1999년도 한국음향학회 학술발표대회 논문집 제18권 제2(s)호

웹상에서의 HMM을 이용한 한국어 음성인식

최광국° , 이재왕, 김 철, 최승호
동신대학교 정보통신공학과

Speech Recognition using HMM over the WWW

Kwang-Kook Choi° , Jae-Wang Lee, Cheol Kim, Seung-Ho Choi
Dept. of Information and Communication Eng. , Dongshin University
e-mail: shchoi@dongshinu.ac.kr

요 약

본 논문에서는 웹상에서의 음성인식 시스템을 구현하기 위해 자바애플릿과 연속분포HMM을 이용하여 단어단위 인식을 실행하였다. 이 시스템은 Browser-embedded 모델로 구성되었으며 클라이언트컴퓨터에서는 애플릿으로 음성을 처리하여 특징파라미터들을 인터넷을 통해 서버컴퓨터로 보내고, 서버의 음성인식기는 전향 알고리듬을 적용하여 인식된 결과를 클라이언트컴퓨터에게 보내어 문자로 출력하도록 설계하였다. 훈련DB는 자동차 항법시스템에서 사용되는 22개 단어로 구축되었다.

1. 서론

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

웹상에서는 시스템 상호간에 대화식의 형태로 처리되는 요소로 HTML언어가 있다. 이러한 HTML은 텍스트, 사운드, 비디오 등의 정보를 포함한 멀티미디어 정보들을 통합하여 웹에서 바로 억세스할 수 있어야 하고 멀티미디어 순서들을 인식하지 못하는 웹데이터들은 hleper application 이나 third party S/W application을 사용하여 독립적인 특징을 지닐 수 있도록 통합 처리되어야 한다. 이러한 개발방식들을 통합 처리한 것이 자바 O/S이다.

자바의 특징은 제한적인 요소 없이 인터넷을 통해 정보들을 상호적으로 억세스하여 공유하고 자바 코드와 클래스 등은 다른 프로그램들에 저장을 주지 않는다. 또한, 자바는 분산컴퓨터 구조와 통신환경에 따른 응용프로그램 개발에 적합한 언어로써 음성분야에 접목되어 여행안내시스템, 호텔 예약시스템 등과 같은 멀티미디어분야의 새로운 정보통신서비스의 창출이 가능하다.

이러한 점에 착안하여 본 논문에서는 단어단위의 HMM 음성인식 시스템을 웹상에서 처리하기 위해 자바를 이용하였다.

2. 시스템의 구조

본 논문에서는 음성인식시스템을 Stand-alone과 browser-embeded 모델로 구분하여 설계하였다.

2.1 Stand-alone 모델

이 모델의 클라이언트는 음성을 record/send 하고 인식된 text를 출력할 수 있는 프로그램을 필요로 하며, 이러한 프로그램은 자신의 유저머신인터페이스를 통해 유저들에게 인식된 결과를 보여준다. 또한 이 모델은 클라이언트의 H/W와 S/W 자원을 공유할 수 있기 때문에 사운드카드의 제어가 쉽다는 장점이 있으나 프로그램의 갱신 및 이식성이 어렵다는 단점이 있다[1]. 그림2.1은 Stand-alone의 모델을 나타낸 것이다.

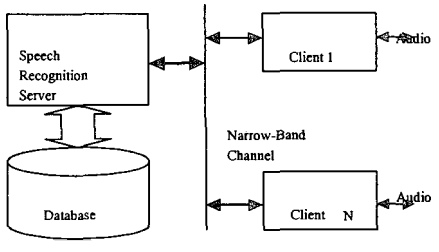

그림2.1 Stand-alone 모델

2.2 Browser-embeded 모델

이 모델은 시스템을 쉽게 갱신하고 이식성이 높아 다른 S/W를 제공할 필요가 없어 Stand-alone의 문제점을 해결할 수 있다. 특히, HMM을 이용한 단어 음성인식 시스템에 적합하도록 음성인식 서버를 억세스하는 웹브라우저에서 자바애플릿으로 음성을 record/send하고 그에대한 결과를 클라이언트에 나타날 수 있도록 설계되어져야 한다. 그러나 자바애플릿으로는 음성을 recoding할 수 없고 local machine은 프로그램과 H/W에 대한 보안성때문에 제어가 불가능하다.

