특 집

멀티미디어환경에서 멀티모달 휴먼인터페이스 기술

李 憲 柱
LG電子 技術院

石 玟 秀
LG電子 技術院

I. 개 요

도래하고 있는 정보사회에서 다양한 멀티미디어 환경을 서비스해야 하는 시스템들은 그래픽 사용자 인터페이스(Graphic User Interface) 뿐만 아니라 정보의 입출력 방식을 보다 더욱 다양화 시킬 수 있는 소리, 동작, 동화상, 에니메이션 등을 포함한 다양한 매체를 이용하여 인간의 총체적 잠제성능을 최적화할 수 있고 만족시킬 수 있도록 시스템을 상호 통합하는 방향으로 바뀌어 가고 있다. 인간들은 다양한 모드로 상황을 폭넓게 파악함으로써 세밀한 인식이 가능한데 멀티모달(multimodal)은 이에 가장 가까운 시도라고 말하고 있다. 이러한 이유로 휴먼인터페이스 분야에서 멀티모달이라는 기술이 주목되고 있다. 음성, 동작, 얼굴의 표정 등 다양한 형태를 구사하여 기계로의 정보 입출력을 자연에 가까운 조건으로 실현하고자 하는 시도인데 멀티미디어 시대에는 기계와 인간의 사이에 정보의 교환을 원활하게 하는 새로운 인터페이스가 필요하게 되는 바 멀티모달은 그 유력한 후보중 하나라고 할 수 있다. 이러한 멀티모달 인터페이스는 여러 가지의 입출력 수단을 가지게 되며, 각각으로부터 입력 정보를 얻고, 이를 통합하여 사용자의 의도를 인식하게 된다.

결국 기계는 사용자의 의도에 따라 동작하게 되고, 선택된 출력 수단을 통해 출력된다. 멀티모달이 주목되고 있는 배경으로는 키보드에 대신하는 입력 수단으로 보여지고 있는 음성입력의 인식수준이 아직 불충분하다는 사정도 있는데 예를 들어 〈한개〉라는 언어가 잡음 등으로 잘 들리지 않아도 손가락을 하나 세운 모습의 의미를 기계가 읽을 수만 있다면 인식도는 크게 높아지게 된다. 그러나 멀티모달 시스템은 유니모달 시스템과 달리 자연스러우면서 안정된 성능을 갖는 시스템을 설계하는 것이 단순한 직관에 의해서는 불가능하다. 따라서 유니모달 시스템보다 우수한 성능의 멀티모달 시스템을 설계하기 위해서는 주어진 환경에 따라 인간이 어떻게 modality를 선택하고 통합하는지에 대한 심도깊은 연구가 요구된다.

본고에서는 현재 세계적으로 주목 받고 있는 멀티모달 사용자 인터페이스에 대한 기술에 대해 최근 연구되고 있는 응용 시스팀을 중심으로 설명하고 향후 개발 방향에 대해 전망하고자 한다.

II. 차세대 휴먼인터페이스 기술

기계와 인간의 상호작용에 있어 인간은 외부로부터 특정 에너지 형태의 자극을 감지할 수 있는 5개의 수용 기관을 갖고 있고, 이들 시각(눈), 청각(귀), 촉각(피부), 후각(코), 그리고 미각(혀) 등의 자극을 관련 감각기관을 이용하여 정보를 감지한다. 이들 중 멀티미디어 시스팀과 인터페이스하는 작용에는 시각과 청각 그리고 촉각이 주로 유용하게 사용된다. 현재의 사용자 인터페이스는 입력을 위해서는 마우스를 출력을 위해서는 여전히 프린터를 사용하고 있다. 그러나 장래에는 멀티미디어 출력과 펜이나 동작, 음성 입력 등 다양한 형태로 나타나게 될 것이다. 최근까지 GUI는 컴퓨터에 접근하기 쉬운 형태로 디자인되고 실현되어 왔다. 예를 들면 마우스와 윈도우는 컴퓨터에 보다 쉽게 적응토록 하여 사용자 스스로 만족을 느끼도록 설계되었던 개념들이며, 사용자는 그것들의 제약사항을 스스로 인정하면서 사용하여 왔다. 그러나 멀티모달 인터페이스는 연속된 다양한 모드를 사용할 수 있으며, 이를 통해 보다 자연스러운 사용자 인터페이스를 제공해줄 수가 있다. 본 장에서는 향후 이러한 멀티모달 인터페이스 시스팀 구현을 위해 필수적인 음성/사운드, 얼굴 표정, 동작 인터페이스 기술에 대해 우선 서술하고자 한다.

1. 음성/사운드 인터페이스(Speech/Audio Interface)

휴먼인터페이스 분야에서 다양한 기술들이 괄목하게 발전하고 있지만, 그동안에는 주로 시각 인터페이스(visual interface) 분야에 집중되었고, 상대적으로 음성/사운드 인터페이스 분야는 발전 속도가 느렸다. 그 첫째 이유가 천공카드와 문자 인테페이스를 위해 개발된 CRT가 그래픽 인터페이스로 발전이 자연스럽게 진행되었고, 둘째가 청각 인터페이스로서 음성이 기술적으로 정복하기가 어려운 매체였기 때문이다. 현재 기계의 음성 인식/이해 능력이 사람에 비해 훨씬 모자라지만, 사용자들이 지금 수준의 장점만이라도 이용하기 위해 음성인식 인터페이스를 점차 채택하는 경향이 있다. 그리고 인식과는 반대로 기계가 인간이 알아들을 수 있는 음성으로 말할 수 있게 하는 음성합성 기술은 현재 많은 성공적인 시스팀들이 발표되고 있다. 본 절에서는 인간과 기계와의 인터페이스에서 음성과 사운드 인터페이스 기술에 대해 기술하고자 한다.[1]

