2000년도 한국음향학회 학술발표대회 논문집 제 19 권 제 2(s) 호

# 전화음성인식을 위한 멀티채널 음성인식 시스템 구현

이승훈, 서영주, 강동규
㈜코아보이스

# Implementation of the Multi-Channel Speech Recognition System for the Telephone Speech

Siong-Hun Yi, Youngjoo Suh, Dong-Gyu Kang
CoreVoice, Inc.
shyi@corevoice.com

## 요 약

본 논문은 전화음성 서비스 시스템의 핵심 기술인 멀티채널 음성인식 시스템의 구현에 대해서 기술하고자 한다. 구현한 시스템은 전화망 인터페이스 모듈, 음성입력 모듈, 음성인식 모듈, 및 서비스 제어 모듈로 구성되어 있다. 전화망 인터페이스 모듈은 전화망을 이용한 교환기와의 호 처리 및 이벤트 처리를 담당하며, 전화망 접속카드와 밀접한 관계를 가지고 있다. 음성입력 및 인식 모듈은 호 접속이 이루어진 채널로부터 음성을 입력 받아 단어인식 기능을 수행하는 부분으로서 멀티 채널을 수용할 수 있는 구조로 설계되어 있다. 음성인식 모델은 문맥 종속형 CHMM 모델이며, 각각의 HMM 모델은 3-state, skip path 로 구성되어 있다. 음성인식 모듈내의 함수들은 모두 re-entrant 하도록 구성함으로써 멀티 채널이 가능하며, 각각의 채널은 모두 독립적인 메모리 공간에서 동작하도록 되어있다. 이와 같은 멀티채널 전화음성인식 시스템은 Dialogic 보드를 이용하여 Windows NT 에서 동작하도록 구현하였다. 실험결과, 구현된 시스템은 실시간으로 상용서비스가 가능한 인식율을 보였으며 원활한 멀티채널 지원이 가능하였다.

## 1. 서 론

CTI 분야는 1990 년 초기에 콜 센터를 시작으로 각광 받기 시작하여 VMS, UMS 등에 적용되어 왔다. 초기의 CTI 는 교환기를 기본으로 사용하였으나 최근에는 교환기 없이 컴퓨터 만으로 CTI 를 구축하는 기술들이 선보이고 있으며, CTI 와 인터넷을 연결하여 자원을 보다 효율적으로 활용하는 기술들도 활발히 개발되고 있다. 예를 들면 인터넷 전화는 전화, 컴퓨터 및 인터넷이 결합된 새로운 서비스이다. 또한 ARS 분야에서도 인터넷을 이용하여 서비스할 수 있는 VXML 기반 시스템, Voice Browser 등에 대한 연구가 세계적으로 대두되고 이에 대한 표준화 작업이 진행되고 있는 실정이다. CTI 의 기술이 ITI 로 발전하면서 가장 중요하게 대두되는 원천 기술들은 음성인식, 음성합성, 대화관리기술 등을 들 수 있다.

전화음성인식은 현재 CTI 를 이용한 음성서비스 시스템에서 음성합성기와 함께 중요한 서비스 기능으로 대두되고 있다. 특히 다수의 사용자에게 서비스를 제공해야 하므로 멀티채널 기능은 필수적이다. 본 논문에서는 이와 같은 전화망 음성서비스를 위한 멀티채널 음성인식기의 구현에 대해서 기술하였으며 다음과 같은 순서로 구성되어 있다. 먼저, 제 2 절에서는 음성 인식기의 구성 및 세부 사항에 대해서 설명한다. 제 3 절에서는 음성인식 엔진을 이용한 멀티채널 음성서비스 시스템의 요구사항 및 실제 구현 기술에 대해서 설명한다. 제 4 절에서는 구현된 시스템의 성능을 평가하기 위하여 수행한 실험에 대해서 기술하고, 제 5 절에서 결론

을 맺고자 한다.

## 2. 음성 인식기

본 논문에서 구현한 음성 인식기는 크게 나누어 음향모델 모듈, 인식단어 모듈, 및 인식엔진 모듈로 나눌 수 있으며, 그림 1과 같은 구조로 입력음성에 대해서 인식결과를 생성한다.

