음성기반 멀티모달 인터페이스 및 표준

홍기형 (성신여대)

<차 례>

<Abstract>

Speech Based Multimodal Interface Technologies and Standards

Ki-Hyung Hong

In this paper, we introduce the multimodal user interface technology, especially based on speech. We classify multimodal interface technologies into four classes: sequential, alternate, supplementary, and semantic multimodal interfaces. After introducing four types of multimodal interfaces, we explain standard activities currently being activated.

* Keywords : multimodal interface, standards, XML, VocieXML

1. 서 론

사용자인터페이스는 텔레비전, 오디오, 비디오, 세탁기, 청소기, 자동차와 같은 일상생활에 쉽게 접하는 수많은 기계 기구와 컴퓨터 하드웨어 및 소프트웨어에서 사용자의 편의를 도모하는 매우 중요한 구성요소이다. 실제로 사용자 인터페이스의 편리성이 제품의 마케팅이나 판매에 매우 큰 영향을 준다. 대부분의 기기에서 기기가 가지고 있는 기능의 사용에 있어 현재는 하드웨어로 구현된 버튼으로 구성되어 있다. 그러나 기기들이 가진 기능이 점점 복잡해지고 다양해짐에 따라 하드웨어로 구현된 버튼으로는 한계에 도달하고 있다.

사용자의 사용의 편리성과 자연스러운 사용자 인터페이스를 위하여 음성인식을 이용한 음성 인터페이스가 대안으로 등장하였다. 음성은 전화 단말이 컴퓨팅 능력을 전혀 갖지 못하는 유선 전화 환경에서 전화를 이용한 정보 시스템 접근을 위하여 사용되었으나, 최근에는 자동차와 같이 손과 눈을 이용한 인터페이스가 불가능한 상황에서 가장 편리하며, 필요한 인터페이스로 관련 기술 개발이 이루어지고 있다. 또한 한정된 공간에 기능을 지시하기 위한 버튼의 수는 제한적일 수밖에 없으므로, 음성인식 문법을 이용한 기능 제어를 위한 명령은 사용의 편리성 뿐 아니라 기기의 외관 및 크기의 결정에 자율성을 줄 수 있다.

2개 이상의 입.출력 모달리티를 동시에 사용하는 멀티모달 인터페이스는 최근 모바일 장비의 급속한 확장과 더불어 그 필요성이 크게 높아지고 있다[1][2][3]. 휴대전화나 PDA와 같은 개인 휴대 단말의 기능이 다양해짐에 따라, 사용자가 원하는 기능을 선택하고, 사용자와의 인터랙션이 많이 필요한 기능이 추가되고 있다. 예를 들면, 과거의 휴대전화가 단순히 상대방에게 전화를 걸고, 음성 대화 채널을 연결해주는 기능만을 가지고 있었다면, 최근에는 인터넷을 접속하여, 사용자가 필요한 정보를 검색하며, 게임, MP3 등 멀티미디어 데이터의 접속과 실행, 디지털 방송 수신기, LBS (Location based Service) 등의 기능이 추가되고 있다. 그러나 이러한 개인 휴대 단말은 휴대하기에 적합하여야 하므로 기본적으로 소형으로 한손에 쥐고 사용할 수 있어야한다. 따라서 개인 휴대 단말이 가질 수 있는 입력 장치는 소형의 키패드, 터치스크린, 마이크 정도이며, 일반적인 컴퓨터에서 사용할 수 있는 많은 키를 가진 키보드나 다양한 기능을 선택하기 위한 많은 수의 버튼을 갖추기 어렵다. 출력 장치의 경우에도 작은 크기의 LCD 스크린(현재, 지상파 DMB 수신 단말의 경우, 320x240의 해상도를 가진 7인치)과 소형 스피커 정도이다. 따라서 음성을 기반으로 하여 다른 모달리티를 결합한 인터페이스의 필요성이 점차 증가하고 있으며, 관련 기술의 개발도 활발해 지고 있다.

본 논문에서는 음성 기반 멀티모달 인터페이스의 기술 현황을 소개하고, 멀티모달 인터페이스를 크게 순차, 단순결합, 보조결합, 의미결합의 4가지로 분류하였다. 그리고, 현재 진행되고 있는 단순결합 음성기반 멀티모달 인터페이스 관련 표

준을 조사하였다.

본 논문의 구성은 다음과 같다. 먼저 2장에서 멀티모달 인터페이스의 정의와 장점을 기술하였고, 3장에서 멀티모달 인터페이스 기술을 모달리티 사이의 결합 방법에 따라 분류하였다. 4장에서는 음성 기반 멀티모달 인터페이스 관련 표준을 소개하고, 표준 중에서 XHTML+Voice에 대하여 상세히 기술하였다. 마지막으로 5장에서 국내 기술 개발 현황과 앞으로의 연구 과제를 기술하여 결론을 맺는다.

2. 멀티모달 인터페이스의 개요

2.1. 정의 및 장점

모달리티(modality)는 개체 사이의 의사 전달을 위한 채널로 정의한다. 여기서, 개체는 임의의 컴퓨팅 시스템, 기계기구, 사람을 의미한다. 시스템과 시스템 사이, 사람과 사람 사이, 시스템과 사람 사이에 존재하는 모달리티는 일반적으로 하나가 아니며, 다양한 형태로 다수의 모달리티가 존재한다. 사람의 경우, 시각, 청각, 촉각과 음성, 제스처와 같은 다수의 채널을 동시에 활용하여 서로 대화하는 것이 일반적이며 자연스럽다.

모달리티는 개별 개체의 입장에서 의사를 다른 개체에 전달할 때 사용할 수 있는 *출력 모달리티*와 다른 대체의 의사가 전달되는 *입력 모달리티*로 구분할 수 있다. 입력 모달리티는 사람의 경우에 시각, 청각, 촉각 등을 들 수 있으며, 출력 모달리티는 말소리, 제스처를 들 수 있다. 컴퓨터나 휴대 단말의 경우에는 키보드, 마우스, 터치 스크린, 마이크(음성), 카메라(비전)이 입력 모달리티에 해당하며, 디스플레이, 스피커, 햅틱(Haptic) 장비 등이 출력 모달리티라 할 수 있다.