1) 음성인식

음성인식이란 음성속에 내재되어 있는 언어 정보를 자동으로 추출하는 과정이다. 음성인식에 대한 연구는 사람의 음성을 인식할 수 있는 지능을 가진 로보트나 타자기를 개발할 목적으로 수십년 전부터 수행되었다. 그러나 오랫동안의 연구에 비해 최근에서야 비로서 인간과 기계사이의 음성 통신이 부분적으로 가능하게 되었다. 현재의 인식 수준은 인간의 능력에 비해 보잘 것이 없는데, 그 이유중의 하나는 아직 인간이 음성을 인식하는 정확한 방법을 모른다는 것이다. 이러한 어려움에도 불구하고 각나라의 학계와 업계에서는 다양한 방법으로 독자적인 인식 방법들을 연구하고, 음성인식 기능을 기계에다 부여하려는 이유는 무엇인가? 그 이유는 음성인식이 다음과 같은 다양한 장점들에 기인한 것으로 생각된다.

- 자연스러움(Naturalness) : 음성은 인간의 가장 자연스러운 통신수단으로 음성으로 기계에다 명령을 입력하는 것이 매우 쉽다.
- 신속성(Speed) : 음성은 글이나 자판 입력 보다 훨씬 빠르게 입력할 수 있다.
- 동시성(Simultaneity) : 인간들은 말을 하면서도 동시에 눈으로 보거나, 몸동작을 중단없이 행할 수 있다.
- 경제성 : 원격지에서도 전화 등을 통해 음성으로 명령을 내리는 것이 가능하므로 여행 경비 등을 절감할 수 있다.

음성인식은 크게 단어단위로 인식하는 고립단어

인식과 연속으로 발음한 문장을 인식하기 위한 연속음성 인식 기술로 나누어질 수 있다. 연속음성 인식은 connected word 인식과 대화체 음성인식으로 세분화될 수 있다. 전자는 또박 또박 발음하는 단어를 인식하는 것이고 후자는 대용량 단어를 단어단위가 아니라 문장의 의미를 파악하는 것으로 기술적으로 상당히 구현이 어렵다. 따라서 최근에는 문장중에서 선별적으로 지정된 단어만을 선택적으로 인식할 수 있는 word spotting에 대한 연구가 활발히 진행되고 있다.[2]

그리고 음성인식은 훈련된 특정화자만을 인식할 수 있는 화자종속인식과 어떠한 사람의 음성도 인식할 수 있는 화자독립 인식으로 대별되기도 한다. 일반적으로 후자가 전자보다 구현히 훨씬 어려우나 이용범위가 넓기 때문에 많이 연구 개발되고 있다.

위와 같은 음성인식에서 훈련과정과 인식과정에 사용하는 기술로서 dynamic time warping, HMM (hidden marcov model), 그리고 neural network 등이 있다. 이중에서 1970년대부터 많이 사용되고 있는 HMM 알고리즘이 빠른 인식 시간과 높은 인식률 때문에 인기가 있다. 최근 neural network도 기존의 방법들과 병용하여 인식하려는 연구가 진행되고 있다. 다음 그림은 phonetic unit을 사용한 대어휘 음성인식 시스템의 개략도이다.

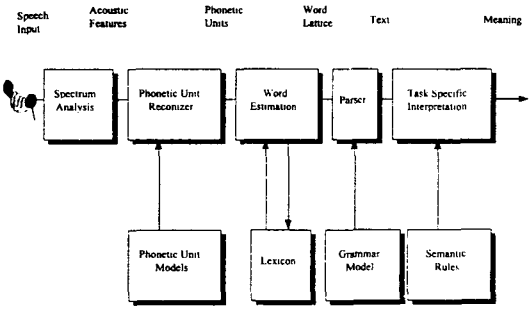
결국 연구와 활용은 별개의 것으로 현재의 수준에서 사용자들이 편리함을 느낄 수 있는 방향으로 기계와의 인터페이스에 적용해야 할 것이다.

2) 화자검증

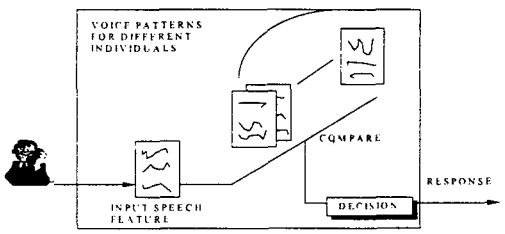
인간은 자기가 이전에 알고 있던 사람에 대해서는 목소리만 몇 마디 들으면 누구인지를 알 수 있다. 최근 개인정보나 proprietary databases와 같은 접근이 제한된 데이타베이스의 접근이나, 계좌이체, 전화 신용 결재 등과 같은 본인인지 아닌지 (personal identity)의 진위를 확인하는 거래에 대한 요구가 증가하고 있다. 이 경우 접근이나 거래를 허용하기 전에 개인음성의 독특한 특성을 이용하여 사용자의 신분 확인을 하는데 보조적으로 사용할 수 있으며, 이러한 기술을 화자 검증(talker verification)이라 부른다.