그림2.2는 음성인식 시스템을 Browser embeded모델로 구성한 것으로써 클라이언트컴퓨터의 마이크로폰는 음성을 record하기위해 사용하고, 로컬프로세스는 웹브라우저로 실행되어 다운로드되어진 애플릿에서 음성데이터를 전송하거나 recording된 음성을 확인한다. 또한, 서버컴퓨터는 웹페이지를 갖고있는 웹서버, 음성인식을 관리하는 음성인식서버, 음성을 인식하는 음성인식기 등으로 구성된다. 여기에서 소켓1은 클라이언트컴퓨터의 웹브라우저와 웹서버를 연결하고 소켓2는 애플릿과 음성인식서버를 연결해준다[2].

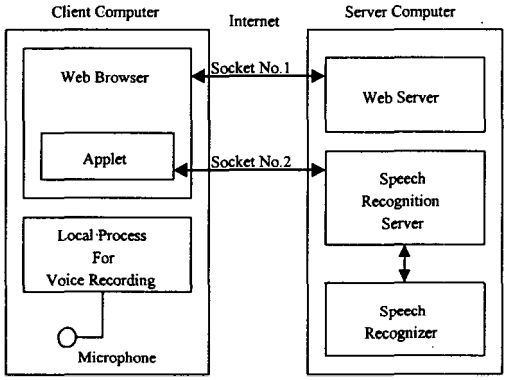

그림2.2 Browser-embeded 모델

3. 자바를 이용한 음성인식시스템의 구현

3.1. 자바애플릿

자바애플릿은 Netscape Navigator나 Microsoft Internet Explorer같은 웹브라우저 안에서 자바 프로그램을 실행할 수 있도록 디자인 되었지만 웹상의 사용자 컴퓨터에 대한 보안성을 지니고 있기 때문에 본 논문의 음성인식시스템에서는 음성신호를 recoding하기 위해 미국 Scrawl, Inc. (http://www.scrawl.com)의 third party program인 SoundBite 클래스를 사용하여 해결하였다[2].

그림3.1은 애플릿으로 처리된 사용자 인터페이스이다.

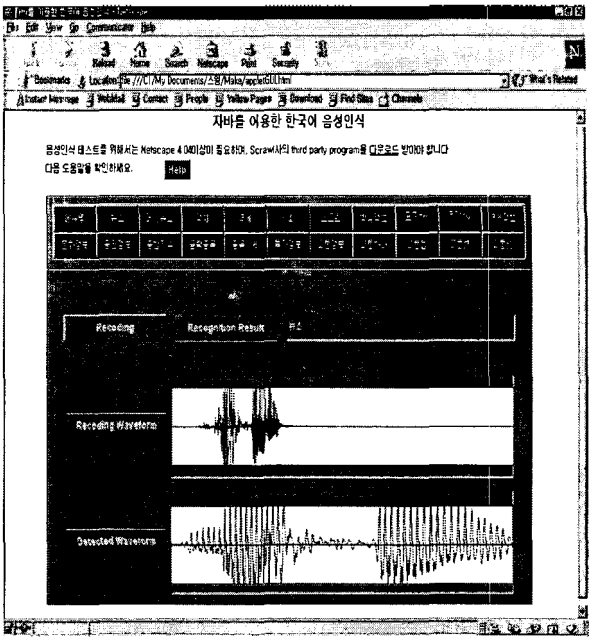

그림3.1 자바를 이용한 음성인식시스템의 사용자인터페이스

3.2 애플릿의 처리과정

애플릿은 크게 3가지 역할을 수행하고 있다. 첫째, 사용자의 음성 recoding 둘째, 음성의 특징파라미터 추출을 위한 전처리 셋째, 인식서버에게 특징파라미터를 전송하기 위한 사용자의 소켓 생성부분이다.

