음성인식기술을 이용한 증권정보 안내시스템의 실험적 실용시험

o
도 삼주, 김 우성, 장 두성, 구 명완
한국통신 소프트웨어연구소 자동통역연구팀

An Experimental Field Trial of Stock Information Retrieval System Based on Speech Recognition

Sam-Joo Doh, Woo-Sung Kim, Du-Seong Chang, Myoung-Wan Koo
Automatic Interpreting Telephony Team, S/W Research Laboratories, Korea Telecom

< 요 약 >

이 논문은 대어휘, 화자독립 음성인식 시스템인 KT-STOCK과 이 시스템에 대한 전화망을 통한 실험적 실용시험에 대해 기술하였다. KT-STOCK은 현재 주식시장에 상장된 712개 회사의 현재주가를 음성을 이용하여 검색할 수 있는 시스템이다. 이 시스템은 hidden Markov model 기술에 기반을 둔 고립단어 인식 시스템이며 유사음소를 기본 인식단위로 사용한다. KT-STOCK은 1994년 6월 24일 부터 실험적 실용시험 중에 있다. 중간 결과에 따르면 모의 실험 결과는 실제 환경에서의 시험과 차이가 있는 것으로 나타났다. 실제 환경에서 이 시스템의 인식률은 현재 61.9%이다.

1. 서론

전화망을 통한 음성인식은 음성인식 기술의 주요 응용분야 중 하나로서, 최근 이 분야에 많은 진전이 이루어졌다. 일본 NTT사는 화자독립 음성인식 기술과 음성합성 기술을 결합하여 ANSER (Automatic Answer Network System for Electrical Request)라는 전화정보검색 시스템을 개발하였다. 이 시스템은 1981년 도입된 이래 은행업에 대한 정보 서비스를 제공해 왔다[1]. 1985년초 AT&T사는 현재 교환원에 의해 처리되는 호의 일부를 자동화하기 위해 제한된 단어의 화자독립 음성인식 기술을 이용할 수 있는지를 조사했다. 1992년 실용시험후 AT&T사는 음성인식을 이용한 호처리를 도입하겠다고 발표했다[2]. Bell Northern Research (BNR)는 1989년 Ameritech와 지역 전화회사를 통해 자동화된 과금 서비스를 개시했다[3].

BNR은 1992년 중반부터 주식시세 서비스를 시작했다. 이용자는 이 시스템에 전화를 걸어서 종목명을 말하면 현재의 주가를 알 수 있다[4]. 이 시스템은 서브워드(sub-word) 단위의 음성인식 기술을 이용함으로써 원칙적으로 각 단어를 모두 녹음하지 않고도 수백 또는 수천 단어를 인식할 수 있다.

이 논문은 대어휘 화자독립 음성인식 시스템인 KT-STOCK과 이 시스템에 대한 전화망을 통한 실험적 실용시험에 대해 기술하였다. 2절은 KT-STOCK의 개요에 대해 기술하였다. 3절은 실험적 실용시험에서 시스템의 성능을 분석하기 위한 데이타 수집 장치에 대해 기술하였다. 그리고, 4절은 실제 환경에서 수집된 음성데이타를 분석하고 자동끝점검출기와 인식기의 성능을 평가하였다. 마지막으로 5절은 결론을 맺었다.

2. 시스템 개요

2.1 시스템 개요

그림1은 KT-STOCK의 개략적인 구성도이다[5]. KT-STOCK은 크게 전화망 인터페이스부, 인식기, 데이타베이스 관리부로 구성되어있다. 사용자가 지정된 전화번호로 전화를 걸면, 전화망 인터페이스부를 통해 KT-STOCK과 접속된다.

음성 인식기는 IBM-PC 상에서 Texas Instrument 사의 TMS320C40 디지탈 신호처리기(DSP)를 이용하여 구현되었으며, 실시간에 가까운 인식속도를 보인다. 그림2는 음성인식기의 구성도이다. 음성인식기는 4개의 DSP를 이용하여 구현하였다. 그 중 하나는 끝점검출, 특징추출 및 벡터양자화에 사용되었다. 이 DSP에는 전화망을 통해 A/D 변환과 D/A 변환을 변환을 하는 전화망 인터페이스부가 접속되어 있다. 또한, 안내 음성 도중에도 음성을 인식할 수 있도록 하기 위해, least-mean-square 알고리즘을 이용하여 안내음성 제거기능을 구현하였다[6]. 그림3은 안내음성 제거의 효과를 보여준다. 나머지 세개는 Viterbi 검색에 사용되었다. 이 DSP는 병렬처리를 위해서 TMS320C40의 통신포트를 통해 특징추출용 DSP에 접속되어있다.

