

TV 가이드 영역에서의 음성기반 멀티모달 사용 유형 분석*

김지영(성신여대), 이경님(서강대), 홍기형(성신여대)

<차 례>

- | | |
|------------------------------|--------------------------|
| 1. 서론 | 3. 멀티모달 사용 유형 분석 |
| 2. 멀티모달 사용 유형 수집 | 3.1. 멀티모달 수집 사용 유형 |
| 2.1. 음성 및 터치 동작 사용
수집 시스템 | 3.2. 음성 및 터치 동작 사용 유형 분석 |
| 2.2. 음성과 터치 동작 사용
수집 과정 | 3.3. 음성 및 터치 동작 시간관계 분석 |
| 2.3. 수집 대상 및 시나리오 | 3.4. 연관된 터치 동작 발생시간 |
| 2.4. 수집 환경 | 3.5. 분석 결과 |
| | 4. 결론 및 향후 연구과제 |

<Abstract>

Speech-Oriented Multimodal Usage Pattern Analysis for TV Guide Application Scenarios

Ji-young Kim, Kyong-Nim Lee, Ki-Hyung Hong

The development of efficient multimodal interfaces and fusion algorithms requires knowledge of usage patterns that show how people use multiple modalities. We analyzed multimodal usage patterns for TV-guide application scenarios (or tasks). In order to collect usage patterns, we implemented a multimodal usage pattern collection system having two input modalities: speech and touch-gesture. Fifty-four subjects participated in our study. Analysis of the collected usage patterns shows a positive correlation between the task type and multimodal usage patterns. In addition, we analyzed the timing between speech-utterances and their corresponding touch-gestures that shows the touch-gesture occurring time interval relative to the duration of speech utterance. We believe that, for developing efficient multimodal fusion algorithms on an application, the multimodal usage pattern analysis for the given application, similar to our work for TV guide application, have to be done in advance.

* Keywords: Speech-oriented multimodal interface, Multimodal usage pattern.

* 이 논문은 산업자원부 지원 “자동차용 음성 HMI 시스템 기술 개발” 과제의 일환으로 수행되었음

1. 서론

사람과 시스템의 상호작용에 있어서 음성과 다른 입력을 동시에 사용하는 멀티모달 인터페이스가 사용의 편리성과 자연성으로 최근 많은 연구가 이루어지고 있다. 음성, 제스처, 영상 등의 다양한 입력 모달리티 중에서 특히 음성은 사람이 가장 편리하게 느끼며 정보 전달에 있어 주된 접근 방식이다[1]. 하지만 음성인식 기술의 발달에도 불구하고 음성만을 사용할 경우 잡음 환경에서 취약하며, 해당 객체를 말로만 표현하는 것보다 실제 객체를 가리키는 제스처가 훨씬 효과적일 때가 많다. 따라서 음성을 기반으로 다른 모달리티를 활용하는 것이 가장 이상적이라고 할 수 있다.

멀티모달 인터페이스를 활용하는 국내외 관련 연구로는 음성을 기반으로 한 터치 동작, 필기체 인식을 함께 활용하는 멀티모달 인터페이스 시스템[2][3][4], 입술 모양을 인식하는 립리딩(lip-reading)과 음성인식을 함께 결합하는 시스템[5], 음성인식과 손 동작인식을 통합한 시스템[6][7][8] 등이 있다. 또한 각각의 모달리티를 결합하기 위한 기본 연구를 비롯하여 멀티모달 상호작용이 일어나는 동안 사용자의 다양한 유형에 대한 연구[9][10][11][12]가 이루어지고 있다.

본 연구에서는 음성을 기반으로 터치 동작의 멀티모달 입력 결합(fusion)을 위한 멀티모달의 사용 유형을 분석하였다. 효과적인 음성과 터치 동작의 결합을 위해서는 특정 시점의 음성 발화가 어떤 터치 동작과 결합을 해야 하는지 음성과 터치 동작 간의 시간관계 분석이 필요하다. 이러한 분석은 로봇이나 휴대장치 등의 멀티모달 입력 결합에 이용되며, 실시간으로 연속적인 음성과 터치 동작의 입력이 있을 때 특정 시점의 발화와 연관된 동작이 무엇인지 판단하기 위한 기초 정보이다.