1) record와 play의 설계

SoundBite 클래스를 사용하여 음성의 record/play를 설계한 구성도를 그림3.2와 같이 나타내었다.

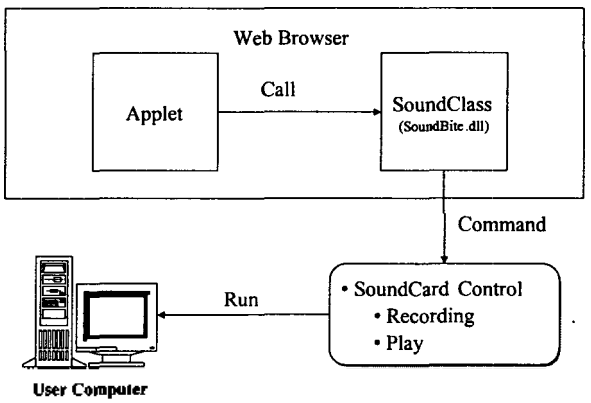

그림 3.2 음성의 record/play 구성도

2) MFCC(Mel Frequency Cepstral Coeff.)의 추출

음성의 실시간 검출 후 MFCC를 구하는 과정을 그림3.3에 나타내었다[3][4].

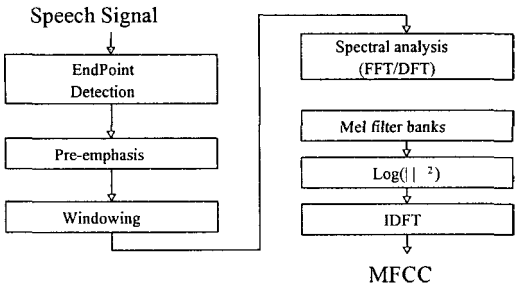

그림 3.3 MFCC의 추출 흐름도

3) 소켓생성의 설계

특징파라미터가 구해진뒤 애플릿은 인식서버와의 데이터 전송을 위해 소켓을 생성한다. 소켓이 정상적으로 연결되면, 애플릿은 파라미터를 인식서버로 전송하고, 인식서버가 인식결과를 애플릿으로 전송할때까지 소켓을 유지한다. 또한, 소켓은 클라이언트컴퓨터와 서버컴퓨터간의 음성처리 프로그램을 TCP/IP로 연결한다. 그림 3.4는 애플릿과 인식서버간의 소켓연결을 나타낸 것이다.

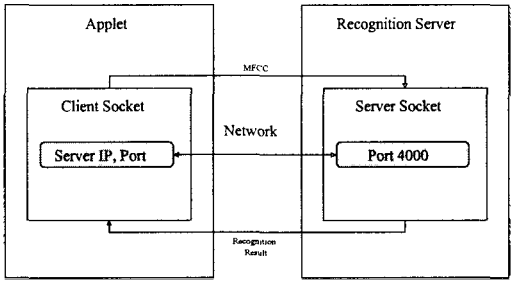

그림 3.4 애플릿과 인식서버와의 소켓연결 구성도

3.3 JNI(Java Native Interface)의 설계

자바로 응용프로그램을 설계시 자바로 표현 불가능한 부분을 해결하기위해 만들어진 JNI는 C, C++과 어셈블과 같은 다른 언어로 작성되어진 라이브러리나 응용프로그램과 함께 자바가상머신(Java Virtual Machine: JVM) 안쪽에서 자바 코드와 같이 실행될 수 있도록 설계된 것으며 그림 3.5는 일반적인 JNI을 이용한 응용프로그램 설계과정을 나타낸것이다[5].