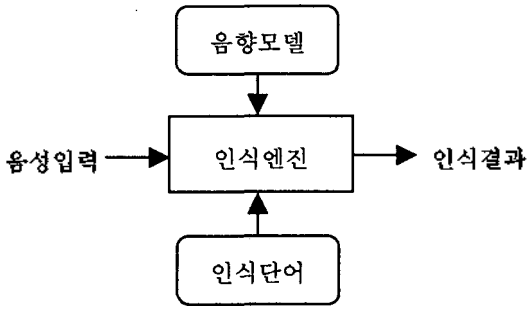


그림 1. 음성 인식기의 구성

### 2.1 음향모델 모듈

음향모델 모듈은 입력음성에 대해서 인식과정을 수행하는데 필요한 모델들에 대한 HMM 파라미터들을 가지고 있는 부분으로서 인식기의 훈련과정과 밀접한 관계를 가지고 있다. 음향모델링은 context-dependent triphone 모델로 구성하였으며, tree-based clustering을 이용한 state-tying을 적용하여 사전에 정해진 수의 모델들을 추출하였다[1]. 또한 각각의 모델들은 3-state, skip path, multiple mixtures를 가지는 continuous density HMM으로 구성하였다. 이와 같은 음향모델들에 대한 훈련은 Baum-Welch 알고리즘을 사용하였다[2][3].

### 2.2 인식단어 모듈

인식단어 모듈은 인식엔진 내부에서 word network를 구성할 때 필요한 발음사전의 대상어휘를 가지고 있는 부분으로서, 단어목록을 변경함으로써 즉시 새로운 단어에 대한 인식을 할 수 있다. 현재 구현된 인식기의 단어목록의 크기는 수 천 단어급이다.

### 2.3 인식엔진 모듈

인식엔진 모듈은 음향모델과 인식대상 어휘를 이용하여 입력음성에 대한 인식을 수행하는 부분으로서, 현재 화자독립/고립단어 인식이 가능하다. 인식엔진 모듈에서 사용하는 특징벡터는 MFCC, 이에 대한 first 및 second regression coefficient를 사용하였다. 또한 전화망 채널 보상 파라미터로는 CMS를 적용하였다. 음성인식 방법으로는 viterbi beam search알고리즘을 이용한 decoding기법을 사용하였다.

## 3. 멀티채널 시스템 구현

전화음성인식을 위한 멀티채널 음성인식 시스템을 구현하기 위해서는 다음과 같은 사항들을 먼저 고려하여야 한다. 첫째, 지원하고자 하는 채널 수의 규모에 따라서 아날로그 또는 디지털 전화망 인터페이스 보드를 선택하여야 한다. 둘째, 시스템 자원의 효율성, 처리속도, 대용량 시스템으로의 확장성, 및 OS별 호환성 등을 고려하여 서비스 시스템 프로그래밍 모델을 결정하여야 한다. 셋째, 시스템의 핵심인 멀티채널 음성인식 엔진을 최적화하여야 한다. 이는 인식속도, 시스템 자원, 및 인식성능 모두를 고려하여 결정한다. 본 논문에서는 이와 같은 사항들을 고려하여 그림 2와 같은 구조의 멀티채널 시스템을 Dialogic보드를 이용하여 Windows NT에서 구현하였다.

