# 음성기반 멀티모달 인터페이스 기술 현황 및 과제

이지근[O] 이은숙 이혜정 김봉완 정석태 정성태 이용주 한문성
원광대학교 전기전자 및 정보공학부, 한국전자통신연구소

# The Status and Research Themes of Speech based Multimodal Interface Technology

ChiGeun Lee[0] EunSuk Lee HaeJung Lee BongWan Kim SukTae Joung
SungTae Jung YongJoo Lee MoonSung Han
Dept. of Computer Engineering, Wonkwang University
E-mail : (lcg74, enlion, redrose, stjoung, stjung, yjlee)@wonkwang.ac.kr
bwkim@sitec.or.kr msh@etri.re.kr

## Abstract

Complementary use of several modalities in human-to-human communication ensures high accuracy, and only few communication problem occur. Therefore, multimodal interface is considered as the next generation interface between human and computer. This paper presents the current status and research themes of speech-based multimodal interface technology. It first introduces about the concept of multimodal interface. It surveys the recognition technologies of input modalities and synthesis technologies of output modalities. After that it surveys integration technology of modality. Finally, it presents research themes of speech-based multimodal interface technology.

## I. 서론

일반적으로 인간과 인간이 대화할 때, 음성뿐만 아니라 얼굴표정의 변화, 제스처, 시선의 움직임 등과 같은 다양한 모달리티(Modality)를 사용하여 의사전달을 함으로써 서로간에 원만한 이해를 하도록 도와준다. 반면, 인간과 컴퓨터간의 대화는 키보드, 마우스, 디스플레이에 한정되어 있기 때문에 전문지식이 없는 일반 사용자가 자유롭게 컴퓨터를 사용할 수 없게 하는 원인이 되고 이미지, 디자인, 음악과 같은 감성정보를 입력하는데 치명적인 장해 요인이 되고 있다. 이러한 문제점을 해결하기 위하여, 인간이 보다 친숙하고 자연스럽게 사용할 수 있는 인간과 컴퓨터간의 인터페이스에 관한 연구가 활발히 진행되고 있다. 그 중 인간의 정보전달 수단을 적용하여 전문지식이 없는 일반 사용자도 사용하기 쉬운 차세대 사용자 인터페이스 즉, 멀티모달 인터페이스(Multimodal Interface)에 대한 연구가 활발히 진행되고 있다[1]. 다시 말하면 멀티모달 인터페이스는 여러 가지 모달리티를 이용한 인간과 컴퓨터간의 대화 방법이라고 정의할 수 있다.

멀티모달 인터페이스의 장점은 다음과 같다. 첫째, 다양한 모달리티를 사용한 입/출력은 정보의 정확도를 높여 준다. 둘째, 다양한 모달리티의 사용은 각 모달리티의 이점을 살리는 시스템을 구현을 가능하게 해 준다. 셋째, 다양한 모달리티의 사용은 새로운 응용 분야를 만들어 낼 수 있다. 예로써 인터랙티브(Interactive) TV가 있다. 넷째, 사용자의 기호에 따라 모달리티의 선택이 가능하다. 다섯째, 다양한 모달리티의 사용은 인간과 컴퓨터의 상호작용에 자연스러움을 제공한다. 여섯째, 외부 환경에 따라 모달리티의 변환이 가능해진다. 소음이 있는 환경에서는 음성정보를 시각정보로 변환하고, 어두운 환경에서는 시각정보를 음성정보로 변환하여 제공할 수 있다.

## II. 멀티모달 입력 처리 기술

음성기반 멀티모달 인터페이스에서 음성과 함께 사용할 수 있는 입력 모달리티로는 시각 정보, 촉각 정보, 미각 정보, 후각 정보 등이 있을 수 있는데, 현재의 기술로는 시각 정보가 널리 사용되고 있고 촉각 정보는 부분적으로 사용되고 있으며 미각 및 후각 정보는 아직 사용되고 있지 않다. 본 논문에서는 시각 정보로 가장 널리 사용되는 제스처, 얼굴 및 입술 정보에 대해서 기술한다.