사용자들은 자신의 특정한 음성 샘플을 발성하여 시스템에 우선 등록한다. 그 다음 한 사용자가 제한된 DB에 접근을 시도하면, 기계는 본인 확인을 위해 우선 음성으로 본인 이름이나 계좌번호 입력을 요구하고, 그 다음 등록된 특정 문장과 같은 문장을 발음하길 요구하여 DB 접근 여부를 최종적으로 판단한다. 아래 그림은 사용자의 화자검증 과정을 보여주는 전형적인 시스템이다. 거래의 중요도에 따라 검증의 엄격 정도가 조절 가능하며 현재 수준에서 약 90%의 검증 정밀도의 성능을 보여주고 있다.[3,4] 음성인식은 문장을 입력한 사람이 누구인지에 관심이 있는 것이 아니고 무엇을 말했는지를 알아 내는 것이다. 이와 반대로 화자검증은 말한 내용에 관계없이 누가 말했는지를 알아내는데 그 목적이 있다. 이러한 화자검증을 위하여 보통 개인의 음성 특징을 잘 표현할 수 있는 cepstral coefficients 파라메타을 사용하고 있다.

자동화자검증 인터페이스는 transaction processing, 신용 결재, 계좌 이체 등에 보조 보안 수단으로 바람직하다. 기술이 계속 발전 하지만 아직 구현에 있어 많은 계산이 필요하여 이에 따른 비용



〈그림 1〉 대어휘 음성인식 시스템 개략도



〈그림 2〉 화자검증 시스템구조

증가로 기계에 채택은 활성화 되지 못하고 있다. 그러나 검증은 인식에서와 유사한 신호해석과 하드웨어를 필요로 하기 때문에 일반적으로 두 가지 기능을 통합하는 것이 경제적이다.

3) 음성합성

인간이 알아 들을 수 있는 기계의 음성합성은 합성 대상 어휘에 따라 제한 어휘 합성과 무제한 어휘 합성으로 분류되며 합성 방식에 따라 파형 코딩법과 음원 코딩법, 혼합 코딩법으로 분류할 수 있다.[5]

제한 어휘 합성은 합성하고자 하는 어휘들을 미리 분석하였다가 이들의 조합에 의해 음성을 합성하는 방법으로써 합성 대상 단위가 제한된다. 주로 단어 또는 소문장 단위의 음편들을 연결하여 말을 합성하는데 현재 지하철 안내방송, ARS(Audio Response System) 등에 이용되고 있다. 구현이 용이하며 무제한 어휘 합성에 비하여 높은 음질을 얻을 수 있으나 음편들의 연결이 부자연스러우며 합성 대상이 어휘가 바뀔 때 마다 다시 녹음, 분석해야 되는 단점이 있다. 반면에 무제한 어휘 합성은 언어의 기본 단위인 음소, 음절 등의 조합에 의해 말을 합성해 내므로 합성 대상 어휘에 제한이 없으며 주로 TTS(Text-to-Speech) 장치 및 CTS (Context-to-Speech) 장치 등에 적용된다. 그러나 음소, 음절 등의 연결시 상호 조음현상의 처리 및 자연스러운 운율처리 등이 아직 미흡하여 현재까지는 제한 어휘 합성방법에 비해 음질이 떨어지는 단점이 있다.

합성 방법에 의한 분류에서 파형코딩 방법은 구현이 용이하고 음질이 좋은 반면 저장해야 할 데이타양이 많아 제한 어휘 합성기에 많이 쓰인다. 음원 코딩법은 인간의 성도 특성을 모델링하여 특징 파라메타의 시간적 변화 정보에 의해 음성을 합성한다. 파형코딩 방법에 비해 연산량이 많고 음질도 떨어지나 데이타 압축률이 높고, 특징 파라메타의 변환에 따라 말의 속도 음높이 변환 등이 용이하여 주로 무제한 어휘 합성에 응용된다.

TTS 합성용으로 성공적으로 여러 개가 구현되고 상품화가 된 상태이고, 계속해서 성능을 향상시키고 있다. 이 분야의 최종 목표는 사람과 같은 음성을 합성해 내는 것인데, 이를 위해 현재 파싱 규칙, 운율계산, 그리고 부드럽고 자연스러운 조음생성에 대한 연구를 주로 수행하고 있다.

4) 사운드

사운드는 인간들이 일상 생활에서 사용하고 있는 가장 자연스럽고 친밀한 정보전달의 수단이다. 따라서 기계와 인간과의 인터페이스에서 이미 사운드를 사용하고 있지만 대체로 경보를 위한 간단한 부저만을 도입하였다. 연구 결과 많은 사람들이 사운드의 효용성에 긍정적인 생각을 하고 있는데, 그 이유는 시각 다음으로 정보전달의 중요한 수단으로 여기기 때문이다.[6] 예를 들면 사운드는 실제 환경에서 다음과 같은 정보를 사용자에게 제공할 수 있다.

