# 웹 브라우징을 위한 음성 인터페이스 설계 및 구현

이 승 호*, 육 상 조*, 권 영 미**, 이 극*
*한남대학교 컴퓨터공학과 AI & HCI 실험실
**목원대학교 컴퓨터공학과

# Design & Implementation of Voice-Interface for Web-Browsing

Seung-ho Lee*, Sang-cho Youk*, Young-mi Kwon**, Geuk Lee*
*AI & HCI Lab, Department of Computer Engineering, Hannam University
**Department of Computer Engineering, Mokwon University

요 약

WWW은 무한한 확장 가능성을 지닌 HTTP(Hyper-Text Transfer Protocol)와 편리한 웹 브라우져를 통해 질적, 양적 성장 계속해 왔으며 특히 GUI(Graphic User Interface) 환경에서 동작하는 웹 브라우져는 WWW이 수많은 이용자를 확보하는데 일익을 담당했다. 본 논문에서는 이 웹 브라우져에 음성인식 기술을 접목하여 WWW의 이용자가 자신의 음성으로 편리하게 웹 브라우징을 할 수 있도록 하는 음성 인터페이스를 설계, 구현 한다. 본 음성 인터페이스는 계속적으로 입력되는 음성 정보 중 화자의 발성음을 추출하여 음성 인식기에 전달하는 음성 입력기와 화자의 발성을 인식하는 음성 인식기 그리고 인식결과를 웹 브라우저에게 처리 하도록 하는 결과 처리기로 구성되어 있다.

## 1. 서론

인터넷을 기반으로 WWW(World Wide Web)은 현재 가장 대중적이고 일반적인 정보망으로서 다수의 사용자를 확보하고 있으며 사회 각 층에서 이를 이용한 사업을 추진 중이다. WWW가 이처럼 단기간에 폭발적인 성장을 할 수 있었던 요인으로 HTTP(Hyper-Text Transfer Protocol)와 편리한 웹 브라우져(Web Browser)를 꼽을 수 있는데, 특히 GUI(Graphic User Interface) 환경에서 동작하는 웹 브라우져는 이전까지 키보드를 통해 명령을 입력하는 대신 간단히 마우스 클릭을 통해 원하는 정보에 접근할 수 있도록 함으로써 전문지식을 지니지 않은 일반 사용자도 직접 정보망에 접근하여 손쉽게 정보를 이용할 수 있는 기회를 제공하였다.

본 연구에서는 사용자가 마우스 대신 음성으로 웹 브라우징을 할 수 있도록 하는 음성 인터페이스를 설계하여 보다 손쉽게 WWW을 이용할 수 있는 방법을 제시한다.

이를 위해 2장에서 음성 인터페이스가 웹 브라우져 상에서 동작하기 위한 기반 기술인 Plug-in 기술과 MIME에 대해 조사하고 3장에서는 음성 인터페이스가 동작하는 모습을 보임으로써 음성 인터페이스의 이해를 돕고자 한다. 4장에서는 음성 인터페이스의 전체 구조와 이를 구성하고 있는 음성 입력기, 음성 인식기, 결과 처리기의 세 모듈을 설계하고 5장에서 음성 인터페이스를 활용하는 예를 보인다. 끝으로 6장에서 결론을 맺는다.

## 2. MIME과 Plug-ins

본 음성 인터페이스가 웹 브라우져 상에서 동작하기 위해 기반으로 하고 있는 Plug-in 기술은 MIME(Multipurpose Internet Mail Extensions)과 매우 밀접한 연관을 가지고 있다.

MIME은 멀티미디어 메일 전송의 필요성에 따라

부각된 인터넷 표준으로 인터넷 상의 특정 정보 형태를 구분하기 위해 사용된다. 따라서 인터넷 상의 정보들은 MIME에 따라 그 종류가 구분되고 객체로서의 성격을 띄게 된다. 넷스케이프사는 Plug-in 기술을 공개하여 제3의 개발자들이 자사의 웹 브라우져 상에서 동작하여 MIME 객체를 처리할 수 있는 소프트웨어를 제작할 수 있도록 하였다.

MIME객체는HTML의 <EMBED> 태그에 의해 HTML문서에 삽입된다. 이렇게 삽입된 MIME객체는 웹 브라우져에 의해 자동으로 처리되는데 그 과정은 다음과 같다. HTML 문서에서 MIME객체가 발견되면 웹 브라우져는 MIME에 따라 이와 관련된 플러그-인 프로그램을 결정하고 이를 기억장소에 적재한 다음 이 MIME객체를 플러그-인 프로그램에 전달함으로써 실행 시킨다.