### 2.1 제스처 인식 기술

#### 2.1.1 제스처 인식 방법

제스처 인식 방법은 입력 데이터 획득시 생성되는

데이터를 차원적으로 분류할 때 크게 2D 제스처 인식 방법과 3D 제스처 인식 방법으로 나누어지게 된다.

**1) 2D 제스처 인식**

2차원 제스처 인식방법의 대표적 예로서는 Pen writing과 Hand writing이 있는데, 인식 방법은 하나 이상의 필체와 이것을 구성하고 있는 각 좌표의 연속된 순서를 이용하여 생성될 수 있는 가능한 모양 중 하나로서 필체의 조합을 분류하는 방법을 사용하고 있다. 현재 pen writing은 PDA 등에서 보편적으로 사용될 수 있을 정도로 그 기술이 발달하였으며 hand writing 인식 기술도 상당한 수준에 올라와 있다. 이러한 2D 제스처 인식의 방법은 다음과 같은 방법을 적용하고 있다. 템플릿 기반 방법과 특징 기반 방법이 있다. 템플릿 기반 제스처 인식은 형판 정합 기법을 이용하는 방식으로서 입력되는 제스처의 패턴과 이미 정해진 원형 템플릿(Prototype Template)을 비교하여 가장 유사한 패턴을 선택하는 방식으로 각각의 제스처는 패턴 분류에 의해 특성화되고 입력 제스처는 해당되는 하나 이상의 원형 템플릿 제스처로 표현되는 방식이다. 특징기반 제스처 인식은 입력 제스처의 좌표에서 특징(Feature)를 추출하고 주로 통계적 패턴 분류 알고리즘을 이용하여 이미 정의된 제스처 카테고리의 집합 중 하나로 할당하여 매칭 시키는 방법이다[2].

**2) 3D 제스처 인식**

손이나 몸의 움직임을 인식하는 3D 제스처 인식 방법은 크게 두 가지 방법으로 나뉘게 된다. 첫 번째는 움직임의 좌표를 캡쳐할 수 있는 센서를 사람의 신체에 부착하고 입력되는 좌표를 패턴 분류 알고리즘을 이용하여 인식하는 장치기반 제스처 인식 방식이 있고 다른 하나는 컴퓨터 비젼 기술을 이용하여 하나 이상의 카메라로 입력되는 이미지 데이터의 Feature vector를 분석하고 패턴 분류 알고리즘을 적용하여 인식하는 방법이 대표적이다[3].

장치를 이용하는 방법은 현재 컴퓨터 애니메이션이나 방송 또는 영화 촬영에서 주로 사용되고 있는 방법으로 입력되는 데이터 값이 비교적 정확하다는 장점은 있지만 고가의 장비가 요구되고 사람의 신체 부위에 직접 장착해야 한다는 번거로움과 이로 인해 측정대상들이 거부감을 느낀다는 단점을 가지고 있다. 컴퓨터 비젼 기술을 이용한 인식 방법은 장치를 이용한 인식 방법의 단점을 보완하기 위한 방법으로 현재 많은 연구가 진행되고 있다. 범용적으로 사용되는 카메라를 이용하여 인식하고자 하는 대상을 촬영하고 이로 인해 생성되는 원시 데이터에서 움직임의 거리나 방향, 위치, 윤곽선 등을 특징 데이터로 변환하여 추출하게 된다. 추출된 특징 데이터를 통계적 학습을 통한 패턴 분류 알고리즘에 적용시켜 새로 입력되는 제스처와 비교하여 제스처를 인식하게 하는 방법이다. 근래에 컴퓨터 성능의 향상과 보급화로 컴퓨터 비젼을 이용한 인식 방법은 영상처리기술을 이용하여 넓은 영역에서 연구가 진행되고 있기 때문에 향후 인식을 위한 방법으로서의 전망이 밝다고 할 수 있다.