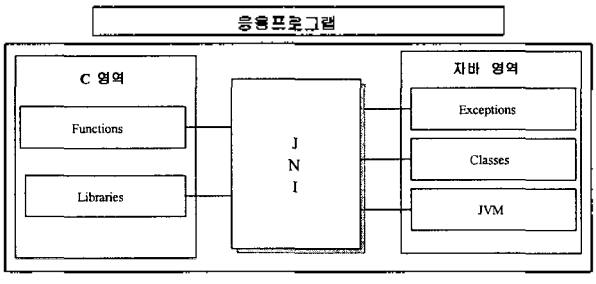

그림 3.5 JNI을 이용한 응용프로그램 구성도

따라서, 본 논문에서는 애플릿에서 클라이언트컴퓨터의 사운드 카드 제어부분과 인식서버와 인식기와의 연결부분에서 JNI을 사용하였다. 웹상에서 클라이언트컴퓨터의 사운드카드를 제어하기위한 JNI는 SoundBite 클래스를 사용했고 이 클래스는 자바를 지원하는 웹브라우저와 MS Window 95/98/NT 운영체제에서도 이용가능하도록 설계되었다. 그림 3.6은 SoundBite의 JNI 구성도이다.

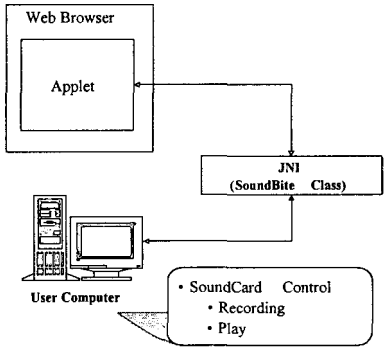

그림 3.6 SoundBite의 JNI 구성도

또한, 그림 3.7은 인식서버와 인식기 사이에 JNI을 적용한 구성도이다.

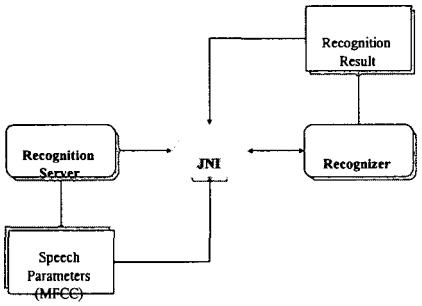

그림 3.7 인식서버와 인식기의 JNI 구성도

3.4 애플릿과 인식서버의 데이터 처리

인식서버의 소켓은 클라이언트측의 애플릿이 접속할 수 있는 특정포트를 활성화 시키고, 클라이언트가 접속할 때 사용할 소켓을 생성하여 접속요구에 응답한다.

애플릿은 인식서버의 IP어드레스와 포트번호를 가지고 초기화하며, 서버에게 접속을 요청한다. 접속이 성공적으로 이루어지면 애플릿은 인식서버에게 전송할 패킷을 생성하고 전송한다. 인식서버는 입력되는 패킷을 분석하고 패킷이 전달되어야 할 부분에 패킷을 전송한뒤 인식결과를 소켓을 통해 애플릿에 전송한다. 모든 처리가 끝나면 클라이언트의 소켓은 소멸되고, 서버의 소켓은 새로운 요청에 대비할 수 있도록 초기상태로 돌아간다. 그림3.8은 애플릿과 인식서버의 데이터 처리과정을 보여주고있다.

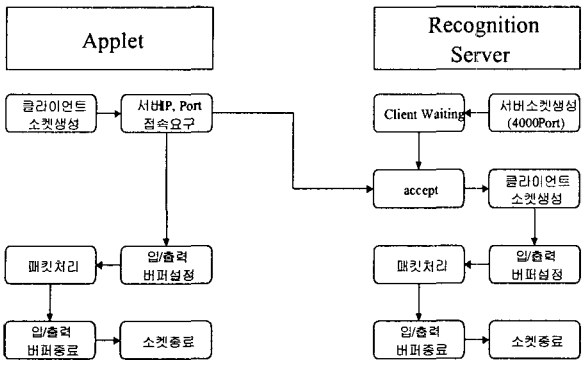

그림 3.8 애플릿과 인식서버의 데이터 처리과정 흐름도

3.5 음성인식기

음성인식기는 애플릿에서 전처리된 특징파라미터와 학습된 DB와의 최대 확률을 추정하여 인식된 결과를 애플릿에 문자로 전송하며 연속분포 HMM을 이용하여 단어인식을 수행한다. 인식기의 훈련용 DB는 자동차 항법 시스템에서 사용되는 22개 단어를 52명의 20대 남성화자가 조용한 실험실 환경에서 발성한 것을 수집하였다.