- 물리적 사건 정보 : 컵을 떨어 뜨릴때 발생되는 튀거나 깨지는 소리
- 보이지 않는 구조 정보 : 벽에 걸린 액자의 위치를 단지 벽을 두드리므로서 알 수 있음
- 동적 변화에 대한 정보 : 컵에 물을 채울 때 소리만으로 수위를 알 수 있음
- 비정상적 구조 정보 : 소리만으로 엔진의 비정상을 알 수 있음

실제로 애플은 휴대용 정보단말기인 뉴톤(Newton)에 현실감 증가를 위해 사운드와 애니메이션을 결합한 인터페이스를 구현하여 사용자로부터 좋은 반응을 얻고 있다.[7] 그러나 이 분야에서는 뉴톤보다 앞서 제록스 PARC에서 매킨토시용으로 다양한 소리를 낼 수 있는 SonicFinder를 개발했다.[8] 예로서 아이콘을 선택하여 끌 때 끄림소리를 내고, 파일 아이콘을 선택할 때에는 파일 크기에 따라 소리 크기가 다르게 난다. 이러한 소리들을 auditory icon이라 하며, 기본 개념은 소리와 실제 오브젝트를 연결하여 사용자가 쉽게 그 내용을 파악할 수 있게 하는 것이다. SonicFinder는 애플 내부에서 충분히 효용성 실험을 한 결과 사운드 인터페이스가 없을 때 보다 훨씬 현실감을 느낀다고 말하였다. 반대로 실험에 참여한 사용자들에게 사운드를 없애고 사용하게 한 결과 매우 답답함을 표시했다. 이와 같이 사운드는 기계을 운용할 때 자연스러운 환경을 제공하는 것 이외에 기계의 동작 상태를 사용자에게 제공하는데 아주 효과적이다.

예를 들면 막대그래프의 변화로 파일 복사 진척도를 표시하던 것을 컵에 물을 채울 때 발생되는 소리로 대치 할 경우 사용자는 계속해서 화면을 주시하는 일로부터 해방될 수 있다.

상기와 같은 사운드 인터페이스의 장점에도 불구하고 현재 상용화를 적극적으로 추진하지 않고 있다. 그 첫째 이유는 작업장에서 여러 사용자들의 기계에서 동시에 발생되는 사운드는 정보전달 수단보다는 오히려 소음으로 느낄 수 있다. 그 다음 이유는 인위적으로 합성된 사운드가 아직 자연스럽지 못하고, 실제 환경의 운용상의 상황을 어떤 소리로 표현해야 되는지에 대한 아이디어가 충분하지 않다는 것이다. 이러한 장애 요소에도 불구하고 상용화를 계속 추진하기 위해서는 사용자와 사용 환경에 따라 변화하는 사운드의 효과를 반영할 수 있는 보다 유연한 인터페이스 툴을 제공하는 것이 바람직하다. 실제로 SonicFinder에서는 사용자가 사운드 인터페이스를 조정 가능하도록 하여, 최종적으로 사용자 개인의 주관적인 취향에 따라 스스로 새로운 사운드를 추가하거나 필요에 따라 사운드를 ON/OFF 할 수 있도록 할 계획으로 있다.

2. 얼굴 인터페이스(Facial Expression Interface)

인간과 인간 사이의 통신에서 얼굴 표정(Facial Expression) 만큼 확실한 정보전달의 수단은 없을 것이다. 일본의 소니는 인간과 기계와의 인터페이스에서 얼굴 에니메이션의 효과를 검토하기 위하여 자사의 상품을 소개하는 멀티미디어 시스팀을 구현하였는데, 여기서 사용한 3-D 에니메이션은 현실감을 제공하기 위하여 500poligon face를 사용하였으며, 출력정보의 내용에 따라 26개의 다양한 얼굴표정을 출력할 수 있도록 하였다.[9,10] ClearBoard는 얼굴표정 인터페이스를 도입한 또 다른 방법이다.[9,11] 이것은 그룹웨어 시스팀으로 칼라 펜과 whiteboard을 동시에 사용할 수 있으며, whiteboard 배면에서 프로젝션된 상대방의 얼굴도 동시에 볼 수 있도록 실현하였다. 일반적으로 비디오 윈도우와 whiteboard 윈도우를 가진 원격회의 시스팀 환경에서는 공유 작업공간인 whiteboard와 상대방의 제스추어와 얼굴표정을 관찰할 수 있는 공간인 비디오 윈도우가 서로 분리되어 있다. 공동작업을 수행하는 사람들 간의 시각 통신(visual communication)에서 가장 중요한 요소가 시선(gaze)이라는 것이 실험결과 입증되었다. 여기서 시선이라는 것은 상대 작업자가 어디를 주시하고 있는 지를 알 수 있는 중요한 정보 전달의 한 수단을 의미한다. 이러한 결과를 고려할 때 작업공간과 얼굴을 볼 수 있는 비디오 공간이 분리된 공동 작업시스팀은 통합된 시스팀보다 작업 능률의 저하가 예상된다.

서로의 시선을 감지하기 위하여 개발된 메타포는 통신하는 상대방 사이에 있는 clearboard 위에 얼굴을 보면서 drawing 할 수 있는 메타포이다. 이 시스팀의 목적은 whiteboard를 사용하면서 상대방의 얼굴표정을 볼 수 있기 위한 것으로 중단하지 않고 계속 집중해서 일을 할 수 있다는 장점이 있다.

3. 동작 인터페이스(Gesture Interface)