단일모달 사용자인터페이스 (Uni-modal User Interface)는 개체가 가진 모달리티 중에서 하나만을 사용하여 상호 의사전달을 하는 인터페이스를 말한다. *멀티모달* 사용자인터페이스는 2개 이상의 모달리티를 사용하는 인터페이스를 말한다.

단일모달만을 사용한 커뮤니케이션에 비하여 다수의 모달리티를 활용하는 멀티모달 인터페이스는 다음과 같은 장점이 있다[1].

■ 모달리티 결합을 통한 시너지 효과

일반적으로 음성인식의 경우, 무소음 환경에서의 인식 성능은 경우에 따라서 95% 이상의 정확도를 보이지만, 도로나 공공장소와 같이 소음이 심한 경우에는 정확도가 50% 이하로 낮아지는 경우가 일반적이다. 이러한 경우, 음성과 함께 다른 모달리티, 예를 들면 비젼을 이용한 입술모양 인식을 함께 사용함으로써, 인식의 정확도를 향상시킬 수 있다.

■ 각 모달리티 별 장점 극대화

일반적으로 개별 모달리티는 각기 제 나름의 장점을 가지고 있다. 예를 들면, 특정 개체를 지칭하는 경우, 어떤 개체인지를 말로 설명하는 것 보다는 손이나 포인팅 장치를 이용하여 직접 어떤 개체인지 가리키는 것이 훨씬 효과적이다. 예를 들어, 탁자위에 있는 3개의 컵 중에서 빨간색 컵을 의미하고자 한다면, 포인팅 제스처는 그저 그 컵을 가리키는 것으로 충분하지만, 말로만 설명한다면, "탁자위에 있는 3개의 컵 중에서 빨간색 컵"이라고 설명하여야 한다. 만일, 탁자위에 빨간색 컵이 2개라면, 말로 하게 되면 더 복잡해질 것이다.

■ 사용자 선택의 자율성

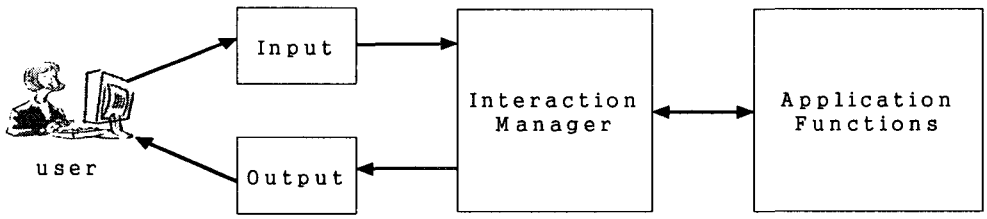
다수의 모달리티가 지원된다면, 사용자에게는 자신이 현재 처한 상황에 따라, 가능한 모달리티를 선택할 수 있다. 예를 들면, 운전 중에는 키보드나 포인팅 장치를 이용한 다는 것이 매우 위험하므로 음성 인터페이스의 사용을 매우 선호할 것이다. 청각장애인의 경우에는 GUI (Graphical User Interface)를 시각 장애인의 경우에는 음성 사용자 인터페이스, 즉 VUI (Voice User Interface)가 유용하므로 하나의 시스템에서 GUI와 VUI를 모두 제공한다면, 사용자의 환경에 무관하게 시스템의 기능을 사용할 수 있는 인터페이스 선택의 자율성이 보장 된다.

■ 커뮤니케이션의 자연성

일반적으로 사람이 일상생활에서 다른 사람과 상호 작용의 대부분이 멀티모달이다. 따라서, 가장 자연스러운 커뮤니케이션은 멀티모달이라고 할 수 있으며, 다양한 모달리티를 동시에 사용한 멀티모달 인터페이스의 지원은 시스템을 사용하는데 있어 사람이 가장 자연스럽게 시스템의 접근과 사용을 가능하게 한다.

2.2. 멀티모달 인터페이스 프레임워크

<그림 1>는 멀티모달 인터페이스 프레임워크[4][5]의 기본 구성요소를 보여주고 있다.



<그림 1> 멀티모달 인터페이스 프레임워크 구성요소

■ 사용자(user): 사용자는 시스템에 정보를 입력하거나 말하고, 시스템이 내놓은 정보를 듣거나 본다.

■ Input: 대화형 멀티모달 구현에서 오디오, 음성, 수시(handwriting), 키보드 등과 같은 다중 입력을 의미한다.

■ Output: 대화형 멀티모달 구현에서 음성, 문자, 그래픽, 오디오 파일, 애니메이션 같은 하나 혹은 그 이상의 출력 모드를 사용한다.

■ 대화 관리자(Interaction manager): 대화관리자는 사용자와 응용 함수 사이에서 정보 흐름을 관리한다. 다음과 같은 대화 방법을 지원한다.

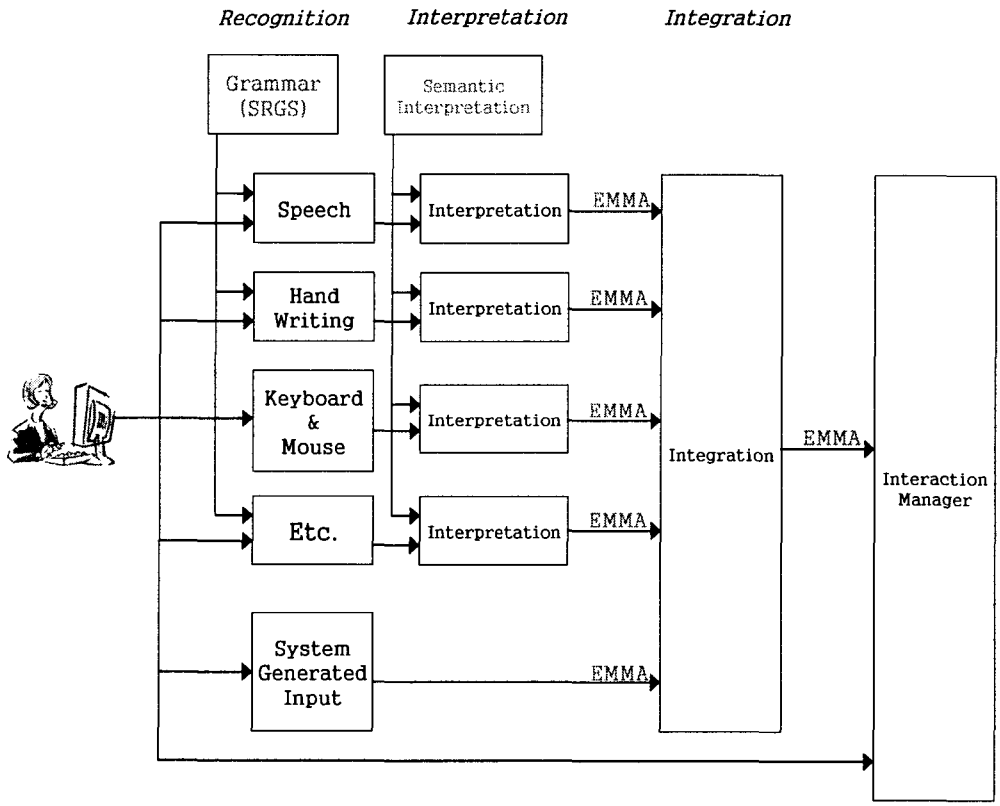
- 시스템 주도 대화 - 시스템은 질문하는 방식으로 유저에게 프롬프트하고, 사용자는 질문에 답하는 방식으로 응답한다.
- 사용자 주도 대화 - 사용자는 액션을 수행하도록 컴퓨터에게 지시하고, 컴퓨터는 행동의 결과를 내놓음으로서 유저에게 응답한다.
- 상호 주도 대화 - 경우에 따라 사용자와 시스템이 다이얼로그를 주도하는 것으로 시스템 주도와 사용자 주도 대화를 섞어 놓은 것이다.