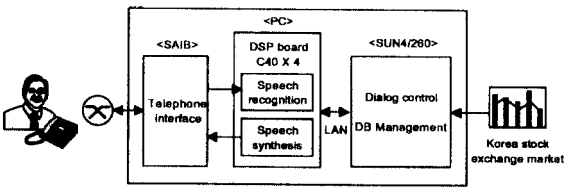


그림1 KT-STOCK의 전체 구성도

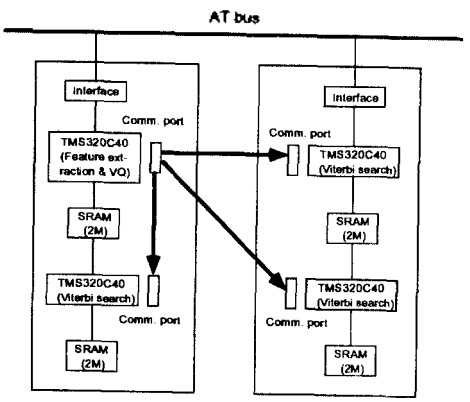
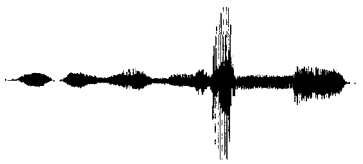
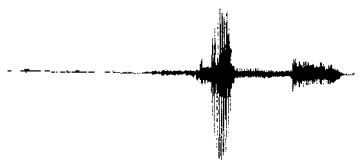


그림2 음성인식기의 구조



(a) Echo cancellation 이전



(b) Echo cancellation 이후

그림3 Echo cancellation 효과

데이타베이스 관리 프로그램은 워크스테이션에서 수행된다. 이것은 한국증권전산의 컴퓨터와 전용선으로 연결되어 있으며, 여기서 전송되어 오는 현재의 주식 정보를 관리한다. 주식시장에 새로운 회사가 상장될 때마다 새로운 어휘를 추가해야 하므로 시스템 관리자가 새로 상장되는 회사이름을 간단히 추가할 수 있도록 단어보다 작은 단위를 사용하는 음성인식 기술을 사용하였다. 또한, 한국어 단어를 음소의 열로 자동적으로 변환시키기 위해 한국어 음소생성기 또한 개발하였다.

2.2. 특징 추출

전화망을 통해 들어온 음성은 8kHz로 표본화되고, $(1 - 0.95 z^{-1})$의 전달함수를 갖는 필터를 사용하여 pre-emphasis 된다. 이 음성은 20 msec의 길이의 frame 단위로 분할된다. 이 frame은 10 msec씩 중첩된다. Autocorrelation 방법을 사용하여 14차 LPC 분석을 수행하고, 이 LPC 계수를 이용하여 cepstral 계수를 구한다. 이 계수에는 아래 수식의 window $W_c(m)$을 사용하여 weighting을 한다[7].

$$W_c(m) = 1 + \frac{Q}{2}\sin\left(\frac{\pi m}{Q}\right), \qquad 1 \le m \le Q$$

사용되는 음성특징으로는 이렇게 구한 weighted LPC cepstral 계수 외에 이것의 delta, delta-delta, 로그 파워의 일차, 이차 delta 값 등이 있다. 이 계수들은 각 종류별로 벡터 양자화(Vector Quantization)된다. 이때 4개의 벡터 codebook을 사용하는데, 로그 파워 부분은 64개의 codeword를 갖으며, 나머지 3개는 각각 256개의 codeword를 갖는다. 우리의 VQ 알고리즘은 Linde-Buzo-Gray (LBG) 방식에 기반을 두었다. 각각의 출력확률은 서로 독립적이라고 가정하고, 사용된다.