## 2.2 얼굴 및 입술 정보 처리 기술

음성 기반의 인터페이스에서 음성인식을 강화하고 사용자의 청각 결함을 보충하기 위한 방법으로 립리딩(lipreading)이 제시되었다. 립리딩은 잡음이 심한 공간에서 음성 인식률을 향상시킬 수 있으며 또한 사용자의 청각결함 보충에 매우 효과적인 것으로 판명되고 있다. 립리딩을 위해서는 먼저 캡쳐된 영상에서 얼굴 영역을 검출하고, 그 다음 얼굴 영상에서 입술 부분을 검출한다. 그다음 입술의 특징 파라미터를 추출하여 신경 회로망이나 HMM과 같은 패턴 인식 방법을 사용하여 음성을 인식한다.

### 2.2.1 얼굴 검출 방법

얼굴 검출 방법에는 크게 얼굴의 전체적 검출과 얼굴 특징 검출의 2가지 주된 기술로 나눌 수 있다. 주로 사용되는 얼굴의 전체적 검출 방법은 피부색이 모여있는 영역과 이 색깔 정보를 포함하고 있지 않는 배경의 영역을 구분하는 방법이다. 이 방법은 컬러정보를 이용하기 때문에 빛의 조명이나 카메라의 시점 변화에 의해 왜곡 될 수 있기 때문에 다른 기법들과 혼합하여 사용할 때 효과적인 방법이 될 수 있다. 특징기반 검출 방법은 이미지 처리 기술을 이용하여 얼굴의 부분적 또는 전체적 형태나 질감, 피부색, 움직임 정보 등의 특징을 혼용하여 얼굴영역을 검출하는 방법이다.

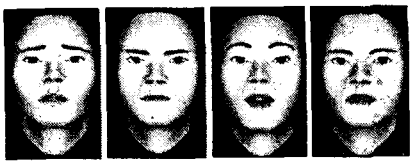
### 2.2.2 입술 특징 추출

입술 특징 추출 시 고려 사항은 조명의 변화에 독립적인 정확한 특징 점을 추출해야 하고, 사람의 발음 습관이 다르더라도 그 발음의 특징을 잘 묘사할 수 있는 특징을 구해야 한다. 입술 특징 추출 방법은 크게 픽셀 기반 방법과 윤곽선 기반 방법으로 분류할 수 있다. 픽셀 기반 방법은 입술 영역 윈도우 안의 픽셀 값에 대하여 이진화 변환, 이산 푸리에 변환, 이산 코사인 변환, 이산 웨이블릿 변환, 주성분 분석, LDA(Linear Discriminant Analysis)등의 기법을 이용하여 입술 영역 이미지로부터 주요 특징을 추출한다. 윤곽선 기반 방법은 영상 처리 기법과 능동적 윤곽선 모델, snake 모델, 변형가능 템플릿 모델 등을 이용하여 입술의 윤곽을 추출함으로써 입술의 높이, 너비, 면적, 돌출 정도, 입술 윤곽선의 움직임 정보 등을 특징으로 사용하는 방법이다. 이와 같이 추출된 입술 특징 파라미터들은 음성 인식률 향상을 위한 보완 방안으로 제시되고 있으며 음성과 얼굴 모델을 합성하는데 있어 자연스러운 입 움직임을 표현하는데 사용되고 있다.

## 2.3. 얼굴 모델링과 애니메이션 처리 기술

음성기반 멀티모달 인터페이스에서의 출력 형태로는 텍스트, 그래픽, 비디오, 음성, 애니메이션과 같은 전형적인 멀티미디어 출력, 발성 메카니즘을 인공적으로

구현하여 음성을 만들어 내는 음성합성, 가상현실에서 촉감을 느끼도록 하는 Force Feedback [1], 얼굴합성과 음성(합성)을 결합한 토킹 헤드(Talking Head) [4][5]등이 있다. 이중 토킹 헤드는 <그림1>과 같이 인간의 표정, 제스처, 입술모양, 시선 등을 의인화된 캐릭터 에이전트가 자연스럽게 재현하는 기술로써 차세대 음성기반 멀티모달 인터페이스로 주목받고 있다.