대화 관리자는 사용자의 의도와 초점을 파악할 뿐 아니라, 대화 컨텍스트(context)와 어플리케이션의 상태 유지, 입력의 구조와 입력 모드사이의 동기, 비지니스 로직과 결합 등을 담당한다. 대화 관리자가 시스템에 따라서는 여러 시스템 구성 요소에 기능적으로 분산되어 존재할 수 있다.

■ 응용 시스템 기능(Application functions) - 데이터베이스 접근, 트랜잭션 프로세싱, 응용 종속 계산 등 여러 가지 응용 시스템을 위한 기능을 말한다. 응용 시스템과의 정보 교환은 응용 별로 정해진 형식에 따라 이루어진다.

2.2.1. 멀티모달 입력 시스템 구성 요소

<그림 2>은 멀티모달 입력 시스템의 구성 요소를 보이고 있다.



<그림 2> 멀티모달 입력 구성요소

■ 인식부(Recognition component) - 사용자로부터 자연스럽게 입력을 인지하고 다음 프로세싱을 위해 입력을 폼(form) 형태로 구성한다. 인식부는 사용자의 입력을 보다 정확히 파악하기 위하여 특정 시점에 사용자의 입력 패턴을 지정하는 문법(Grammar)를 사용한다. 대표적인 인식부의 예는 다음과 같다.

- 음성인식(Speech) - 말해진 음성을 텍스트로 바꾼다. SRGS (Speech Recognition Grammar Specification)은 W3C에서 권고하는 문법 기술을 위한 표준이다.
- 필기인식(Handwriting) - 사용자의 필기를 텍스트로 변화한다. 필기 인식의 구성요소는 handwritten gesture model, language model, 필기 인식 문법으로 구성된다.
- 키보드입력(Keyboarding) - 글자 key를 눌러서 입력한다.
- 포인팅 장치(Pointing device) - 2차원 표면의 x-y 위치를 기반으로 입력을 받는다.

또 다른 입력 인식부로는 시각, 기호 언어, DTMF, 생체, 촉각 입력, 화자 인증 등을 들 수 있다.

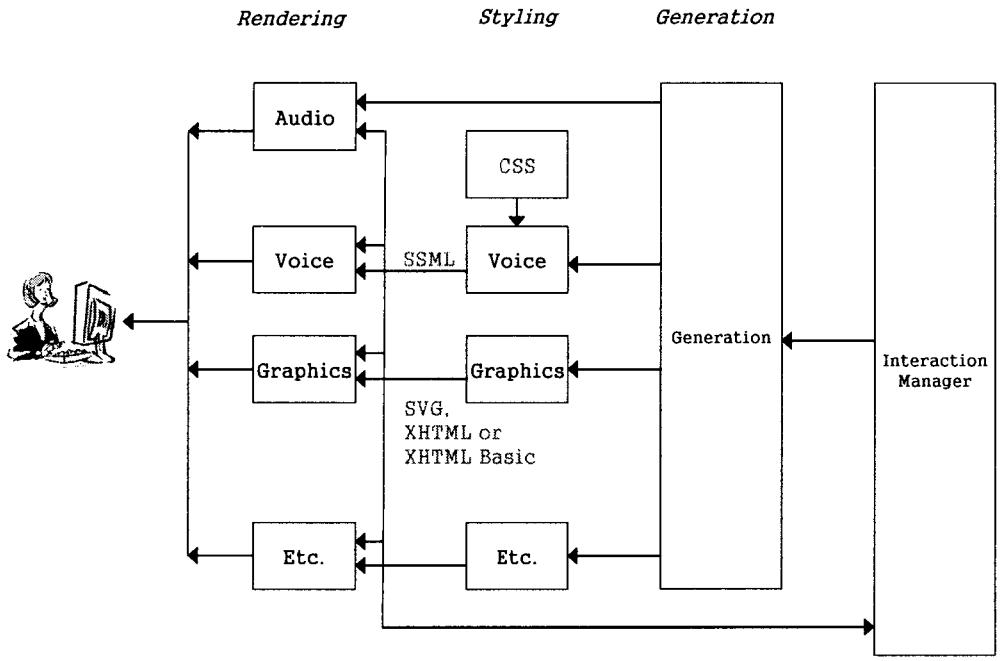
■ 해석부(Interpretation component) - 인식의 결과를 처리한다. 해석부는 사용자의 입력으로부터 의미(의도)를 식별한다. 예를 들면, "yes", "affirmative", "sure", "I agree" 같은 많은 단어들이 "yes"로 표현 될 수 있다.