2.3. 음소 모델

HMM에 기반을 둔 음성인식 시스템을 위해서는 기본단위가 필요하다. 우리는 유사음소(phoneme-like phone)를 기본단위로 사용하였다. 먼저 61개의 문맥독립(context-independent) 음소 모델을 사용하여 시작했다. 그림4는 우리가 사용한 모델의 구조이며, 이는 Lee등이 사용한 모델과 유사하다[8].

이 모델은 7개의 state와 12개의 transition으로 구성된다. transition은 3개의 그룹으로 묶여진다. 같은 그룹에 있는 transition은 같은 출력 확률을 가진다. 문맥독립 음소 모델을 문맥종속(context-dependent) 음소모델로 확장하였다. 각 DSP에서 사용가능한 메모리의 크기를 고려하여 150개의 문맥종속 음소모델을 선정하였다. 통계적으로 신뢰성있는 모델을 생성하기 위하여 인식단위 축소법칙(unit reduction rule)을 사용하였다[9].

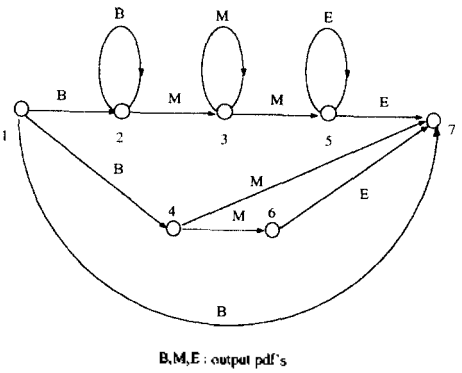


그림4 음소모델의 topology

2.4 모의실험 결과

서울에 거주하는 일반 남녀 100명에게 전화를 통해 전체 892 단어 중 1/3 씩을 각각 읽게 하였다. 이들에게 각 단어를 자연스럽게 읽도록 하였다. 이들의 연령은 20대 부터 40대 까지 분포한다. 100명 중 80명 분의 음성데이타는 음성인식을 위한 훈련에 사용하였으며 나머지 20명분은 인식시험에 사용하였다. 모든 음성은 끝점을 검출하여 인식시험을 위해 컴퓨터에 저장하였다.

음성데이타는 성별과 나이를 고려하여 표1에 보인 것 처럼 두 그룹으로 분류하였다.

표1 사용한 데이타 베이스의 특성

	훈 련 용		시 험 용	
연령	남성 (명)	여성 (명)	남성 (명)	여성 (명)
20	13	13	4	3
30	14	13	3	4
40	13	14	3	3
계	80 명		20 명	
	23,414 Token		5,772 Token	

음성인식 훈련에는 Baum-Welch 알고리즘을 사용하였다. 그리고 인식에는 Viterbi beam search 알고리즘을 사용하였다.

표2는 모의실험 환경에서 150 개의 context-dependent 음소 모델을 사용하였을 때의 인식률을 나타낸다.

표2 모의실험환경에서의 인식률

Top 1	Top 2
93.2(%)	97.8(%)

모의실험 환경에서는 첫번째 후보단어(Top1)에 대해 93.2%, 두번째 후보단어(Top2) 까지에 대해 97.8%의 인식률을 각각 얻었다.

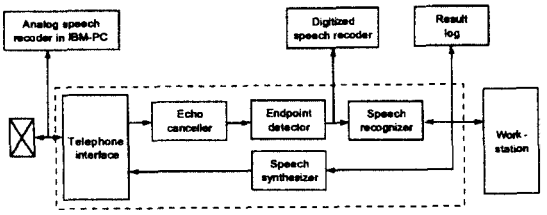


그림5 KT-STOCK의 데이타 기록점

표3 첫 일주일간 수집된 음성 데이타

	입력 음성	
	갯수	비율 (%)
고립단어 (자연스러운 발음)	367	75.4
배경 잡음	37	7.6
비허용 단어 (문장 포함)	31	6.4
숨소리	23	4.7
부자연스러운 발음	16	3.3
배경 대화	8	1.6
기침소리	5	1.0
계	487	100.0

75.4%의 사용자 만이 고립단어를 바르게 말하였으며 많은 사람들이 인식대상단어가 아닌 단어나 문장을 말하였다. 상당수의 배경잡음이 검출되었다. 이 결과는 사용자의 편이성을 위해서는 문장으로 부터 핵심 단어를 인식하는 기능이 필요 불가결하며, 배경잡음에 대처할 방법이 필요하다는 것을 보여준다. 표4는 이 시스템에 대한 사용자들의 행동을 보여 준다.