수집된 단어리스트는 그림 3.1의 사용자인터페이스 상단에 나타나 있다. 수집된 데이터는 샘플링주파수가 8Khz, 양자화레벨 16Bit, 채널은 Mono 형태의 Wave파일로 작성되며 끝점검출 알고리즘으로 자동검출하여 저장된다.

음성신호는 단어를 25msec의 프레임 단위, $1-0.97z^{-1}$ 의 FIR 필터로 전강조, Windowing은 해밍창을 사용하여 15msec씩 중첩되어 전처리된 후에 각 프레임은 12차 MFCC, 12차 Delta-MFCC, 1차 에너지, 1차 델타에너지 등 26개의 특징파라미터가 추출된다[3][4].

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

또한, HMM의 토플로지는 left-to-right모델로써 모든 단어에 대한 초기화 값을 설정해 준다. 이때, 초기상태 천이확률은 현재상태로 천이 할 확률 0.7과 다음상태로로 천이할 확률 0.3, 마지막 상태의 확률은 현재상태로 천이할 확률 1.0과 다음상태로 천이확률 0.0을 갖는다.

이러한 HMM 초기화 출력값들은 가우시안 혼합밀도함수, 전·후향 알고리즘, Baum-welch 재추정 알고리즘에의해 문턱값이 0.0005가 될 때까지 수렴하여 파라미터를 갱신한다. 이 데이터가 훈련DB에 저장되어 사용된다.

여기에서 서버컴퓨터는 Pentium III 450 H/W와 NT 4.0 O/S로 구성되었다. 이때, 서버컴퓨터의 음성인식기는 훈련DB와 시험데이타간의 최대확률을 추정하기 위해 전향알고리즘을 사용하였다.

4. 실험 및 결과

인식실험은 Stand-alone과 Browser-embeded모델로 구분하여 행하였다. 전자에서는 훈련에 참가한 22(단어/명)×18명= 396단어를 시험데이터로 사용하여 인식한 경우 1%의 오인식율이 구해졌으며, 후자에서는 훈련에 참가하지 않은 화자 5명이 22단어를 2회 발성한 220개를 시험데이터로 사용하여 9.1%의 오인식율이 나타났다.

이러한 결과를 유추해보면 Browser-embeded모델로 시행된 자바를 이용한 음성인식시스템에서는 실험에 참가한 클라이언트컴퓨터의 마이크로폰의 특징, 실험 당시의 상황, 적은 시험데이타등의 영향으로 높은 에러율을 가져왔다고 생각된다.

5. 결 론

본 논문에서는 한국어에 대한 음성인식 시스템을 웹상에서 구현하기 위해 자바애플릿을 사용하였다. 이 시스템은 Stand-alone모델보다 높은 에러율을 갖고 있지만 애플릿을 사용하여 멀티미디어 분야에 접목시켰다는 의미를 갖고 있다. 따라서 좀더 나은 음성처리에 대한 분석적 연구와 시스템의 H/W의 조건을 만족할 수 있다면 향후 새로운 정보통신서비스가 창출될 것으로 전망된다.

참 고 문 헌

[1] V.Digalakis, L. Neumeyer, Perakakis, "Quantization of Cepstral Parameter for Speech Recognition Over the World Wide Web,"Proceeding ICASSP 98, pp. 989-992, 1998

[2] ZhemlnTu, philips C. Loizou, "Speech Recognition Over the Internet Using JAVA," Proceeding ICASSP 99, pp. 2267-2370, 1999

[3] C. Becchetti, Prina Ricotti, *Speech Recognition*, John Wiley & Sons, 1999

[4] Steve Young, *The HTKBook (for version 2.2)*, Entropic Ltd., 1999

[5] SUN Web Site: *"http://www.java.sun.com/products/"*

본 논문은 동신대학교 1999년도 교내학술연구비에 의해 연구되었습니다.