■ 통합부(Integration component) - 다수의 해석부로 부터의 결과를 결합한다. 통합부의 일부 혹은 전체 기능은 인식부, 해석부 및 대화 관리자에서 실행될 수도 있다. 예를 들면, audio-visual 음성 인식은 입술움직임 인식과 음성 인식이 통합된 것이다. 다른 예로, 음성인식과 포인팅의 두 입력 모드가 "put that (point to an object), there (point to a location)"와 같이 사용된다. 다른 시스템에 의해 발생된 정보 역시, 대화관리자에 의해 사용자 입력과 통합될 수 있다. 예를 들면, 사용자의 현재 위치를 나타내는 GPS 시스템이 있다.

2.2.2. 멀티모달 출력 시스템 구성요소

<그림 3>은 출력 구성 요소를 보이고 있다.

■ 생성부(Generation component) - 생성부는 대화관리자에서 사용자에게 줄 정보를 어떤 출력 모드를 사용하여 전달할 것인지 결정한다. 단일 출력 모드를 선택할 수도 있고, 다중 출력모드나 보조 출력 보드를 선택할 수 있다.



<그림 3> 멀티모달 출력 구성 요소

■ 스타일부(Styling component) - 정보가 어떤 'lay-out'을 이용하여 출력될 것인지를 결정한다. 예를 들어 디스플레이를 위한 스타일부는 그래픽 객체(graphical objects)를 어떻게 캔버스에 위치시킬 것인가에 대해 기술한다. 오디오를 위한 스타일부는 묵음의 삽입, 음성합성 시의 억양제어 등이 삽입된다. HTML에서 사용하는 CSS (Cascading Style Sheet) 역시 음성 합성을 제어하는데 사용될 수 있다.

■ 렌더링부(Rendering component) - 렌더링부는 스타일부로부터 사용자가 쉽게 인지할 수 있는 포맷으로 정보를 변환 시킨다. 예를 들어, 그래픽렌더링부는 사각형을 점의 백터로 변환하고, 그리고 음성 합성 시스템은 텍스트를 합성된 음성으로 변환한다. 각 출력 모달리티는 고유의 스타일부와 렌더링부를 가진다.

■ 음성 스타일링부는 어떻게 단어가 발음되어야 하는지를 나타내는 SSML (Speech Synthesis Markup Language) 문서를 하나의 출력으로 할 수 있다. 이것은 음성 렌더링부(Text-To-Speech)에 의해 음성으로 전환된다. 또 음성 스타일부는 음성 렌더링부로 재생되기 위해 미리 녹음된 오디오 파일을 선택하기도 한다.

■ 그래픽 스타일부는 그래픽이 어떻게 전환되어져야 하는지 기술한 XHTML, XHTML Basic, SVG markup 등으로 표시할 수 있다. 그래픽 렌더링부(일반적인 웹브라우저)는 사용자에게 보여 지는 그래픽으로 이를 변환한다.

2.2.3. 멀티모달 시스템 사용 예

멀티모달 인터페이스 프레임워크의 구성요소를 보다 잘 이해하기 위하여 간단한 사례를 생각해보자. 사용자는 화면에 표시된 지도를 보고 포인팅 제스처와 음성으로 "이 곳(가리키며)의 지명은 무엇입니까?"라고 질문한다. 멀티모달 인터페이스 시스템은 결과로 "Wobegon 호수, 미네소타"라는 음성을 합성하여 들려주고, 동시에 "Wobegon 호수, 미네소타"라는 텍스트를 화면에 표시한다.

이 과정에서 이루어지는 각 구성요소 사이의 일련의 작업은 다음과 같이 이루어진다.

① 사용자 - 지도상의 위치를 찍고, "이곳의 지명은 무엇입니까?"라고 말한다.
② 음성 인식부 - "이곳의 지명은 무엇입니까?"라는 단어들을 인식한다.
③ 포인트 인식부 - 사용자가 지도상에 지정한 위치를 x-y 좌표 상에서 인식한다.
④ 음성 해석부 - "이곳의 지명은 무엇입니까?"라는 문장을 내부 표기법으로 전환시킨다.
⑤ 포인트 해석부 - 사용자가 지정한 위치를 내부 표기법 x-y좌표로 변환 한다.
⑥ 통합부 - x-y좌표와 함께 "이곳의 지명은 무엇입니까?"의 음성 해석부의 결과를 통합한다.
⑦ 대화관리자 - 사용자 요구를 데이터베이스 질의로 변환하고, 결과 "Wobegon 호수, 미네소타"라는 값을 DBMS로 부터 수신한다. 대화관리자 내부 자료구조로 결과 값을 변환하고, 생성부로 응답을 보낸다.
⑧ 생성부 - 음성과 그래픽 둘 모두 상호 보완하는 모드로, 결과를 제공하는 것을 결정한다. 생성부는 "Wobegon 호수, 미네소타"라는 것을 음성 스타일부로 보낸다. 그리고 그래픽 스타일부로 내부 자료구조가 나타내는 "Wobegon 호수, 미네소타"의 위치를 보낸다.
⑨ 음성 스타일부 - "Wobegon 호수, 미네소타"를 SSML로 변환한다.
⑩ 그래픽 스타일부 - "Wobegon 호수, 미네소타"를 표현하는 지도의 위치 정보를 HTML로 변환한다.
⑪ 음성 렌더링부 - 사용자가 들을 수 있도록 SSML을 음성으로 변환한다.
⑫ 그래픽 렌더링부 - 사용자가 볼 수 있도록 HTML을 비주얼 그래픽으로 변환한다.

3. 멀티모달 인터페이스의 분류

멀티모달 인터페이스 기술은 서로 다른 모달리티의 결합 정도에 따라 다음과 같이 분류할 수 있다.

- 순차 멀티모달 (Sequential)
- 단순결합 멀티모달 (Alternate, or Simultaneous)
- 보조결합 멀티모달 (Supplementary)
- 의미결합 멀티모달 (Semantic)

여기서, 순차와 단순결합 멀티모달은 시간적으로 서로 다른 모달리티를 선택적으로만 사용하고, 서로 다른 모달리티의 결과를 하나의 정보로 결합하는 것은 아니다. W3C에서의 분류[4]는 순차 멀티모달, 동시지원 멀티모달 (단순결합 멀티모달), 그리고 복합 멀티모달(Composite)로 나누고 있다. W3C에서의 복합 멀티모달은 보조결합 및 의미결합을 모두 통칭한다.