표4 시스템에 대한 사용자의 행동
(사용자들이 단어를 말하는 시점)

단어수	안내음성 도중	안내음성이 끝난 후
367	321	46
100 %	87.5 %	12.5 %

87% 이상의 사용자가 안내음성도중 단어를 말하였으며 13% 만이 전체 안내음성을 다 들은 후에 단어를 말하였다. 이 결과로 부터 안내음성 도중에도 음성을 인식할 수 있는 기능이 효과적이라는 것을 알 수 있다.

3. 데이타 수집 장치

실제 환경에서 데이타 수집의 주된 목표는 시스템의 성능을 분석하고 인식률 향상을 위해 음성데이타를 저장하기 위한 것이다. 이 목적을 위해 그림5에 보인 것처럼 3개의 기록 위치를 선정하였다.

첫번째 기록위치는 사용자와 시스템간의 전체 대화를 기록할 수 있도록 KT-STOCK 앞쪽의 two-wire channel에 위치한다. 대화만을 녹음할 수 있는 장치를 개발했다. 이 녹음장치는 시스템이 전화를 받을 때부터 동작하기 시작하며 전화가 끊어지면 동작을 멈춘다. 녹음기구로서 디지탈 오디오 테이프 (DAT) 녹음기를 사용하였다. 녹음된 대화는 상호대화에서의 여러가지 timing을 분석하기에 유용하도록 시스템의 안내음성에 사용자의 음성이 겹쳐있다. 두번째 기록점에서는 echo-canceller와 끝점검출기를 거친 디지탈 음성신호를 얻을 수 있다. 이 음성데이타는 인식과정으로 바로 들어가기 때문에 echo-cancellation과 끝점검출의 성능을 평가할 수 있다. 마지막으로 인식결과와 시스템 안내 문자를 포함한 시스템 동작 정보를 세번째 기록점에서 기록한다.

4. 실험적 실용시험

4.1 실제환경에서 수집된 음성 데이타 베이스

이 시스템은 1994년 6월 25일 이래로 한국통신 연구센터에서 시험운용에 들어 갔다. 사용자들은 526-5900번으로 전화를 걸어 이 시스템을 사용할 수 있다. 회선 수는 1회선 만을 운용하였다. 시험운용 전에 시스템 사용법을 설명하기 위해 설명회를 개최하였다. 이 시스템은 고립단어만을 인식하므로 사용자에게 주의사항을 설명하기 위해 설명회가 필요하였다 현재까지는 시험운용의 기본적 데이타 만을 얻은 상태이다. 표3은 두번째 기록점에 기록된 입력음성을 보여준다.

4.2 자동 끝점 검출기

자동 끝점검출기의 목적은 입력신호로 부터 의미있는 부분을 음향적으로 분리하는 것이다. 여기서 사용한 끝점 검출기는 기본

적으로 Lamel의 끝점 검출기에 기반을 두고 있다[10]. 그러나, 실시간으로 음성을 검출하기 위해 알고리즘을 수정하였다[11]. 두번째 기록점에 기록된 음성으로 끝점검출기의 정확도를 조사하였다. 고립단어의 경우 3.3%의 단어만이 전체단어 중 한 두 음절을 빠뜨렸다. 그러나 대부분의 음성은 고립단어에 말각하는 소리, 숨소리나 배경 잡음 등이 첨가되어 있었다. 일단 끝점검출기가 음성의 일부분을 빠뜨리면, 인식기는 그 오류를 복구할 수 없다. 그러나 고립단어가 잡음을 포함하는 경우는 인식과정에서 잘 인식이 되었다. 그러므로, 끝점검출기의 동작특성을 삭제오류를 줄이는 쪽으로 조절하였다.

4.3 실제환경에서의 인식결과

인식률은 세번째 기록점에 기록된 데이타와 두번째 기록점에 기록된 디지탈 음성 데이타로 부터 구할 수 있다. 결과를 표5에 나타내었다.