플러그-인에 인수를 전달하기 위해 <EMBED>태그를 옵션 파라메터를 이용하는데 이는 HTML에서 정의하고 있는 것 외에 플러그-인 개발자에 의해 확장 가능하다. 플러그-인이 적재되고 초기화될 때 웹 브라우져는 EMBED태그 내에 있는 파라메터를 분석하여 이름, 값의 쌍으로 플러그-인에 전달한다. 본 음성 인터페이스에서는 이 기능을 이용하여 단어와 사용자가 단어를 발음했을 때 이동하게 될 URL의 쌍을 입력 받는다.

Plug-in 기술은 웹 브라우져 상에서 동작하는 소프트웨어를 제작하기 위한 인터페이스 기술이라고 할 수 있다. 자바 애플릿(JAVA Applet) 형태의 소프트웨어도 웹 브라우져 상에서 작동하지만 JVM(Java Virtual Machine) 위에서 동작하기 때문에 실행 속도가 느리며 아직 멀티미디어에 대한 지원이 미흡하여 사용자의 음성 입력이 불가능하다. 또한 보안 문제로 인해 본 음성 인터페이스에 적용하기 어렵다. 음성 인터페이스는 사용자의 음성을 학습하고 이를 사용자의 컴퓨터에 저장하고 있어야 하지만 자바 애플릿의 보안 기능은 이를 금지하고 있다.

## 3. 음성 인터페이스의 동작

기존의 웹 브라우져에서 사용자는 다른 웹 페이지로 이동하기 위해 마우스로 앵커를 클릭하였다. 그러면 웹 브라우져는 이 앵커의 URL정보를 이용해 해당 페이지로 이동하게 되는데 음성 인터페이스는 마우스 클릭 대신 사용자가 발음한 단어의 음성 신호를 입력 받고 이를 인식하여 단어에 지정된 URL을 웹브라우져의 새로운 위치로 지정함으로써 마우스로 앵커를 클릭한 것과 같은 역할을 수행하게 된다.

사용자가 본 음성 인터페이스가 동작 가능하도록 작성되어 <EMBED> 태그가 포함된 웹 페이지에 방문하면 웹 브라우져는 음성 인터페이스 플러그-인 소프트웨어를 적재한다. 이때 <EMBED> 태그의 옵션을 분석하여 이동할 URL을 단어와 함께 플러그-인에 전달한다. 이와 같은 초기화가 끝나면 사용자의 음성을 입력 받기 시작한다. 사용자가 단어를 발음하면 음성 인터페이스는 이를 인식하여 인식결과가 적절하지 않으면 오류를 통보하고 그렇지 않으면 해당 URL로 이동한다. 플러그-인은 플러그-인이 포함된 웹 페이지를 떠나면 기억장소에서 자동으로 삭제된다.

## 4. 음성 인터페이스의 설계

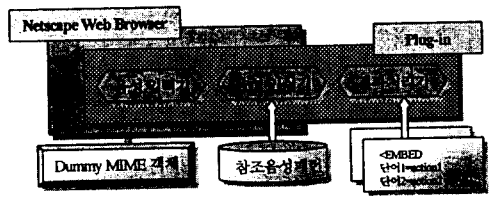
### 4.1. 음성 인터페이스의 구조

음성 인터페이스는 플러그-인으로서 넷스케이프 웹 브라우져 상에서 동작하며 <그림1>과 같이 음성 입력기, 음성 인식기, 결과 처리기의 세 모듈로 구성된다.

음성 입력기는 플러그-인의 시작과 함께 음성을 입력 받기 시작하여 사용자의 발화음을 검출한다. 그리고 이를 음성 인식기에 전달한다.

음성 인식기는 음성 입력기로부터 전달된 발화음을 정보를 미리 학습된 참조 음성 패턴과 비교하여 인식 결과 얻는다. 그리고 이 결과를 결과처리기에 전달한다.

결과 처리기는 <EMBED>태그의 인수로 전달된 URL중 인식 결과와 사상되는 URL을 웹 브라우져의 새로운 위치로 지정한다.




<그림 1> 음성 인터페이스의 구조

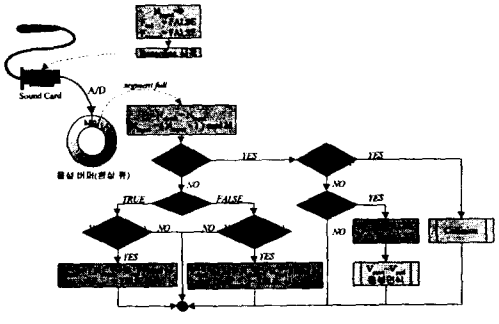
### 4.2. 음성 입력기의 설계

본 음성 인터페이스는 음성의 입력과 인식이 실시간으로 처리되므로 이미 입력된 음성 데이터를 처리

하는 것과는 다른 방식으로 동작한다. 파일 등의 저장매체에 음성이 저장되어 있는 경우, 파일이라는 한정된 양의 데이터만을 처리 대상으로 하지만 실시간 처리에선 발화자의 음성 정보가 입력되는 시기와 길이를 예측할 수 없다. 따라서 계속적으로 음향 데이터를 입력 받으면서 발화한 음성 정보가 입력되는지를 검사하여야 한다. 이를 위해 음성 입력기의 버퍼를 일정 크기의 세그먼트로 나누고 이를 환상큐의 형태로 구성한다.