순차 멀티모달 (Sequential Multimodal)은 특정 시점에는 하나의 모달리티만을 사용하지만, 시간에 따라 다른 모달리티를 활용함으로써 개체가 가지고 있는 다수의 모달리티를 사용하는 인터페이스를 말한다. 현재 휴대전화에서 SM (Short Message) 수신 후, SM발신자와 통화를 하게 되는 형태를 들 수 있다. SMS (Short Message Service)를 이용하여, 발신자가 상대방에게 메시지를 전달한다. 이때 사용하게 되는 모달리티는 키패드(입력)이며, 수신자는 수신된 SMS를 키패드를 이용하여 관리할 수 있으며, 휴대전화의 스크린으로 수신된 SM를 확인할 수 있다. SM을 확인한 후 발신자와 바로 통화를 시도할 수 있는데, 통화를 시도하여 연결이 되면, 그 다음 부터는 마이크와 스피커를 통하여 음성 통화를 하게 된다. 이 예에서 보면, 사용자는 SM을 이용하고 난 후, 음성 통화를 하였다. 크게 보면, 키패드(스크린), 음성의 2 가지 모달리티를 시간에 따라 하나씩 연속적으로 사용한 것이다.

단순결합 멀티모달 (Alternate Multimodal)은 특정 시점에서는 하나의 모달리티만 사용되지만 동일한 기능을 위하여 다수의 모달리티가 제공되는 형태이다. 피자를 주문하는 웹사이트를 생각해보자. 집에서 PC를 사용할 수도 있고, 전화를 사용할 수도 있다. PC를 사용하는 사용자의 경우, 피자를 주문하기위하여 웹에 접속하여, HTML로 구성된 웹페이지를 보면서, 키보드와 마우스의 조작으로 원하는 피자를 주문할 수 있다. 또, 전화만을 사용하여, 주문를 받는 피자 전화번호로 전화를 걸어, 음성으로 피자를 주문할 수 있다. 이러한 예에서 보면, 피자 가게에서는 전화로든 인터넷으로든 피자를 주문 받는 것은 동일하다. 음성(전화)로 주문을 받

은 경우이든, 웹으로 주문을 받은 경우이든 피자 가게에서는 동일한 기능을 수행하는 것이 된다. 4장에서 기술할 X+V와 SALT의 경우처럼, 웹페이지에서 GUI와 음성 인터페이스를 동시에 제공하는 경우도 여기에 해당한다. GUI, 즉 화면을 보고 키보드나 마우스를 이용하여 피자를 주문할 수도 있고, 음성으로 피자의 종류를 듣고 원하는 피자와 수량을 말함으로써 주문할 수도 있다. 이때, 음성으로 입력한 것이 바로 GUI에서 화면상에 표시되게 할 수 있다. 여기서, 단순결합이란 하나의 모달리티의 결과가 다른 모달리티에 그대로 동기시키면 되는 형태이기 때문이다. 각각의 모달리티에서 입출력되는 정보의 양이 동일하다.

보조결합 멀티모달 (Supplementary Multimodal) : 여기서부터는 다수의 모달리티가 동시에 사용되어 서로 다른 정보를 나타내고, 이들이 결합되어 하나의 의도를 파악하게 하는 것이다. 보조결합은 하나의 모달리티를 통한 정보의 전달은 문자로 표현할 수 있는 형태의 정보이고 다른 모달리티는 감정이나 표정, 또는 소리를 냄에 따라 움직이게 되는 입 주의의 근육의 동작 등과 같이 일반적으로 문자화하기 어렵거나, 동일한 정보를 표현하지만 다른 모달리티의 결과를 의미한다. 예를 들면, 음성인식과 함께 입술모양의 인식을 병행하는 오디오-비주얼 인식을 들 수 있다. 출력의 예로는 감정과 함께하는 음성 합성을 들 수 있다.

의미결합 멀티모달 (Semantic Multimodal) : 일반적으로 음성으로는 액션을 기타 제스쳐로는 해당 액션에 필요한 인자를 받아들여, 이 2 가지 모달리티의 결과가 의미있게 결합되어 하나의 정보를 표현하는 경우를 말한다. 예를 들면, 터치스크린 상에 나타난 지도상의 특정 위치를 가리키면서, 음성으로 '여기서 가장 가까운 주유소는?'이라고 말하는 예를 들 수 있다. 음성 인식을 통하여 '여기서'라는 단어를 인식하였다고 하더라도 시스템 입장에서는 '여기'가 어디 인지 명확하지 않다. 그러나 터치스크린을 이용하여 가리킨 곳이 입력됨으로써 '여기'는 '명동역'과 같이 구체화될 수 있다. 음성과 터치스크린 상의 제스쳐 인식을 동시에 사용하여, '명동역에서 가장 가까운 주유소는?'이라는 완전한 의미가 전달되는 것이다.

<표 1>에는 여기서 분류한 각 멀티모달 인터페이스의 예를 보이고 있다. 또한 현재의 연구개발 수준을 표시하였다.

<표 1> 멀티모달 인터페이스 분류 별 예제

멀티모달 분류		입 력	출 력	연구개발 수준
순 차		휴대전화(키패드->마이크)	휴대전화(액정-->스피커)	상용화
결 합	단순	X+V, SALT	X+V, SALT	시제품
	보조	Audio-visual recognition	감정 TTS	연 구
	의미	음성은 행위, 포인팅 제스처는 행위에 필요한 인자 ("이것을 들려줘" [음악파일 포인트])	프로그램 내레이션 (화면에 그림, 소리로 설명)	초기연구 (단, 출력은 SMIL 등, 상용화단계)

4. 음성기반 멀티모달 인터페이스 표준

4.1. 표준화 활동

현재 존재하는 멀티모달 인터페이스 표준 및 표준화 활동은 <표 1>에서 보인 바와 같이 상용화 직전단계에 이른 단순결합 멀티모달 인터페이스를 위한 표준이며, 보조결합이나 의미결합을 위한 표준은 초기 연구 단계에 있다. 멀티모달 인터페이스를 개발하기 위한 표준을 제정하기 위한 활동은 크게 3가지가 있다.