<그림1. 토킹헤드의 예>

### 2.3.1 얼굴 모델링 방법

초기에는 다양한 얼굴표정을 디지털화해서 저장한 다음, 필요한 얼굴표정을 연속해서 보여줌으로써 얼굴 모델링과 애니메이션을 생성하였다[6]. 이 방법은 간단하지만 시간이 많이 걸리는 단점이 있다. 그 후 다음과 같은 방법이 제안되었다.

**1) 파라미터 모델 (Parametric Model)**

이 모델은 얼굴형태를 나타내는 파라미터와 얼굴표정을 나타내는 파라미터를 이용하여 얼굴을 모델링한다. 전자는 눈과 코의 위치나 크기, 얼굴의 크기와 같은 얼굴 위상에 대하여 작용하는 파라미터이고, 후자는 눈썹, 입, 눈동자 등의 움직임을 표현하는 파라미터를 말한다 [5]. 애니메이션은 파라미터의 값을 변경시킴으로써 생성되어진다. 이 모델은 간단하고 데이터양이 적어 메모리 공간이 적게 필요하나, 다양한 표정의 얼굴을 표현하기 위해서는 많은 수의 파라미터가 필요하기 때문에 제어하기가 힘들어 진다.

**2) 구조화 모델 (Structural Model)**

이 모델은 얼굴을 앞머리, 뺨, 입술, 이마 등과 같은 영역과 윗(아랫) 입술, 왼쪽(오른쪽) 입술 모퉁이 등과 같은 서브 영역으로 나누어 얼굴을 모델링 한다[7]. 구조화 모델 중 FACS(Facial Action Coding System)가 가장 대표적인 방법이다[8]. 이 방법은 얼굴 움직임을 44개의 기본 동작(AU: Action Unit)으로 분해해서, 역으로 AU를 조합함으로써 얼굴표정을 합성한다. 단점으로는 주름살이나 피부의 구김, 피부의 볼륨 등을 표현할 수 없다는 것이다.

**3) 근육기반 모델 (Muscle-based Model)**

이 모델은 머리, 피부, 근육과 같은 해부학상의 특징과 피부의 탄성, 근육 수축과 같은 얼굴의 특성을 가지고 얼굴을 모델링 한다[9]. 인간의 얼굴 근육의 구성은 모두 같기 때문에 다른 모델에 적용하기 쉽다. 그러나 피부의 탄성, 근육 수축과 같은 얼굴의 특성을 정의하고 제어하기가 힘들다.

### 2.3.2 애니메이션 기법

수작업으로 얼굴표정 애니메이션을 수행하는 것은 지루한 작업이고 능숙한 기술을 갖은 애니메이터가 필요함으로 자동화할 수 있는 기법이 필수적이다.

**1) 규칙기반 애니메이션 (Rule-based Animation)**

얼굴표정은 감정, 말의 억양, 말의 의미와 관계가 있는데, 이 규칙기반 애니메이션은 이러한 관계를 규칙의 집합으로 표현한다. 또한 어떤 규칙에 대한 동작이 순서적 혹은 동시적으로 일어나는가에 대한 기술은 스크립트 언어로 정의한다. 이 스크립트 언어가 수행됨으로써 애니메이션이 생성된다 [10]. 이 기법은 간단하지만 상호작용을 구현하기 힘들며, 반복적인 애니메이션을 생성하므로 만화 영화에서 많이 사용한다.

**2) 성능기반 애니메이션**

이 기법은 애니메이션을 위한 정보를 얻기 위해 마커(Marker)나 트랙커(Tracker)를 사용하여 사물의 움직임을 추적한 다음, 3차원 공간에서의 위치를 파악해서 애니메이션을 생성한다 [11]. 자연스러운 움직임을 만들어 내지만 정보를 처리하는데 많은 시간이 들며 제작비도 비싸다.