환상 큐의 형태로 버퍼를 구성하면 버퍼의 끝까지 음성을 입력 받게 되었을 때 다시 버퍼의 처음 위치에 음성이 입력되므로 무한정의 시간을 대비해야 하는 상황에서 버퍼의 용량을 절약할 수 있다. 단, 음성 정보의 길이가 버퍼 전체 용량 보다 클 경우 버퍼의 입력 위치가 순환하여 입력된 음성 정보의 처음부터 덧씌워지게 되므로 음성 정보를 상실할 수 있게 된다. 이를 막기 위해 전체 버퍼의 크기를 입력 받을 음성의 최대 길이보다 충분히 크게 할당한다. 본 연구에서는 길지 않은 고립 단어를 인식 대상으로 하기 때문에 전체 버퍼의 길이를 음성을 3초간 기록할 수 있는 크기로 할당하였다.

발화자의 음성이 입력되고 있는 지를 검사하기 위해 세그먼트 내의 에너지를 계산하는데 계산된 에너지 값이 특정한 수치보다 크게 되면 사용자의 음성이 입력되고 있다고 판단하고 이 세그먼트를 음성의 시작으로 설정한다. 그리고 계속적으로 음성을 입력 받으며 에너지가 일정한 값 보다 작은 세그먼트를 음성의 끝으로 판단하고 지금까지 입력된 음성을 음성 인식기에 전달하게 된다.




<그림 2> 음성 입력기의 구조

## 4.3. 음성 인식기의 설계

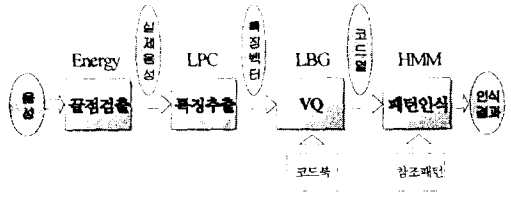
음성 인터페이스의 음성 인식기는 화자 종속의 고립단어를 인식하며 일반적인 음성인식기와 구조가 동일하다. <그림 3>은 음성 인식기의 구조와 각 과정별 처리 방법을 보이고 있다.

음성 입력기로부터 전달된 음성은 세그먼트 단위로 받아들여지기 때문에 음성의 실제 구간과 차이가 있으며 따라서 세밀한 음성 구간의 검출이 필요하다. 끝점 검출기는 세그먼트 단위 보다 적은 구간의 에너지를 계산함으로써 보다 정밀한 음성 구간을 검출한다.

이렇게 검출된 음성 정보는 특징 추출기를 거치며 대용량의 음성 데이터는 다차원 벡터로 변환된다. 이를 위해 비교적 계산 속도가 빠르면서도 전달 함수를 정확히 추정할 수 있는 선형 예측 부호화 방법을 사용한다.

다차원 벡터로 추출된 음성 특징은 LBG알고리즘에 의해 미리 생성된 코드북에 의해 벡터 양자화되어 코드 패턴열로 변환된다.

이 코드 패턴은 HMM알고리즘에 의해 참조 패턴과 비교되어 인식 결과를 도출하게 된다.

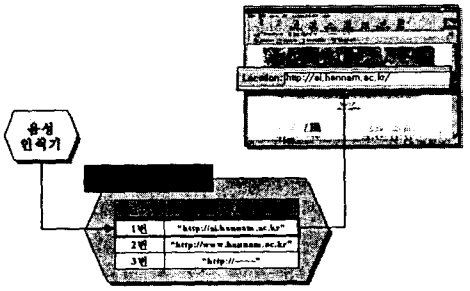



<그림 3> 음성 인식기의 구조

## 4.4. 결과 처리기

결과 처리기는 음성 인식기로부터 인식결과를 받아 <EMBED> 태그의 인수로 지정된 URL 정보들에 사상하여 해당 URL을 선택한다. 그리고 이 URL을 웹 브라우져가 새로운 위치로 지정하도록 한다. <그림 4>는 이와 같이 동작을 하는 결과처리기를 보인것이다.

웹 브라우져의 위치를 선택된 URL로 이동시키기 위해 Plug-in 기술에서 제공하는 API를 이용한다.