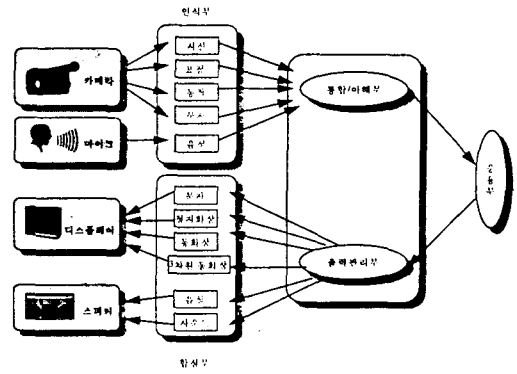
동작(gesture)이란 정보을 내포하고 있는 신체의 운동을 의미한다. 매일 사람들은 사람들 뿐만 아니라 심지어 고양이나 개 등과 같은 동물과도 동작을 사용하여 정보를 통신한다. 특히 동작은 의사표현이 어렵거나 자기의 생각을 강조하거나 감정을 표현할 때 주로 유용한 수단으로 사용된다.[12] 그러나 사람들은 일상생활의 일부가 되어 버린 동작의 중요성에 대해서 특별한 관심을 갖지 않고 있다. 예를 들면 헤어질 때 안녕이란 말만하고 손을 흔들지 않거나 길을 가르킬 때 손가락으로 지시하지 않을 경우 대단히 어색할 것이다. 현재의 컴퓨터의 사용자 인터페이스가 사용하기 편하고 만족할 만한 수준이라면 동작을 사용자 인터페이스로 사용하려는 시도를 하지 않을 것이다. 따라서 이러한 사용상의 불편함을 보완하기 위한 인터페이스 가운데 관심을 끄는 것이 동작인식시스팀이다. 동작 인터페이스를 효과적으로 사용할 수 있는 일들은 많이 있다. 우리는 3-D 그래픽 프로그램을 사용할 때 물체를 제자리에 위치시킬 때 많은 어려움을 느낀다. 심한 경우 원하는 물체를 위해 스케일링하고 회전 정도와 좌표를 직접 입력하도록 인터

페이스가 불편하게 되어 있다. 물론 마우스를 사용하여 물체를 끌어 원하는 위치에 놓고 모드변환을 이용해서 물체의 크기 조정과 회전을 수행하는 조금 편리한 프로그램도 있다. 이러한 소프웨어 사용자의 대부분은 아마 화면에 접근하여 물체를 집어 원하는 위치에 놓고 물체의 크기를 줄이기 위해 누르고 싶은 생각을 하게 된다. 이런 생각의 실현은 3차원 공간에서 gesture를 이용하여 gesture와 조작할 물체를 적절히 mapping 한후 손을 움직여 크기와 회전를 조작하므로서 가능할 것이다. 또 gesture는 다른 입력 수단을 명확하게 해주는데 탁월한 효과가 있다. 예를 들면 MIT에서 개발한 "Put That There"는 입력으로 음성과 동작을 같이 사용한 인터페이스 시스팀이다.[13,14,15] 사용자는 마이크와 동작추적(gesture-tracking) 장치를 장착하고 화면 앞에 앉아 음성과 pointing을 조합하여 물체를 생성, 명명, 이동, 복사 그리고 지우기도 할 수 있다. 실체 물체를 이동시키 위해 우선 음성으로 "Put That…"이라는 명령을 한후 위치를 명확히 하기 위해 손가락으로 위치를 가라키며 "There"라고 명령하면 된다. 동작 인터페이스를 구현하기 위해서는 사용자의 동작을 추적하는 입력 장치외에 동작을 인식할 수 있는 소프웨어 기술이 필요하다. 인식 소프트웨어의 구현은 the number of degrees of freedom, 인식할 동작의 수, 연속동작인가 이산동작인가에 따라 난이도가 다르다. 특히 정적인 동작에 비해 동적인 동작의 인식은 시간 dimension 추가로 구현이 어렵다. 현재 동작인식 시스팀은 세계적 컴퓨터 업체인 미국의 DEC사와 일본의 NEC에서 손동작을 이해할 수 있는 수준의 시스팀까지 개발되고 있다. 물론 아직은 초보적인 수준에 지나지 않지만 최근의 기술발전 속도를 감안할 때 앞으로 인간의 희로애락을 감지할 수 있는 시스팀도 나올 것으로 기대해 볼 만하다.

III. 멀티모달 인테페이스 시스팀 연구 현황

지금까지 앞장에서 얼거한 다양한 인터페이스 기술들은 상호 배타적 관계가 아니라 필요에 따라 몇 개의 첨단인터페이스를 결합해 새롭고 편리한 사용자 환경을 구축하려고 하는 멀티모달 휴먼 인터페이스의 연구가 활발해지고 있다.

여기서 멀티모달의 개념을 예를 들어 쉽게 설명하면, 도로안내를 가르켜 주는 경우 동시에 방향을 손이나 시선으로 표시하는 것처럼 인간이 무엇을 전달하고자 할 경우 음성이나 몸동작이 함께 이루어지는 바 음성(청각), 동작 및 표정(시각), 접촉(촉각) 등을 총동원하여 의사전달 효율을 높이고자 하는 것이 멀티모달의 생각이다. 이것을 실현하기 위해서는 그림 3과 같이 음성, 화상을 인식/합성하는 모듈 외에 다수의 정보를 통합관리하는 기술이 필요하다. 아직 실용화된 상품도 없고 인식결과를 통합하여 내용을 이해하는 연구가 겨우 시작되었을 뿐이지만 가상현실 기술로 만들어진 가상비서(virtual agent)에게 멀티모달로 명령을 내려 컴퓨터를 작동시키는 일도 완전히 꿈같은 이야기는 아니게 될 것이다. 따라서 본 장에서는 대표적인 연구 예로서 미국의 MIT에서 수행 중에 있는 화상인식시스팀 ALIVE를 기본으로 멀티모달 인터페이스를 위한 몸과 얼굴의 시각인식에 대한 연구내용과 일본 히다치제작소 중앙연구소에서 연구 중인 방의 가구배치 변경을 윅스테이션의 디스플레이위에서 지시하는 인테리어 디자인 시스팀을 시험제작하여 다양한 입력모드에서의 사용자 사용성을 비교한 결과에 대해서 설명하겠다.