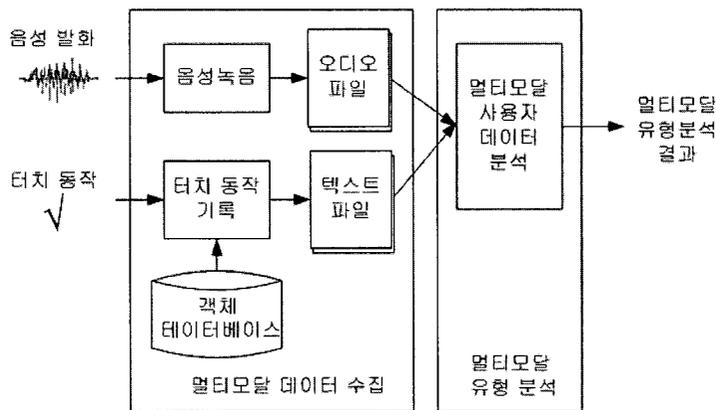
본 연구에서는 TV 가이드 영역에서 ‘특정 TV 프로그램 예약’과 ‘채널 변경’의 구체적인 작업을 수집 대상자에게 제시하고, 음성과 터치 동작을 동시에 사용할 수 있는 멀티모달 사용 수집 시스템을 이용하여 데이터를 수집하였다. 수집된 데이터를 바탕으로 음성과 터치 동작을 동시에 어떻게 사용하는지 사용 유형을 분석하고, 두 모달리티 간의 사용 시간관계를 분석하였다.

본 논문의 구성은 다음과 같다. 먼저 다음 2장에서 멀티모달 사용 유형 수집에 관한 시스템과 수집 과정을 기술한다. 3장에서는 TV 가이드 영역에서 멀티모달 사용 유형과 분석에 대해 기술하고 마지막으로 4장에서 결론 및 향후 연구 과제를 제시하였다.

2. 멀티모달 사용 유형 수집

2.1. 음성 및 터치 동작 사용 수집 시스템

음성과 터치 동작의 멀티모달 인터페이스 사용 유형 분석을 위하여 다음과 같은 멀티모달 사용 유형 수집 시스템을 구현하였다.

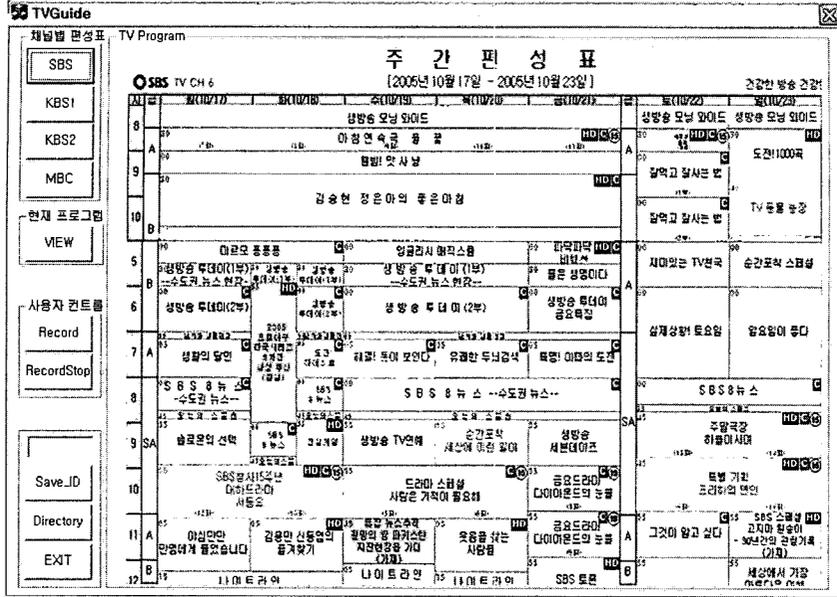


<그림 1> 멀티모달 수집 시스템 개요

음성 및 터치 동작 사용 유형 수집 시스템은 사용자로부터 음성과 터치 동작의 입력이 일어날 때, 각각의 모듈을 수행한다. 음성 녹음 모듈에서는 사용자의 음성 입력을 오디오 파일로 기록 저장한다. 터치 동작 수집 모듈에서는 사용자가 터치스크린 상에 동작을 취했을 때 선택된 좌표 옆을 해당 이미지 맵과 관련 데이터베이스를 참고하여 사용자의 터치 동작 결과를 텍스트 파일로 기록한다.

다음의 <그림 2>는 <그림 1>의 시스템에서 TV 가이드 영역의 사용 유형 수집을 위해 구현한 멀티모달 인터페이스의 실행 화면이다.

