에 이미 정의된 상태 비트 ENLB와 ACK가 사용되는데, ENLB를 체크하여 기존에 보낸 메시지가 없는지(ENLB의 값이 0) 확인하여 그 값이 0이면, 쓰기 영역에 메시지를 쓰고 ENLB의 주소를 액세스한다. 그러면 하드웨어적으로 ENLB와 ACK가 1로 바뀐다. 관리 모듈에서는 ACK 비트를 폴링하면서 대기하던 중에 ACK가 1로 세팅되면 해당 주소를 읽는다. 그리고 나서 ACK의 주소를 액세스해주면 하드웨어적으로 ENLB와 ACK의 값이 0으로 바뀐다. 관리 모듈에서 DSP로 메시지를 보낼 때는 메시지의 전달 방향만 다를 뿐 동일한 방식으로 이루어진다.

3.2 관리 모듈

관리 모듈에서는 음성 인식 모듈로부터 넘어온 인식 코드에 해당되는 정보를 증권 호스트로부터 넘겨받아서 안내 메시지를 작성하고 이를 다시 음성 인식 모듈로 넘겨준다.

관리 모듈의 다른 기능은 시스템의 초기화 과정에서 DSP 프로그램과 음성 인식에 필요한 데이타를 다운로딩(downloading)하고 데이타가 변경되었을 때 수정된 데이타를 다시 다운로딩하는 일이다. 주식 시장에서는 주식의 상장에 의해 새로운 종목이 생겨난다. 따라서 시스템에서는 이러한 변화를 즉각적으로 반영하여 단어 리스트에 종목명을 추가하고 사용자에게 현재 정보를 제공하는 일이 필요하다. 마찬가지로 기존의 종목이 삭제되었을 때에는 없어진 종목에 대한 잘못된 정보를 주지 않도록 단어 리스트에서 그 단어를 제거해 주어야 한다.

인식 대상이 되는 단어들은 모두 단어 사전에 포함되어 있는데, 단어 사전은 각 단어의 코드 번호와 단어를 구성하는 모델들의 명칭으로 이루어져 있다. 하나의 종목 코드에 대하여 여러 개의 명칭을 가질 수도 있는데, 이러한 경우까지 고려하여 단어 사전을 구성하였다.

본 시스템에서는 단어의 추가, 삭제에 따라 단어 사전에 변화가 생겼을 때에는 음성 인식에 필요한 데이타를 자동적으로 갱신하여 준다. 만약 단어는 추가되었지만 해당되는 단어에 대한 음성이 준비되지 않았다면, 합성기가 미리 저장되어 있는 음성도막을 이용하여 해당되는 단어를 합성한다. 사전에 변화가 생겼을 때 파라미터를 다시 훈련하지는 않는다. 왜냐하면 서비스의 실시간적인 특성상 단어의 변동이 있을 때마다 파라미터를 새로 훈련하고 다운로딩하기는 곤란하기 때문이다. 다만 오프라인(off-line)에서 파라미터를 훈련하여 더 좋은 인식률을 보일 경우 이를 기존의 것과 대치할 수 있다. 이 작업은 관리 모듈의 디스크에 새로운 파라미터를 갖다 놓고 시스템을 초기화하면 된다.

4. 서비스 시나리오

본 시스템의 서비스는 두 가지 모드로 나누어져 있다. 첫째는 사용자의 음성을 인식할 수 있는 음성 인식 모드이고, 둘째는 전화기 버튼 입력만을 받아들이는 버튼 입력 모드이다. 두 모드간에는 서로 공통적으로 지원되는 서비스 내용이 있고 각각의 모드에서만 고유적으로 지원되는 서비스 내용도 있다. 음성인식 모드에서의 인식 단어는 종목명과 서비스명인데, 종목명의 경우에는 현재 상장되어 있는 주식과 부가주의 명칭들이다. 그리고 서비스명은 '종합주가 지수', '코스닥 지수', '시황정보', '사용방법 안내' 등이다. 종목명의 경우에는 말을 하지 않고 DTMF로 코드 번호를 눌러도 처리하도록 되어 있다. 버튼 입력 모드에서는 종합 주가 지수와 코스닥 지수, 그리고 몇몇 다른 서비스를 제공한다. 두 모드간의 전이는 '*'를 누르면 된다. 본 시스템의 경우에는 내용이 단순하기 때문에 시나리오가 복잡하지는 않다. 음성 입력 모드나 버튼 입력 모드에 상관없이 단지 들어오는 입력에 대하여 처리를 해서 출력을 내보내고, 다시 입력이 들어오면 그 입력에 대하여 처리를 하는 반복 작업을 되풀이한다. 일정 시간 동안 입력이 없을 때에는 입력이 없었음을 사용자에게 알리고 몇 번에 걸쳐 입력이 없으면 시스템 쪽에서 전화를 끊는 방식을 취한다. 사용자의 입력이 잘못되었을 때는 입력이 잘못되었음을 알리고 다시 안내 음성을 내보낸다.