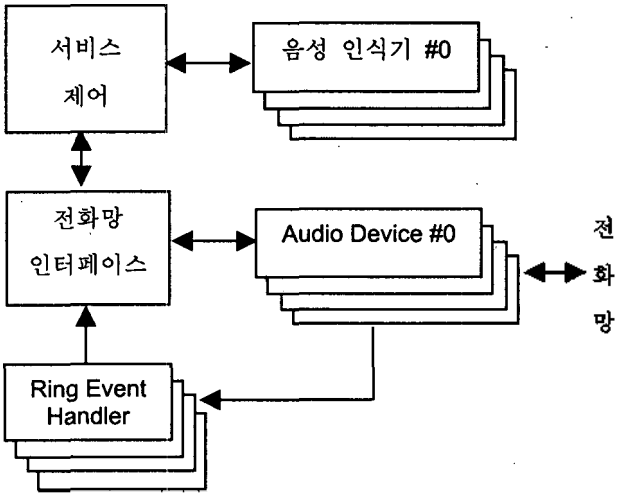


그림 2. 멀티채널 시스템 구조

### 3.1 전화망 인터페이스 모듈

전화망 인터페이스 모듈은 전화망을 이용한 교환기와의 호 처리 및 이벤트 처리를 담당하며, 하드웨어와 밀접한 관계를 가지고 있다. 대부분의 하드웨어 제공업체들은 전화망 인터페이스를 위해서 회사에서 공급하는 자체적인 라이브러리로 구성된 API, MS사의 TAPI 규격을 만족하는 API 등 여러 가지 API를 제공한다. 전화망을 이용한 ARS 서비스 시스템에서는 음성입출력 이외에 전화기와 관련된 호 처리 기능이 필요하다. 전화망 API는 동기 및 비동기 함수가 있으며 비동기의 경우 어떤 작용의 시작은 함수를 호출 함으로써 이루어지고 이벤트에 따르는 callback 함수에 의해 종료된다. 그러므로, 함수 호출 후 함수의 동작이 완료되기 전에 관련된 함수를 호출하면 시스템에 에러가 발생하거나 정상적인 결과를 얻을 수 없으므로 주의가 필요하다. 본 시스템은 프로그램의 변경 없이 다른 CTI 보드를 사용할 수 있는 MS TAPI를 이용한 시스템과 Dialogic API를 이용한 시스템을 구현하였다[4].

### 3.2 음성입력 모듈

음성입력 모듈은 전화망 인터페이스 모듈에서 접속된 전화라인으로부터 음성을 입력 받는 부분으로서, 연속적으로 입력되는 음성을 인식하기 위해서는 여러 개의 메모리 버퍼를 이용하여 오디오장치로부터 순차적으로 입력 받으면서 연속성을 유지한다. Dialogic board 상에서 특히 아날로그의 경우 기존의 음성 입출력 함수를 이용하여 이 같은 연속 음성인식을 수행하기에는 많은 제약이 따른다. 또한 음성 메시지가 출력되는 도중에 사용자가 발성한 음성을 인식해야 할 경우가 있다. 이 경우에는 상대방의 음성이 검출되면 즉시 음성출력을 중지하고 음성 데이터를 받아 들여서 음성인식 모듈로 전송해야 한다. 이러한 기능을 barge-in 이라 한다[5]. 구현된 시스템에서 MS TAPI 로 작성된 버전은 TAPI 규격상 이 기능이 지원되지 않는다. 그러나 vendor specific API 로 작성된 시스템은 이 기능을 지원한다. 즉 barge-in 이 검출되면 요청한 두개의 메모리 버퍼에 번갈아 가면서 내부적으로 녹음이 수행되고 하나의 버퍼가 채워질 때 마다 버퍼 callback 이 호출된다. 버퍼 callback 에서는 녹음된 데이터를 다른 메모리에 복사한 다음 계속적인 녹음을 위해 입력요청을 하면 연속적인 음성을 입력 받아 음성인식을 수행할 수 있다.

### 3.3 음성인식 모듈

음성인식 모듈은 입력 받은 음성에 대한 음성인식을 수행하는 부분으로서, 모듈내의 함수들은 모두 re-entrant하도록 구성함으로써 멀티 채널이 가능하도록 구현하였으며, 각각의 채널은 모두 독립적인 메모리 공간에서 동작하도록 되어있다. 또한 시스템이 점유하는 메모리를 줄이기 위하여 모든 채널이 공유할 수 있는 데이터는 가능한 한 모두 공유할 수 있도록 설계함으로써 메모리를 최적화 하였으며, 사용된 시스템 함수들은 모두 멀티 쓰레드가 지원되는 것들을 사용하였다. 내용이 변경되는 전역 변수를 공유해야 하는 경우에는 변수를 읽거나 쓰는 시간을 동기화 함으로써 정확한 처리가 가능하도록 하였다. 음성서비스 시스템에서 인식성능 다음으로 중요한 요소인 가용 채널 수를 최대화 하기 위하여 인식모듈 내에서 채널별 처리시간을 최소화 하였다. 채널별 처리시간을 최소화 하기 위한 방법으로는 음성 특징 파라미터 최적화, 탐색 알고리즘 최적화, 관측확률 계산 최적화, 고빈도 함수 최적화 등의 기법을 사용하였다[6].