**3) 분석기반 애니메이션**

마커나 다른 디바이스를 사용하지 않고 비디오로부터 애니메이션을 위한 정보를 추출한 다음, 애니메이션을 생성한다 [12]. 애니메이션을 위한 정보 추출을 위하여, 연속적인 이미지 사이의 각 픽셀간의 명암도를 계산하여 사물의 움직임을 벡터로 표현하는 Optical Flow [12] 방법 등을 사용한다.

## Ⅲ. 모달리티의 통합

멀티 모달 인터페이스에서 각 모달리티의 처리 기술에 더불어 모달리티를 통합하는 기술도 아주 중요하다. 모달리티는 여러 가지 형태로 통합될 수 있는데, 보완적 통합, 중복적 통합, 동등적 통합, 전문화 통합, 병렬적 통합, 전이적 통합 등으로 분류될 수 있다. 모달리티를 효율적으로 통합하기 위한 많은 연구가 수행되고 있는데, 모달리티의 사용 형태, 모달리티의 융합 형태, 모달리티를 통합 단계 등이 중요한 요소로 평가되고 있다. 모달리티의 사용 형태란 여러 모달리티를 순차적으로 사용할 것인지, 병렬적으로 사용할 것인지를 의미하며 모달리티의 융합 형태란 배타적 융합, 대안적 융합, 병렬적 융합, 상승적 융합 등이 있다. 모달리티 통합은 원시데이터를 통합하는 신호 단계의 통합이 있고, 특징 데이터를 통합하는 중간적 통합이 있으며, 각 모달리티의 인식 결과를 통합하는 의미 단계 통합이 있다.

## Ⅳ. 연구동향 및 과제

음성인식은 선진국을 중심으로 구체적인 응용분야가 개척되어서 인식과 다른 모달리티들과의 결합을 통하

여 멀티모달 인터페이스 환경 속에서 많은 응용 영역으로 발전하고 있다. 연구 동향을 살펴보게 되면 멀티모달 인터페이스 연구 영역을 컴퓨터 정보 분야에 국한시키지 않고 일상 생활 전반에 적용하여 인간 친화적인 생활을 추구하고자 노력하고 있다. 음성과 사람의 제스처, 입술 인식, 3차원 캐릭터 얼굴 모델링 기술 등을 통합하여 실생활에서 보편적으로 활용되는 가전기기에서부터 정보통신 분야까지 널리 활용하고자 하는 추세이다. 국내외 각 연구 기관에서는 이러한 추세에 따라서 활발한 연구가 진행되고 있는데 국내에 발표된 관련 연구로서는 "손제스처와 음성 인식을 이용한 인터렉티브 가상환경 기술 개발", "인공지능 인터페이스를 위한 헤드 제스처 인식 프로그램", "컴퓨터 비젼 기반의 Biometric 정보를 이용한 3차원 얼굴 에니메이션 프로그램", "영상 또는 센서 기술을 이용한 스포츠 시뮬레이터 에 관한 연구", "응시 위치 추적을 통한 인터페이스", "다중 신경 회로망을 이용한 손제스처 인식" 등, 멀티모달 연구 전반에 걸쳐 영역이 넓어지고 있다. 이와 같은 국내 멀티모달 인터페이스에 대한 연구의 구체적인 성과물을 살펴보게 되면 음성 인식을 이용한 연구 결과물은 이미 많은 분야에서 실용화되어 활용되고 있고, 아직은 완전한 수준은 아니지만 사이버 캐릭터나 어린이용 완구 등에 멀티모달 기술을 적용하여 인간의 음성과 감성, 제스처를 인식하여 표현하게 하고 이에 필요한 표정과 제스처 데이터 베이스가 개발되어 제품 개발에 활용되고 있다. 또한 사람의 눈동자나 얼굴의 움직임에 따라서 사람의 응시 위치를 자동으로 추적하여 손발이 부자연스러운 장애인이 컴퓨터를 사용하게 하거나 멀티 윈도우 환경에서 음성을 통하여 컴퓨터를 제어하는 기술 등이 현실화되었다. 국외에서도 마찬가지로 멀티모달 인터페이스 관련 기술 연구는 역행할 수 없는 거대한 흐름으로 자리잡아 가고 있다. 일례로 얼굴과 머리 제스처에 의하여 사람의 감성상태를 인식하는 기술이나 지휘봉의 움직임에 따라 음악의 강약이나 속도가 변하는 시스템 개발, 음성과 제스처 인식을 통한 어린이 학습 시스템 등, 다양한 연구를 하고 있다. 이러한 멀티모달 기술의 활용 분야로는 각종 멀티미디어 정보기기의 입출력 인터페이스, Car Navigation 시스템 개발, 장애자를 위한 서비스 시스템, 대화형 자판기, 대화형 Robot, 3차 컴퓨터 시스템 개발, 제품의 검사, 멀티모드 의료 서비스, 각종 멀티모달 데이터 베이스 검색, 멀티모달형 인터넷 검색기, 홈쇼핑, 자동 예약/문의 시스템, 음성기반 멀티모달 입출력 PC, 멀티모달형 자동항법 장치 개발 등 그 분야는 이루 헤아릴 수 없다. 이와 더불어 각국간의 자동통역전화를 위한 멀티모달에 관한 연구도 가속화될 것으로 보인다. 하지만 향후 좀더 나은 멀티모달 인식 기술을 위하여 통계적 방법을 기반으로 실재의 대량의 데이터에 기초를 둔 범용적인 인식모델을 구축하는 것과 다수의 다양한 데이터를 기반으로 하여 개인차의 모델을 추출하여 이에 의한 다수 알고리즘을 개발하는 것, 화자 독립적인 시스템을 구축하는 것, 그리고 여러 종류의 노이즈에 자동적으로 적응되는 방법을 찾는 일 등이 멀티모달 인터페이스 연구 영역의 과제로 남고 있다.