5. 운용 결과

시스템이 운용되는 곳은 서울을 비롯한 대도시 7곳이지만, 본 논문에서는 데이타 수집의 편의상 서울 지역에서의 데이타를 이용하여 분석하였다.

그림 6과 그림 7에는 1999년 5월에서부터 10월에 이르기까지의 착신호수와 월별 통화 완료율을 나타내었는데, 90% 이상의 완료율을 보여 준다.

착신호수가 월에 따라 차이가 나는 이유는 주식 시장의 변화에 따른 사용자들의 관심 정도에 따라 전화를 거는 횟수가 변하기 때문인 것으로 여겨진다. 여기서 착신호수는 사용자가 전화를 건 회수를, 통화 완료율은 이 중에서 시스템과 연결된 호수의 비율을 의미한다. 사용자가 말한 단어에 대한 시스템의 응답 시간은 보통 3초 내외이며, 입력이 길거나 음질이 안 좋은 경우에는 4,5초까지 걸릴 때가 있다. 전화가 연결되고 나서 사용자가 끊을 때까지의 평균 사용 시간은 약 1분 30초 정도로서 거의 일정하다.

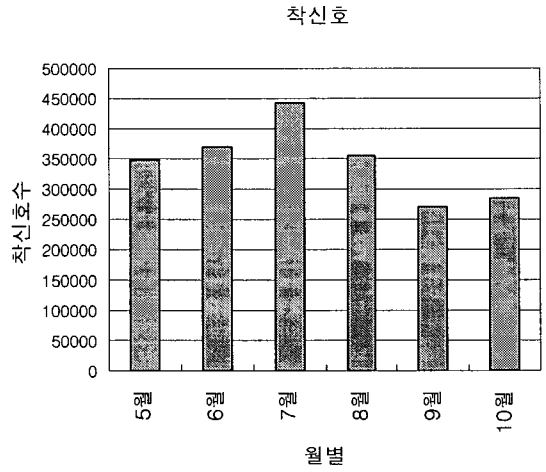


그림 6 착신호수

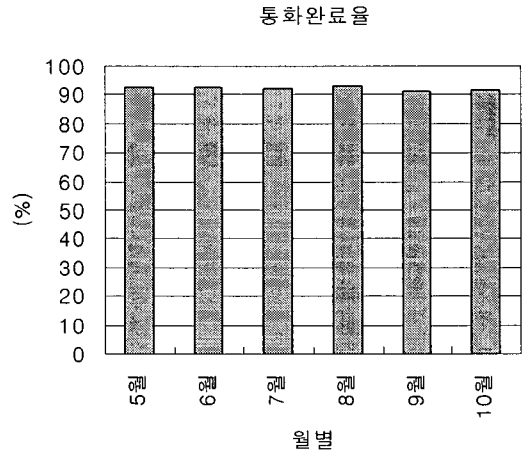


그림 7 통화 완료율

인식률은 사용자가 전화를 건 장소나 전화기의 종류 및 발화 습관에 따라 차이가 나지만, 평균적으로 80-90% 정도의 인식률을 보여 준다. 오인식을 하는 경우는 사용자가 인식 대상에 없는 단어를 말한다든지 잡음이 입력되는 때가 많았으며, 이는 시뮬레이션보다 인식률이 떨어지는 요인이 되었다. 이러한 경우를 제외한다면 시뮬레이션과 비슷한 인식률을 보여 준다. 단어 중에서 빈번하게 오인식되는 것들은, 많은 경우 2음절 단어로서 오류 유형을 보면 유사한 음절이나 음소가 있는 경우(방림 -> 상림, 기린 -> 세진, 한솔 -> 바른손, 협동 -> 극동, 교하 -> 유화 등)가 많으나, 전혀 엉뚱한 단어로 인식될 때도 많았다(세품 -> 북두, 거평 -> 시넥스, 홍창 -> 동서, 원풍 -> 동원 등). 일단 단어의 길이가 짧기 때문에 다른 단어와 구별할 수 있는 특징 데이타가 적다는 점을 오류의 주된 원인으로 꼽을 수 있으나, 보다 체계적인 파악을 위해서는 음성학적인 측면에서의 접근이 병행되어야 할 것으로 생각한다.

국내에서 개발된 다른 음성인식 증권정보 시스템으로서 같은 시기에 상용화된 것이 있다[16]. 시나리오는 조금 다르지만 유사한 방식으로 서비스를 제공하는 것으로 보이며, 인식률도 비슷한 결과를 보여 준다.