### 3.4 서비스제어 모듈

서비스제어 모듈은 위에서 설명한 각각의 모듈들을 관장하면서 실제 서비스를 제공하는 부분으로서, 시스템 통합 측면에서 매우 중요한 역할을 한다. 소규모의 서비스를 제공하는 경우에는 synchronous 전화망 인터페이스 모델이 적합하나, 대규모 서비스인 경우에는 asynchronous 모델 및 분산형 event handler를 사용하여야 한다. 또한 OS별 호환성을 높이기 위해서는 windows callback handler 대신 자체적인 callback handler를 사용하는 것이 유리하다.

## 4. 성능 평가

### 4.1 음성 데이터베이스

실험에 사용한 음성 데이터베이스는 과기부지원으로 원광대 이용주 교수팀에서 제작한 것으로서, 사무실이나 가정집에 있는 유/무선 전화기를 이용하여 녹음되어 있다. 이 음성 데이터베이스는 다양한 음운환경 조합을 고려한 3,813개의 어휘로 이루어져 있으며, 발성에 참가한 화자는 남자, 여자 합하여 500명이며 이들이 76개씩 나누어 발성함으로써 총 38,000개의 단어로 구성되어 있다. A/D 방식은 8kHz, 8bit, u-law 형식으로 녹음되었다. 이와 같은 데이터베이스를 이용한 훈련 및 인식 실험은 5:1의 비율로 나누어 사용하였다.

### 4.2 성능 평가

위의 데이터베이스를 사용하여 음성인식 성능을 평가한 결과, state-tying 을 적용한 triphone 모델이 음소모델에 비해서 약 52%정도의 에러감소 효과를 얻었다. 또한 triphone 모델에서 특징 파라미터를 최적화 함으로써 최대 8.9%정도의 에러감소 효과를 얻었다. 한편, 인식속도 개선면에서는 state-tying 을 적용한 triphone 모델을 기준으로 하여, 이 모델에 여러 가지 최적화 알고리즘을 적용하여 약 16.7%정도의 에러증가에 대해서 80% 정도의 인식속도 감소를 얻을 수 있었다.

## 5. 결론

본 논문은 전화음성 서비스 시스템의 핵심 기술인 멀티채널 음성인식 시스템의 구현에 대해서 기술하였다. 시스템 구현단계에서 자원의 효율성, 처리속도, 대용량 시스템으로의 확장성, 및 OS별 호환성 등을 고려하여 시스템 프로그래밍을 모델링 하였다. 또한 시스템의 핵심인 멀티채널 음성인식엔진은 인식속도, 시스템 자원, 및 인식성능 모두를 고려하여 최적화하였다. 인식 시스템은 Dialogic보드를 사용하여 Windows NT에서 구현하였다. 구현된 시스템은 실시간으로 상용서비스가 가능한 인식율을 보였으며 멀티채널을 지원할 수 있었다. 앞으로 보다 많은 가용 채널 수의 확장, 인식성능 향상을 위한 알고리즘 개선, 및 음성합성기와 연동을 위한 barge-in기능의 성능을 보완하는 작업을 추진할 계획이다.

## 참 고 문 헌


[1] S.Young, J.J. Odell, and P.Woodland, "Tree-based state-tying for acoustic modeling," *DARPA Workshop on Human Language Technology*, pp. 289-291, 1994.

[2] L. Deng, M. Lennig, V.N. Gupta, P. Mermelstein, "Modeling acoustic-phonetic detail in an HMM-based large vocabulary speech recognizer," *Proc. of ICASSP*, pp. 509-512, 1988.

[3] Kai-Fu Lee, *Automatic Speech Recognition*, Kluwer Academic Publisher, pp. 103-106, 1989.

[4] Voice Programmer's Guide for Windows NT, Dialogic Corporation.

[5] Dialogic Barge-in Development Package Reference for Windows NT, Dialogic Corporation.

[6] P.Moreno and R.Stern, "Source of Degradation of Speech Recognition in the Telephone Network," *Proc. Of ICASSP* , pp. 109-112, 1994.