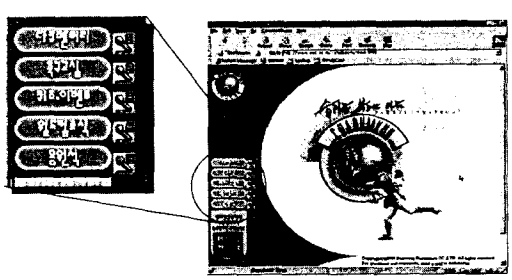



<그림 4> 결과 처리기의 구조

## 5. 음성 인터페이스의 활용

음성 인터페이스는 넷스케이프 사의 Plug-in 기술을 기반으로 설계된 플러그-인 프로그램이기 때문에 <EMBED> 태그에 의해 실행되며 이 태그 안에 사용자의 음성에 따라 이동하게될 URL을 명시한다. 따라서 음성 인터페이스를 사용 가능하도록 하기 위해 웹 페이지는 위의 사항을 포함하고 있어야 하며 이미 작성된 웹 페이지는 수정되어야 한다.

웹 페이지는 앵커를 가지고 있으며 사용자가 앵커를 마우스로 클릭하면 해당 URL로 이동한다. 음성 인터페이스는 웹페이지 전체에 분산되어 있는 앵커들을 자동적으로 인식하지 못하며 따라서 웹 페이지 작성자는 <EMBED>태그의 옵션 파라메터로 단어와 URL쌍을 지정해 주어야 한다. 또한 사용자가 앵커의 URL과 단어가 연관이 있어 단어를 발음하면 URL로 이동함을 알리기 위한 노력이 필요하다. <그림 5>는 일반적인 웹 페이지를 음성 인터페이스가 동작할 수 있도록 수정한 예를 보이고 있다. 왼쪽에 크게 확대된 그림은 사용자가 마우스로 클릭하면 이동하는 앵커들로 각 앵커의 옆에는 마이크 모양의 아이콘이 표시되고 있다. 이는 단어를 발음했을 경우 옆의 앵커의 URL로 이동 됨을 사용자에게 알려주는 것이다. 이 예제에서 사용자가 '1번'을 발음하면 '다큐멘터리' 페이지로 이동하게 된다.



<그림 5> 음성 인터페이스를 위해 수정된 웹 페이지

## 6. 결론

음성은 인간이 사용하는 자연스러운 의사 소통 수단으로 입력장치에 응용하면 키보드나 마우스 등과 같은 기계적인 장치보다 편리하게 사용할 수 있다. 일반인들은 키보드나 마우스에 익숙치 않으며 더욱이 장애자들은 이를 사용할 수 없으므로 정보 검색에 어려움을 겪게 된다.

현대는 정보시대이며 특히 WWW에서의 정보 활동이 활발히 진행되고 있어 현재 생성되는 대부분의 정보들은 WWW을 대상으로 출간되며 기존에 있던 정보들도 WWW으로 재출간되고 있는 실정이다. 또한 산업, 경제, 문화 전반에서 빠른 속도로 WWW과의 접목을 시도하고 있다. 이제 정보 검색은 새로운 형태의 사회 활동이며 이를 제대로 수행하지 못하는 것은 문맹과 같은 수준의 장애이다.

웹 브라우져는 WWW의 정보 검색을 위해 많은 사용자를 보유하고 있으며 이는 점차 확대되어질 전망이기 때문이다. 따라서 본 연구에서는 웹 브라우져 상에서 동작하는 음성 인터페이스를 설계, 구현하여 사용자들이 보다 쉽게 WWW을 이용할 수 있도록 하였다.

본 음성 인터페이스는 넷스케이프 사가 자사의 웹 브라우져를 위해 제공하는 Plug-in 기술을 이용하도록 설계, 구현하였다.

## 7. 참고문헌


[1] J. He, L. Liu and G. Palm, "A New codebook training algorithm for VQ-based speaker recognition", Proceeding of ECSA, pp.1091-1094, 1997.

[2] R. M. Gray, "Vector quantization technique", IEEE ASSP Magazine, Vol. 1, pp. 4-29, 1984.

[3] J. D. Markel, A. H. Gray, Jr., Linear prediction of speech, Springer-Verlag, 1980.

[4] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, "A vector quantization approach to speaker recognition", AT&T Technical Journal, Vol. 66, pp. 14-26, 1987.

[5] E. McDeranmott and S. Katagiri, "LVQ-based shift-tolerant phoneme recognition", IEEE Transaction on Signal Processing, Vol. 39, No. 6, pp. 1398-1411, 1991.

[6] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition", Proceedings of the IEEE, vol. 77, No. 2, pp. 257-286, 1989.

[7] G. Rigoll, "Information theory-based supervised learning methods for self-organizing maps in combination with hidden markov modeling, IEEE, pp. 65-68, 1991.