6. 결 론

본 논문에서는 한국통신에서 개발한 음성인식 증권 정보 검색 시스템의 특징 및 구현된 기술에 대하여 설명하고 시스템의 운용 결과를 살펴보았다. 시스템의 상용화를 위해서는 음성 인식과 관련된 기술, 시스템의 구현 기술, 서비스 및 관리 운용에 필요한 기술이 조화를 이루어야 한다. 음성인식 증권 정보 검색 시스템은 이러한 측면에서 적절한 응용 시스템 중의 하나로 볼 수 있으며, 어느 정도 만족할 만한 결과를 보여 주었다고 여겨진다. 앞으로는 이를 바탕으로 철도 예약 시스템을 비롯한 다른 종류의 서비스를 개발할 예정이며, 아울러 인식률을 향상시키는 작업도 수행해 나갈 것이다.

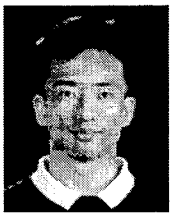
참 고 문 헌

[1] M. G. Rahim, and B. H. Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition," IEEE Trans. On Speech and Audio Processing, Vol. 4, pp. 19-30, Jan., 1996.

[2] V. Zue, S. Seneff, J. Polifroni, H. Meng, and J. Glass, "Multilingual Human-Computer Interactions: From Information Access to Language Learing," in Proc. Int. Conf. Spoken Language Processing, pp. 2207-2210, 1996.

[3] M. W. Koo et al., "An experimental field trial of a

large vocabulary, speaker independent recognition system," in Proceedings of the Second IEEE Workshop on Interactive Voice Technology for Telecommunication Applications, pp. 33-36, Sep., 1994.
[4] 도삼주, 김우성, 장두성, 구명완, "음성인식기술을 이용한 증권정보 안내시스팀의 실험적 실용 시험," 11회 음성통신 및 신호처리 워크샵 논문집1호, pp. 241-244, 1994.
[5] M. W. Koo et al., "A stock information system over the telephone network," in Proceedings of the 6th International Conference on Signal Processing Applications & Technology, pp. 2039-2043, 1995.
[6] 장경애, 김재인, 구명완, "실용적 음성 인식 시스템에서 끝점 검출기의 성능 평가," 한국음향학회 학술발표대회 논문집 제15권, 제 1(s)호, pp. 21-25, 1996.
[7] 박성준, 김재인, 전주식, "음성인식 증권정보 검색시스템의 개선방향에 관한 고찰," 한국정보과학회 '97 추계 학술발표논문집, pp. 639-642, Apr., 1997.
[8] Sung-Joon Park, Jae-In Kim, Chu-Shik Jhon, "The architecture of a speech recognition system for cost reduction," in Proceedings of International Conference on Speech Processing, pp. 549-553, Aug., 1997.
[9] 강명구, 유창동, "입/출력 신호의 상관계수를 이용한 반향제거기," 1998년도 한국음향학회 학술발표대회 논문집 제17권 제1(s)호, pp. 189-192, 1998.
[10] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition*, Prentice-Hall, NJ, 1993.
[11] Y. Linde, A. Buso, R. M. Gray, "An Algorithm for Vector Quantizer Design," IEEE Transactions on Communication COM-28(1), pp. 84-95, Jan., 1980.
[12] Hong Fan and W. Kenneth Jenkins, "An investigation of an adaptive IIR echo canceller : advantages and problems," IEEE Trans. On Acoust. Speech and Signal Processing, Vol. 36, No. 12, pp. 1819-1833, Dec., 1988.
[13] K.-F. Lee, *Automatic speech recognition: the development of the SPHINX system*, Kluwer Academic Publishers, Norwell, Mass., 1989.
[14] C. H. Lee et al., "Acoustic modeling of subword units for speech recognition," in Proc. 1990 IEEE Int. Conf. Acous., Speech, Signal Processing, pp. 721-724, Apr., 1990.
[15] Sung-Joon Park, Myoung-Wan Koo, Chu-Shik Jhon, "An implementation of continuous speech recognition for a stock information retrieval system," in Proceedings of International Conference on Speech Processing, pp. 461-464, Aug., 1999.
[16] 감지은, 이문형 "음성인식 증권정보서비스 구축 사례," 정보처리학회지 제6권 제4호, pp. 67-72, 1999.



박 성 준
1992년 2월 서울대학교 컴퓨터공학과 학사. 1994년 2월 서울대학교 컴퓨터공학과 석사. 1996년 2월 서울대학교 컴퓨터공학과 박사과정 수료. 1996년 2월 ~ 현재 한국통신 멀티미디어연구소 전임연구원.

구 명 완
1982년 2월 연세대학교 전자공학과 학사. 1985년 2월 한국과학기술원 전기 및 전자공학 석사. 1991년 8월 한국과학기술원 전기 및 전자공학 박사. 1996년 12월 ~ 1997년 12월 미국 벨연구소 객원연구원. 1985년 4월 ~ 현재 한국통신 멀티미디어연구소 음성언어연구팀장

전 주 식
정보과학회논문지 : 시스템 및 이론
제 27 권 제 2 호 참조