〈그림 3〉 멀티모달 휴먼인터페이스 처리과정

1. MIT의 ALIVE[16]

1) 개 요

미국의 메사추세츠공대 미디어 연구소(MIT Media Lab.)에서는 멀티모달 인터페이스를 위한 몸과 얼굴의 시각인식에 대한 연구를 수행중이며 그 주요 연구내용은 화상인식시스템인 ALIVE로서 입력장치 앞에서 인간이 손을 들거나 가리키는 등 다양한 포즈를 취하면 기계가 인간의 모습만 배경으로부터 분리하여 이미지 처리를 하는 것이 가능하고, 손등만 클로즈업하여 손을 흔들거나 손가락을 구부리는 등의 동작을 패턴인식하는 것도 가능하며 출력 인터페이스로 인간 얼굴표정의 입체적인 효과를 이미지 상태로 재현한 와이어프레임이라고 불리우는 시스템도 같이 연구하고 있으며, 이것을 사용하여 음성과 입모양 등의 움직임이 연동하는 시스템을 개발중이다.

2) ALIVE 시스템[17]

ALIVE는 Media Lab에서 연구중인 비디오 화면을 통해서 사람과 가상 물체(simulated agent)가 실시간이면서 능동적으로 대화할 수 있는 시스템에서 가장 중요한 기술로서 사람의 움직이는 자세와 손의 위치를 인식하고 손동작(hand gesture)을 인식할 수 있는 영상처리 시스템이다. 이러한 환경에서 사용자와 agent와는 서로 볼 수가 있는데, 사용자는 비디오 화면을 통해 agent를 볼 수가 있고 agent는 컴퓨터 비젼 시스템을 통해서 사용자를 볼 수가 있다. 이런 대화성 있는 비젼 시스템의 개발에는 일반적인 영상처리 기법뿐만 아니라 active/situated 비젼 기법이 필요하다. 여기에서는 agent의 동작선택과 사용자 인식을 위해 behavior-based 기법을 사용하였다. 지금부터는 ALIVE을 실현하기 위한 관련된 주요 사항들에 대해 기술하겠다.

가) 배경으로 부터 사용자 찾기

사용자의 동작을 분석하기 위해서 비젼 시스템은 제일 먼저 배경(background)으로부터 사용자를 분리하는 것이 필요하다. 이를 위해서 가장 간단한 방법은 배경의 칼라을 시스템이 알고 있는 경우에는 입력되는 영상에서 물체와 배경을 쉽게 분리할 수 있다. 그리고 임의의 배경인 경우에도 배경의 평균과 분포에 대한 정보를 계산하여 각 화소가 배경인지 사용자 인지를 정확하게 판단할 수 있다.

나) Scene projection and calibration

일단 배경으로부터 사용자가 분리되면 가상세계에서 대충의 사용자의 위치를 알아야 하는데, 사용자가 바닥에 앉거나 서있다고 가장하고 카메라의 calibration 상태를 알고 있을 경우 사용자의 모습을 외접하는 3차원 박스의 위치를 계산할 수 있다.

다) Hand tracking

가상세계의 agent가 사용자와 대화하는데 사용하는 중요한 특징 중의 하나는 사용자의 손 위치이다. 따라서 입력영상 내에서 손의 위치를 밝히기 위하여 공간탐색패턴(spatial search pattern)을 사용한 알고리즘을 구현하였다. 또한 hand tracking 알고리즘으로 다수의 서로 다른 context-dependent 탐색 휴리스틱을 사용하여 구성하였다. 예를 들면 손은 입력영상에서 수평 에지로 존재하므로 이것을 찾기 위해서는 사용자 영상의 상단 측면부를 따라 정규화 상관 탐색을 적용한다. 상단 측면부를 탐색영역으로 제한하는 이유는 손이 있는 위치가 위쪽 측면에 존재하고, 또한 아래쪽에 있는 발도 수평으로 존재하여 혼동을 일으킬 가능성이 있기 때문이다. 그리고 사용자의 자세를 추측할 수 있는 박스의 크기에 따라 탐색 윈도우와 탐색 패턴들도 틀린 것을 사용한다.

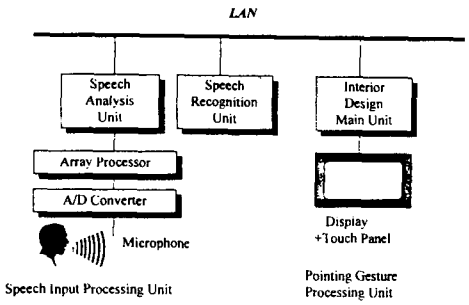
라) Gesture Interpretation

손의 절대 위치와 손이 나타내는 동작패턴은 가상세계속의 agent와 사용자가 대화하는 가장 중요한 수단이다. 여기서는 손이 나타내는 패턴을 간단한 low-level 인식 방법을 사용하여 해석하고 있으며, 사용 가능한 동작의 수를 spatio-temporal 특성을 갖는 pointing과 waving 2개로 제한하고 있다. 위의 두 개의 동작은 손의 상대적 위치정보와 움직이는 정도 및 방향 정보를 이용하여 해석하도록 하였다.

2. Hitachi의 Interior Design System[18]

1) 개 요

일본 히다치제작소 중앙연구소에서는 방의 가구



〈그림 4〉 Hitachi's Interior Design System

배치 변경을 웍스테이션의 디스플레이위에서 지시하는 인테리어 디자인 시스팀을 시험 제작하여 다양한 입력모드에서의 비교를 하였다. 이 시스팀은 HMM을 사용한 음성인식 기능과 화면에 손가락을 접촉하여 지시하는 터치패널을 갖추고 있는 방 배치 시뮬레이션 시스팀이다. 화면을 손가락으로 접촉하면서 〈이 책상을 이쪽으로 옮기시오〉라고 문장으로 지시할 수 있는 외에 〈이동〉, 〈삭제〉 등 짧은 단어로 지시하는 것도 가능하다.