**[참고문헌]**


[1] D. Gibbon, I. Mertins and R. K. Moore. "Handbook of Multimodal and Spoken Dialogue Systems", Kluwer Academic Publishers, 2000.

[2] Newman and R. Sproull (1979). *"Principles of Interactive Computer Graphics"*. McGraw-Hill.

[3] Koons, C, Sparrell and K. Thorisson (1993). "Integrating simultaneous input from speech, gaze, and hand gesture". In: M. Maybury, ed., Intelligent Multimedia Interfaces, pp. 257-275. Morgan Kaufmann.

[4] F. Parke. Control "Parametrization for Facial Animation", Computer Animation'91, pp. 3-14, 1991.

[5] T. Guiard-Marigny, A. Adjoudani and C. Benoit. "A 3-D Model of the Lips for Visual Speech Synthesis", Proceedings of the 2nd ESCA/IEEE Workshop on speech Synthesis, pp. 49-52, 1994.

[6] J. Kleiser. "A fast, efficient, accurate way to represent the human face", ACM Siggraph'89 Course notes, pp. 20-33, 1989.

[7] S. Platt and N. Badler. "Animating facial expressions", Computer Graphics 15(3), pp. 245-252, 1981.

[8] P. Ekman and W. Friesen. "Facial Action Coding System", Consulting Psychologists Press, 1978.

[9] M. Patel and P. Willis. "FACES-The Facial Animation, Construction and Editing System", Proceedings of Eurographics'91 Conference, pp. 33-45, 1991.

[10] J. Beskow. "Animation of Taling Agents", AVSP'97 workshop, 1997.

[11] P. Litwinowicz. "Animating Images with Drawings", Computer Graphics Annual Conference, pp. 413-420, 1994.

[12] D. DeCarlo and D. Metaxas. "The Integration of Optical Flow and Deformable Models with Applications to Human Face Shape and Motion Estimation", Proceedings CVPR'96, 1996.

[13] C. Bregler, Y. Konig. "Eigenlips for Robust Speech Recognition". IEEE ICASSP, Australia, 669-672, 1991

[14] J.Yamato, J.Ohya and K.Ishii, "Recognition human action in time-sequential images using hiddenMarkov model", Proc. CVPR'92, pp.379-385, 1992.