4.1.1. SALT (Speech Application Language Tags) [6]

마이크로 소프트 사가 주도하는 SALT는 HTML과 다른 마크업 언어(XHTML, WML)의 확장으로, 음성만을 위한 브라우저 (VoiceXML[7]과 동일한 목적)와 멀티모달(Multimodal) 브라우저의 2 가지 목표를 달성하기 위한 명세이다. SALT는 기본적으로 PC용 비주얼 웹 브라우저를 위한 HTML 또는 XHTML이나, 휴대전화 또는 PDA용 웹 브라우저를 위한 WML(Wireless Markup Language)에 내포될 수 있도록 설계되었다. 따라서 SALT는 비주얼 페이지를 통해 음성 입출력을 동시 지원할 수 있는 멀티모달 인터페이스를 기술할 수 있는 언어 명세이다.

SALT는 사용자가 여러 가지 방식으로 정보 시스템과 상호 작용을 할 수 있도록 한다. 음성, 키보드, 키패드, 마우스 등을 이용해서 입력을 할 수 있고, 합성 음성, 오디오, 텍스트, 비디오, 그래픽 등과 같은 데이터를 산출할 수 있다. 또한, 비주얼 디스플레이가 없는 경우, SALT는 HTML 이벤트 모델과 스크립팅 모델을 사용하여 다이얼로그의 상호 작용 흐름을 관리하도록 하고 있다. 음성인식 문법은 W3C의 SRGS를 그대로 사용한다.

4.1.2. X+V (XHTML + Voice) [8]

IBM이 주도하는 콘소시움[9]에 의하여, 개발되고 있는 X+V는 SALT와 매우 유사한 형태 및 목적을 가지고 있다. SALT와의 차이는 SALT에서는 음성과 관련한 마크업을 VoiceXML[7]과는 별도로 개발하였으나, X+V에서는 VocieXML을 그대로 사용한다.

이름에서도 나타나듯이, X+V는 XHTML을 호스트 언어로 하여, VoiceXML을 내포 언어로 채용한 것이다. 또한, XML의 이벤트 모델을 그대로 사용할 수 있도록 하였다. X+V와 SALT는 다음에 설명하는 W3C의 멀티모달 인터페이스 표준에 채택되기 위하여 제출된 상태로, 마이크로소프트와 IBM 진영이 경쟁하는 모양을 보이고 있다.

4.1.3. W3C 멀티모달 인터페이스 표준화 활동 [4]

음성과 제스처로 접근할 수 있는 웹 페이지를 모토로 하여, W3C는 인터페이스의 다양한 모드를 지원하는 이동형 장치에서 다양한 인터페이스 모드를 지원하기 위한 표준의 개발을 목표로 표준화 활동을 진행하고 있다.

데스크 탑 PC 시스템은 웹에 접근하기 위해 매우 효과적인 것이 증명되었다. 고해상도 화면, 포인팅 장치와 키보드는 많은 정보와 효율적으로 상호 작용하는 것을 쉽게 한다. 그러나 이동 중인 경우에, 포켓 또는 지갑에 꼭 맞는 작은 경량의 장치를 필요로 한다. 휴대 전화는 매우 대중화된 장치이지만 표시할 수 있는 정보 양이 제한적인 표시 장치, 작은 수의 키 등의 제약이 있다. XHTML, CSS, SMIL 과 SVG 등의 W3C 권고안에서 이동 장치를 위한 모바일 프로파일이 포함되고 있으며, 이동 중에 웹의 접근은 현실화 되었다. 그러나 이동 장치의 작은 키패드만으로는 수 천 개의 문자, 표의언어를 위한 검색, 또는 웹 주소를 입력하는 것이 매우 힘들다.

최근 몇 년 사이에 웹을 전화로 접근하기 위한 수단으로 음성을 사용하는 것에 대한 관심이 집중 되었다. 그 결과로 W3C는 음성 인터페이스 프레임워크를 규정하고, 이에 필요한 표준안을 개발하였다. VoiceXML 기반의 음성 인터페이스는 미리 녹음한 음성이나, 합성음을 이용하며, 단어와 간단한 구를 인식할 수 있는 단계에 까지 이르렀으며, 기술의 발전에 따라, 보다 자연스러운 대화체 인식 및 합성이 가능해 질 것으로 기대되고 있다. 이에 따라, 음성 인터페이스를 다른 다양한 인터페이스 모드 (특히 이동 가능한 휴대장치의 제한적인 입력 시스템 또는 펜, 터치스크린 등)와 결합하고자 하는 연구가 촉진되고 있다. 음성과 다른 인터페이스 모드의 결합으로 나타나는 멀티모달 인터페이스는 사용자에게 말하고, 듣고, 쓰고, 타이핑하고, 보는 보다 다양하고 자연스러운 시스템과의 상호 작용을 가능

하게 할 것이다.

현재 개발 중인 주요한 멀티모달 인터페이스 표준은 다양한 다른 표준화 활동과 긴밀한 협력으로 공동 작업이 이루어질 예정이다. 주요한 표준화 항목은 다음과 같다.

■ Multimodal Interaction Framework : 멀티모달 인터페이스 적용 시스템의 일반적인 구조와 구성 요소 및 사용할 수 있는 표준 및 마크업 언어와 관계를 명기하고 있다. (그림 2, 3 참고) 초안에는 입/출력 구성요소, 상호 장용 관리와 보조 구성요소에 초점을 맞출 예정이며, 향후, 객체지향 모델기반의 마크업 개발과 다른 W3C 마크업 언어와의 통합 방법이 포함될 예정이다.

■ Extensible Multimodal Annotation Markup Language (EMMA) : 입력 장치들과 멀티모달 상호작용 관리 시스템 사이의 인터페이스를 위한 데이터 형식을 표준화하기 위한 활동으로 다음과 같은 계획 하에 진행되고 있다.

- Requirements - 2003년 1월 13일
- First Working Draft - 2003년 12월 18일
- Last Call Working Draft - 2005년 상반기 예정

EMMA에는 인식기가 응용 별 특징 데이터와 부가적인 인식 스코어, 시간, 입력 모드, 그리고 부분 인식 결과 등을 표기할 수 있는 방법을 정의할 예정이다. EMMA는 음성 브라우저 프레임워크 안에서 개발되는 의미 해석 (semantic interpretation specification) 표준의 데이터 형식이기도 하다.

■ 펜 입력 : 여기서는 멀티모달 시스템의 전자펜이나 스타일러스에서 사용되는 잉크를 위한 XML 형식을 정의한다. 필기체 인식, 제스처 및 그림 인식과 수학, 음악, 화학 등에서 사용되는 특수 기호의 인식을 위한 것이다. IBM, Intel, Motorola, 그리고 International Unipen Foundation에서 진행하여 온 연구를 바탕으로 출발하였으며, 다음과 같은 계획을 가지고 있다.