2) 시스팀 구성

시스팀은 위의 그림과 같이 크게 3부분으로 구성되어 있다.

가) Pointing gesture 처리부

이 곳에서는 touch panel에서 입력되는 신호를 초당 180포인트로 샘플링 X,Y 좌표로 변환을 수행한다.

나) 음성입력 처리부

이 곳에서는 사용자가 발성하는 음성을 문자 정보로 변환하기 위해 HMM 기술을 사용하였다. 이 인식 시스팀은 학습부, 음성분석부, 그리고 패턴 매칭을 위한 네트워크 탐색부로 구성되었다. 학습부에서 regular grammar를 사용하여 만든 string을 컴파일하여 HMM의 표준패턴을 생성한다. 음성 분석부에서는 음성신호를 디지틀화 하고 이를 부분석하여 LPC cepstrum, delta cepstrum, 에너지 정보가 VQ(vector quantization) 코드를 추출한다. 네트워크 탐색부에서는 VQ 코드를 사용해서 학습부에서 구축된 네트워크내의 표준패턴 코드와 매칭을 하여 인식된 문자열을 추출해 낸다.

다) 인테리어 설계 부

사용자의 의도를 파악하기 위해서 여러 입력 장치로부터 들어오는 입력 정보를 통합하는 곳이다. 음성 입력 처리부에서 인식된 문장에서 우선 동사를 찾아 내고, 그 다음 동사에 따라 물체와 위치 단어를 추출한다. 그리고 음성처리부에서 추출된 단어와 pointing gesture 처리부에서 추출된 X, Y 좌표를 입력 순서에 따라 조합하여 최종적으로 사용자가 가리키는 물체와 위치를 인식하게 된다.

3) 성능분석

시스팀의 성능을 분석하기 위해서 다음 세 가지 경우에 대해 실험하였다.

- Pointing gesture만을 사용한 경우
- 커맨드 명령과 pointing gesture 병용한 경우
- 문장 명령과 pointing gesture 병용한 경우

사용성 테스트를 위해 실제로 20명의 사람들에게 입력작업을 하도록 해서 그 감상을 들어본 바 커맨드 발성과 터치패널의 병용이 뛰어나다고 대답한 사람이 12명으로 가장 많았으며 문장에 의한 발성과 터치패널의 병용 및 터치패널만을 평가한 사람은 모두 4명씩이었다. 문장에 의한 지식의 평가가 낮았던 것은 음성인식 시스팀이 〈이동〉이라는 짧은 커맨드는 거의 완전히 알아듣지만 긴 문장에서는 오인식이 발생하기 때문으로 보인다. 연구자는 현재의 기술에서도 커맨드를 파악하는 멀티모달 입력시스팀은 충분히 가능하다는 것을 입증하였다.

3. 기 타

일본 전기통신대학은 국제전기통신 기초기술연구소(ATR) 음성번역통신 연구소와 공동으로 인간이 지리안내를 할 때 지도를 사용하는 멀티모달 대화와 전화만으로 대화하는 경우를 비교한 실험결과 멀티모달 대화쪽이 의사소통이 원활하게 진행되었다.[19] 지도를 가르키는 경우 〈여기〉나 〈이것〉이라는 언어가 많이 사용되는 등 언어의 사용빈도가 상당히 바뀌는 것도 확인되었다.

미국의 SRI International에서는 음성과 필기입력을 사용한 Human-computer interaction에서 두 입력 기술의 자연스러운 조합방법을 연구하고

실제 사용자들을 대상으로 그 성능을 평가하였다.[20] 실험 결과 유니모달 인터페이스 보다 두 가지 입력 모드를 상호 보완적으로 사용하는 사용자의 비율이 57%나 되었다.

IV. 결 론

지금까지 2장에서는 멀티미디어 시스템에서 기계와 사용자 사이의 대화를 위한 중요 인터페이스 기술에 대해 기술하였고, 3장에서는 앞서 기술한 인터페이스 기술들을 조합하여 구현한 멀티모달 시스템에 대해 예를 들어 설명을 하였다.

현재 대량 음성이나 화상, 데이터가 초고속으로 전송되는 멀티미디어시대를 맞이해 이를 위한 정보경로 정비는 착실히 진행되고 있지만 인간과 기계의 접점이 여전히 키보드 조작이나 마우스로 되어 있어 사용이 부자연스럽다. 앞으로 이러한 부자연스러운 인터페이스를 보다 자연스럽고 인간지향적인 인터페이스로 대치하기 위해서 인공/가상현실감(Artificial/Virtual Reality), 감성공학적 기술, 인간 감각 성능 측정 기술 등 뿐만 아니라, 인간의 잠재적 감각 기능들을 좀더 유용화할 수 있는 다양한 멀티모달 인터페이스와 적응성이 탁월한 메타포 등이 심도있게 연구되고 설계되어질 것이다.