표5 실제환경에서의 인식결과

Top 1	Top 2
80.8(%)	86.3(%)

실제환경에서 첫번째 후보단어(Top 1)에 대해 80.8%, 두번째 후보단어까지(Top 2)에 대해 86.3%의 인식률을 얻을 수 있었다. 표2에서 모든경우를 고려한다면 인식률은 61.9%로 떨어진다. 즉 61.9%의 질문만이 원하는 답을 얻을 수 있었다. 모의실험 결과와 실제환경에서의 인식률은 큰 차이를 보였다. 그 주된 이유는 아래와 같다.

o 낭독과 자연스런 발성의 차이: 모의실험 환경에서는 단어들이 상당히 정확히 발음되었다. 그러나, 실제환경에서 수집된 음성 데이타에는 더듬거림, 숨소리 등이 포함되어 있었다.

o 배경잡음: 실제환경에서 수집된 음성 데이타에는 배경 잡음이 많이 포함되어 있었다.

이러한 문제에 대해서는 앞으로 더 연구할 계획이다. 실제 데이타를 더 수집한 후에 HMM의 파라미터를 재조정할 계획이다. 그 후에 다시 시험 운용을 할 것이다

5. 결 론

이 논문에서는 데이휘 화자독립 음성인식 시스템인 KT-STOCK과 이 시스템의 시험운용에 대해 소개하였다. KT-STOCK은 전화를 걸어서 종목명을 말하면 현재의 주가를 알 수 있다. 이 시스템은 HMM 기술을 이용한 화자독립 고립단어 인식시스템으로 710 단어를 인식하며, 150개의 분맥종속 한국어 음소 모델을 기본 인식 단위로 사용하였다.

두 종류의 실험을 하였다. 모의 실험에서는 93.2%의 인식률을 얻을 수 있었다. 그리고, 실제환경에서의 성능을 평가하기 위해 시험운용에 들어갔다. 시험운용 결과를 분석하기 위하여 3개의 기록점을 선정하였다. 이 기록점으로 부터 수집된 데이타로 부터 성능을 평가 하여, 61.9%의 인식률을 얻을 수 있었다. 실제 데이타를 더 모은 후에 HMM 파라미터를 재조정하고, 다시 시험운용을 계속할 계획이다.

<< 참고문헌 >>

[1] R. Nakatsu, "Anser: an applaication of speech technology to the Japanese banking industry,", *IEEE computer*, Vol. 23, No. 8, pp. 43-48, Aug. 1990.

[2] D.B. Roe and J. G. Wilpon, "Whither speech recognition: the next 25 years," *IEEE computer*, Vol. 31, No. 11, pp. 54-62, Nov. 1993.

[3] M. Lennig, "Putting speech recognition to work in the telephone network," *IEEE computer*, Vol. 23, No. 8, pp. 35-41, Aug. 1990.

[4] M. Lennig et al., "Flexible vocabulary recognition of speech," in *Proc. 1992 Int. Conf. on Spoken Lang. Processing*, pp. 93-96, Oct. 1992.

[5] M.-W. Koo et al., "KT-STOCK: A speaker-independent, large-vocabulary speech recognition system over the telephone," To be published in *Proc. 1994 IEEE Int. Conf. Acoust., Speech, Signal Processing*, Sep., 1994.

[6] B. Widrow and S. D. Stearns, *Adaptive signal processing*. Prentice-Hall, Inc. Englewood Cliffs, N. J., 1985.

[7] C. H. Lee et al., "Acoustic modeling for large vocabulary speech recognition," *Computer Speech and Language*, No. 4, pp. 127-165, 1990.

[8] K.-F. Lee, *Automatic speech recognition: the development of SPHINX system*, Kluwer Academic Publisher, Norwell, Mass., 1989.

[9] C. H. Lee et al., "Acoustic modeling of subword units for speech recognition," in *Proc. 1990 IEEE Int. Conf. Acoust., Speech, Sigal Processing*, pp. 721-724, April 1990.

[10] L. F. Lamel et al., "An improved endpoint detector for isolated word recognition," *IEEE Int. Conf. Acoust., Speech, Signal Processing ASSP-29*, No. 4, pp. 777-785, Aug. 1981.

[11] S. J. Doh et al., "A real-time endpoint detection algorithm for speech signal," in *Proc. 1992 Korean Signal Processing Conf.* pp. 11-14, Sep. 1992.