- Requirements- 2003년 1월22일
- First Working Draft - 2004년 2월 23일
- Last Call Working Draft - 2005년 예정

4.2. XHTML+Voice

XHTML+Voice[8]는 Opera Software, IBM, Motorola 등이 멀티모달 인터페이스 표준화를 위해 지난 2001년 월드와이드웹 컨소시엄(W3C)에 제출한 웹 표준으로,

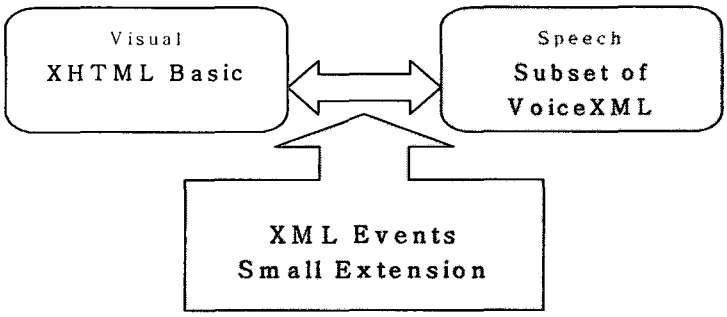
2004년 3월 16일 Profile 1.2 버전이 나온 상태이다. XHTML+Voice는 XML 기반의 마크업(Markup) 언어 두 가지를 통합하여 사용한다. HTML을 XML로 정형화한 언어인 XHTML과 음성 어플리케이션 개발자용 XML 프레임웍인 VoiceXML이 바로 그것이다.

4.2.1. 설계 원칙

XHTML+Voice의 큰 장점은 기존에 있던 표준을 재사용한다는데 있다. XHTML+Voice에서는 XHTML을 주 언어로 사용하고 VoiceXML 2.0의 일부를 차용했다. XHTML+Voice에서 재사용하는 기존 표준은 다음과 같다.

- PDA나 스마트 폰 같은 작은 디바이스에 적합한 요소만을 채택한 XHTML Basic[10]을 재사용한다.
- VoiceXML 2.0[7]에서 음성 인터페이스 처리를 위한 엘리먼트인 <form>태그만을 채택하고 <menu>태그 부분은 제외하여 재사용했다.
- noinput, nomatch, help등 VoiceXML 이벤트 타입과 함께 XML 이벤트[11]들 역시 XHTML+Voice에서 재사용된다.
- 그 외에 비주얼과 음성 모듈들 사이의 동기화를 위해 최소의 확장만이 이루어졌다.

<그림 4>는 GUI를 담당하는 XHTML과 VUI를 담당하는 VoiceXML, 그리고 이들 사이의 동기화를 담당하는 XML 이벤트와 XHTML+Voice만의 확장 모듈의 역할을 그림으로 보이고 있다.



<그림 4> XHTML+Voice 구성요소

4.2.2. XHTML+Voice 문서 구조

XHTML+Voice는 HTML문서와 매우 유사하다. 단 고유의 몇 가지 조건이 있는데 기본적으로 지켜야 할 사항은 다음과 같다.

첫째, XHTML+Voice 문서는 XML 스키마에 준하여 작성되어야 한다.

둘째, XHTML+Voice 문서의 루트 엘리먼트는 반드시 <html>이여야 한다.

셋째, 루트 엘리먼트에서 디폴트 네임스페이스(xmlns)는 XHTML 네임스페이스(http://www.w3.org/1999/xhtml)가 되어야 한다.

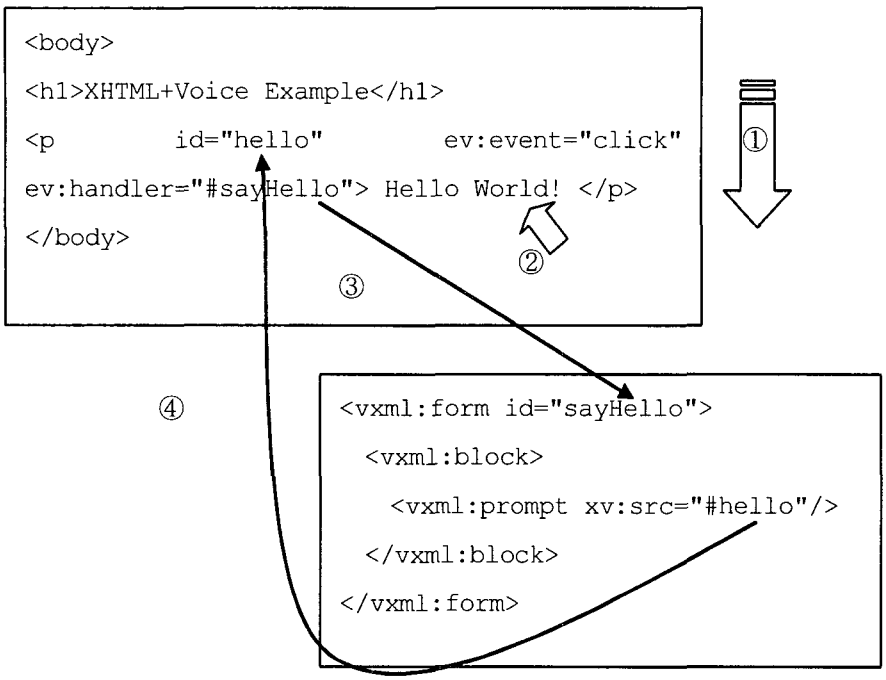
<그림 5>는 XHTML+Voice 문서의 실제 예이다.

```
<?xml version="1.0"?>
<html xmlns=http://www.w3.org/1999/xhtml
  xmlns:vxml=http://www.w3.org/2001/vxml
  xmlns:ev=http://www.w3.org/2001/xml-events
  xmlns:xv="http://www.voicexml.org/2002/xhtml+voice">
<head>
<!-- voice handler -->
<vxml:form id="sayHello">
  <vxml:block>
    <vxml:prompt xv:src="#hello"/>
  </vxml:block>
</vxml:form>
</head>
<body> <h1>XHTML+Voice Example</h1> <p id="hello"
  ev:event="click" ev:handler="#sayHello">
  Hello World! </p>
</body>
</html>
```

네임스페이스 / 음성인터페이스 / 그래픽인터페이스

<그림 5> XHTML+Voice 문서의 예

XHTML+Voice 문서는 크게 네임 스페이스, 음성 인터페이스, 그래픽 인터페이스 세 부분으로 나뉘어 진다. <그림 5>의 문서에서 <html> 태그에 속한 네임스페이스 정의 부분에는 XHTML, VoiceXML, XML Events, XHTML+Voice extension 등 네 가지의 네임스페이스가 있다. <head>내에 있는 VoiceXML form들과 XHTML+Voice extension(sync, cancel 등)을 기술하는 부분이 음성 인터페이스를 담당하는 부분이다. 마지막으로 XML 이벤트를 포함하는 GUI에 관련된 XHTML form들을 기술하는 부분은 <body>태그에 표시된다.