본고에서는 특히 사용자에게 자연스러운 인터페이스를 제공하기 위한 기술들 중에서 멀티모달 인터페이스에 대해 중점적으로 다루고 있는데, 여기서 중요한 것은 지금까지 열거한 다양한 인터페이스 기술들이 상호 배타적 관계가 아니라 필요에 따라 몇 개의 첨단인터페이스를 결합해 새롭고 편리한 사용자 환경을 창조할 수 있다는 것이다. 그러나 실질적으로 사용자 입장에서는 이러한 인터페이스 옵션들을 불규칙하게 조합하여 만든 멀티모달 인터페이스를 사용하기란 아마도 기존의 인터페이스 보다 이용하기가 훨씬 더 어려울지 모른다. 따라서 새로운 사용자 인터페이스는 인간의 능력과 지식의 범주내에서 계획되고 만들어 져야 한다.

가상현실 시스템도 최근 각광을 받고 있는 컴퓨터 인터페이스이다. 실제와 같은 상황을 연출해 컴퓨터를 조작하는 가상현실시스템은 자연에 가장 근접한 사용자 환경을 제공한다는 점에서 컴퓨터 인터페이스의 혁명이라고 할 수 있다. 따라서 향후 멀티모달은 가상현실기술과 함께 멀티미디어를 완결시키는 역할을 담당할 것으로 보인다.

참 고 문 헌

[1] J.L.Flanagan, "Technologies for Multimedia Communications," Proceedings of the IEEE, vol.82, no.4, April 1994.

[2] J.Wilpon, L.Rabiner, C.LEE, and E. Goldman, "Automatic recognition of key words in unconstrained speech using hidden Marcov model,"IEEE Trans. ASSP, vol.38, no.11, Nov. 1990.

[3] F.K.Soong and A.E.Roseberg, "On the use of istantaneous and transitional spectral information in speaker recognition," IEEE Trans. ASSP, vol.36, June 1988.

[4] A.E.Roseberg and F.K.Soong, "Evaluation of a vector quantization talker recognition system in text independent and text dependent modes,"Computer, Speech and Language, vol.22, 1987.

[5] 이 윤근, 안 승권, "음성 합성 기술 분야,"전자공학회지, 제20권, 제3호, 1993년 5월

[6] S.J.Mountford, "Talking and Listening to Computers," The Art of Human-Computer Interaction vol.4, 1989.

[7] A.Gobayashi, "Apple Unveils the Prototypes and Technology of the Personal Digital Assistants Newton," Nikkei Electronics, no.556, June 1992.

[8] W.W.Gaver, "The SonicFinder : An Interface that uses Auditory Icons." Human Computer Interaction, vol.4, no.1, 1988.

[9] Blattner M.M., "In Our Image : InterfaceDesign in the 1990s," IEEE Multimedia, vol.1. no.1, spring 1994.
[10] A. Takenchi and K. Nagao, "Communicative Facial Displays as a New Conversational Modelity," Proc. ACM InterCHI93, ACM press, 1993.
[11] Hirosh Ishii, Minoru Kobayashi, and Kazuho Arita, "Iterative Design of Seamless Collaboration Media," Communications of ACM vol.37, no.8, Aug. 1994.
[12] G. Kurtenbach and E.A. Hulteen, "Gestures in Human-Computer Communication," The Art of Human-Computer Interaction vol.4, 1989.
[13] R. Bolt, "'Put taht there' : Voice and Gesture at Graphics Interface," SIGGRAPH 80 Proceedings, vol.14, no.3, July 1980.
[14] S.J. Mountford, R. Penner, and P. Bursch, "Pilot command Interfaces for Discrete control of Automated Nap-of-Earth Flight," Proceedings of the AIAA/IEEE 6th Digital Avionics System Conference, 1984.
[15] C.M.Schmandt and E.A.Hulteen, "The Intelligent Voice-Interactive Interface," Proceedings of Human Factors in Computing System, March 1982.
[16] Pentland A. P. and Darrel T. "Visual perception of human bodies and faces for multi-modal interfaces," Proceedings of ICSLP 94, vol.2 pp.543~546.
[17] Pentland A. P., Darrel T. and ect. "Visually Guided Animation," Computer Animation "94, Geneva, Switzerland, May 1994.
[18] Haru Ando, Yoshinori Kitahara and Nobuo Hataoka, "Evaluation of Multimodal Interface Using Spoken Language and Pointing Gesture on Interior Design System," Proceedings of ICSLP 94, vol.2 pp.567~570.
[19] K.Loken-Kim, F.Yato, L.Fais, T.Morimoto, A.Kurematsu, "Linguistic and Paralinguistic Differences Between Multimodal and Telephone-only Dialogues," Proceedings of ICSLP 94, vol.2 pp.571~574.
[20] S.Oviatt and Erik Olsen, "Integration Themes in Multimodal Human-Computer Interaction," Proceedings of ICSLP 94, vol.2 pp.551~554.

저 자 소 개

李 憲 柱

1959年 8月 19日生
1983年 2月 경북대학교 전자공학과 졸업
1985年 2月 연세대학교 전자공학과 석사 졸업
1990年 2月 연세대학교 전자공학과 박사 졸업

1990年 1月~현재 LG전자 기술원 지능정보1실 책임연구원

주관심 분야 : 영상신호처리, 비데오신호 압축, 패턴인식

石 玟 秀

1948年 4月 28日生
1968年 10月 서울대학교 전자공학과 재학
1970年 6月 University of California, Davis 학사
1974年 6月 University of California, Davis 박사

1977年 8月～1979年 9月 Rockwell International
1979年 10月～1982年 12月 한국과학기술원 전기 및 전자공학과 교수
1983年 1月～1994年 12月 Syracuse University 교수
1994年 6月～현재 LG전자 기술원 연구위원

주관심 분야 : 정보처리