<그림 6> XHTML+Vocie 수행 순서

4.2.3. XTHML+Vocie 처리 모델

XHTML+Voice 처리 모델의 이해를 돕기 위해 <그림 5> 문서의 수행과정을 살펴보도록 하자. (<그림 6> 참고)

① 먼저 브라우저가 XHTML+Voice 문서를 load한다. 브라우저의 화면에 <body> 부분이 표시된다.

② 사용자가 "Hello World!"를 클릭한다. 이때 XML 이벤트인 'click'이 발생하게 된다.

③ "Hello World!"를 내용으로 하는 <p>가 event listener로 event handler(id값 sayHello)를 호출한다.

④ 'sayHello'를 id로 가진 VoiceXML form이 수행되면서 "Hello World!"를 합성하여 프롬프트한다.

간단하게 예제 문서로 XHTML+Voice의 수행 모델을 알아보았다.

XHTML+Vocie의 수행모델의 특징은 다음과 같다.

- 음성 인터페이스는 XML event에 의하여 활성된다.
- XHTML+Voice에서는 한번에 하나의 음성 대화만이 유효하다
- XHTML 부분(GUI)과 음성 부분(VUI)의 동기화는 전역 JavaScript를 이용하거나, <sync> (XHTML+Voice의 확장 모듈)를 통하여 이루어진다.

5. 결 론

지금까지 멀티모달 인터페이스 기술의 정의와 프레임워크를 설명하였고, 순차 멀티모달, 단순결합 멀티모달, 보조결합 멀티모달, 의미결합 멀티모달로 기술을 분류하였다. 그리고 현재 진행 중인 멀티모달 인터페이스 관련 표준 (X+V, SALT, W3C)및 표준화 활동을 조사하여 기술하였으며, 이중에서 X+V 표준에 대하여 보다 상세히 알아보았다.

유비쿼토스 컴퓨팅, 텔레매틱스, DMB 등 이동형 멀티미디어 단말[3]과 로봇의 위한 인터페이스[12]가 멀티모달 인터페이스이어야 함은 명확하다. 특히, 음성을 기반으로 다른 모달리티의 결합이 일반적이며, 국제적인 연구 개발이 이러한 방향으로 진행되고 있다. 그러나 국내에서는 현재, 보조결합 멀티모달에 해당하는 Audio-visual 분야에서의 일부 연구가 진행되고 있으나, 의미결합 멀티모달에 관련한 연구 개발은 매우 미미한 상태이다.

향후, 음성을 기반으로 한 의미결합 멀티모달 인터페이스 기술은 먼저, 음성과 포인팅 제스처를 결합하는 기술이 단기적으로는 가장 먼저 개발될 것으로 기대된다. 내비게이션[13]이나 LBS와 같은 특정한 영역에서는 일부 시제품이 나오고 있다. 장기로는 음성과 비젼, 감정, 촉각 등의 다양한 모달리티의 의미결합 멀티모달 기술의 개발이 필요하다.

참고문헌

[1] C. Benoit, J. C. Martin et al., "Audio-visual and Multimodal Speech-based Systems," in *Handbook of Multimodal and Spoken Dialogue Systems*, Dafydd Gibbon, Inge Mertinology and Roger K. Moore, Eds., Kluwer Academic Publishers, 2000.

[2] Sharon Oviatt, "Multimodal system processing in mobile environments," Proceedings of the 13th annual ACM symposium on User interface software and technology, November 2000.

[3] Wolfgang Mueller, Robbie Schaefer, Steffen Bleul, "Interactive Multimodal User Interfaces for Mobile Devices," Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04), January 2004.

[4] Multimodal Interaction Activity, http://www.w3.org/2002/mmi, Feb. 2004.

[5] Christian Elting, Stefan Rapp et al., "Multimodal architectures and frameworks: Architecture and implementation of multimodal plug and play," Proceedings of the 5th international conference on Multimodal interfaces, November 2003.

[6] Speech Application Language Tags(SALT) 1.0 Specification, http://www.saltforum.org/saltforum /downloads/SALT1.0.pdf, July, 2002.

[7] Voice Extensible Markup Language (VoiceXML) Version 2.0, http://www.w3.org/TR/2003/ CR-voicexml20-20030128, W3C Candidate Recommendation, January, 2003.

[8] XHTML+Vocie Profile 1.2, http://www.voicexml.org/specs/multimodal/x+v/12/spec.html. 2004.

[9] IBM Multimodal Development Page, http://www-306.ibm.com/software/pervasive/multimodal.

[10] XHTML Basic, http://www.w3.org/TR/xhtml-basic.

[11] XML Events: An Events Syntax for XML, http://www.w3.org/TR/xml-events.

[12] Sebastian Lang, Marcus Kleinehagenbrock, Sascha Hohenner, Jannik Fritsch, Gernot A. Fink, and Gerhard Sagerer, "Providing the Basis for Human-robot-Interaction: A Multi-Modal Attention System for a Mobile Robot," Proceedings of the 5th international conference on Multimodal interfaces, November 2003.

[13] Philip Cohen, David McGee, Josh Clow, "The efficiency of multimodal interaction for a map-based task," Proceedings of the sixth conference on Applied natural language processing, April 2000.

접수일자 : 2004년 8월 20일
게재결정 : 2004년 9월 9일

▶ 홍기형 (Ki-Hyung Hong)
주소: 서울 성북구 동선동 3가 249-1
소속: 성신여자대학교 미디어정보학부
전화: 02-920-7525
Fax: 02-929-7525
E-mail: khhong@sungshin.ac.kr