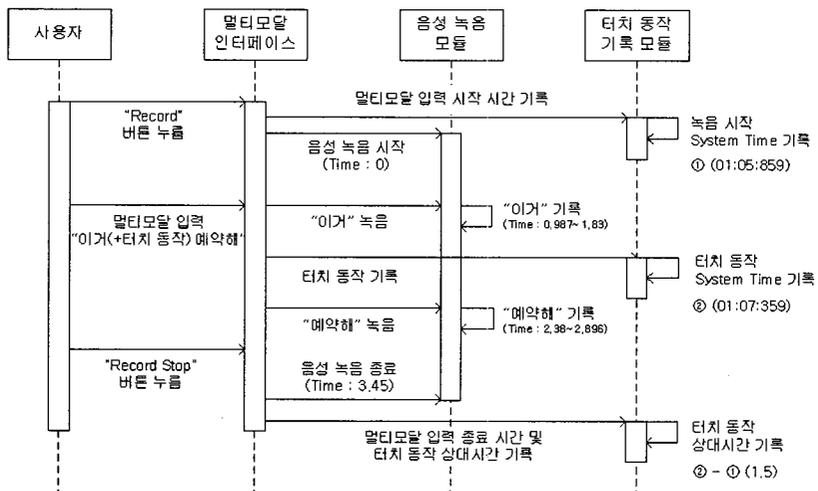
<그림 2>의 인터페이스 상에 나타난 터치스크린 상의 이미지는 TV 프로그램 목록으로 채널별로 다른 정보를 볼 수 있다. 본 논문에서 구현한 수집 시스템은 채널 선택을 위한 버튼 영역과 터치를 위한 이미지의 변경이 용이하므로, 다른 영역에서의 멀티모달 사용 유형 분석에서도 활용될 수 있다.



<그림 2> 사용 유형 수집을 위한 멀티모달 인터페이스 화면

2.2. 음성과 터치 동작 사용 유형 수집 과정

사용 유형 수집과정은 다음 <그림 3>과 같은 방법으로 이루어진다.



<그림 3> 음성과 터치 동작의 사용 유형 수집 예

- ① 사용자가 인터페이스 상에 “Record” 버튼을 눌러 멀티모달 입력의 시작을 알리면, 그 순간의 시스템 시간(1/1000초 까지)이 기록되고, 새로운 오디오파일을 열어 음성 녹음을 시작한다.
- ② 사용자로부터 멀티모달 입력이 있을 경우, 음성 녹음 모듈에서는 음성 발화를 오디오 파일에 기록하고 터치 동작 모듈에서는 터치 동작에 관련된 정보를 텍스트 파일로 기록한다. 이 때, 오디오 파일에 기록되는 음성 발화 시간은 녹음이 시작된 시간으로부터 상대적인 시간(초)이 기록되고, 텍스트 파일에 기록되는 터치 동작 관련 시간은 시스템 시간을 기준으로 기록된다.
- ③ 사용자가 멀티모달 입력을 종료하는 “Record Stop” 버튼을 누르면, 음성 녹음 모듈에서는 오디오 파일을 저장하고, 터치 동작 모듈에서는 종료시간을 기록하고 저장한다. 또한 터치 동작 모듈에서는 입력시작 버튼을 누름과 동시에 오디오 파일의 녹음 시작 시간으로부터 상대적인 시간을 기록한 음성과의 동기화를 위해서 ①의 멀티모달 입력 시작 시스템 시간과 ②에서의 터치 동작 입력 시간의 차이를 구해 터치 동작의 상대적인 시간(초)을 구하여 텍스트 파일에 기록한다.

2.3. 수집 대상 및 시나리오

사용 유형 데이터의 수집은 20, 30대의 남성 및 여성을 대상으로 총 54명이 참여했다. 수집 대상자에게는 <표 1>과 같이 3가지 작업을 제시하고, 음성과 터치스크린을 모두 사용할 수 있음을 알려주었다. 3가지 작업을 수행하기 위해서 음성과 터치 동작에 대한 제약 사항을 주지 않고, 어떤 식으로 시스템에 명령을 할지는 자유롭게 선택하도록 하였다.

<표 1> TV 가이드 시나리오를 기반으로 한 작업(Task)

작업(Task) 번호	내용
T1	선택한 채널의 편성표에서 하나의 프로그램만을 예약하는 주제 (예 : 5월 24일 수요일의 8시뉴스 프로그램이 명시되어 있는 곳을 터치하면서, “8시뉴스 예약해줘” 라고 발화)
T2	선택한 채널의 편성표에서 세 가지 프로그램을 예약하는 주제 (예 : 5월 23일 인간극장 프로그램과 5월 25일 가족오락관 프로그램이 명시되어 있는 곳을 각각 터치하면서, “이거랑 이거 예약해줘”라고 발화)
T3	채널의 변경을 요하는 주제 (예 : 터치 동작 없이 음성만으로 “SBS 보자” 라고 발화)

2.4. 수집 환경

데이터 수집은 방음 시설이 잘 된 음향 스튜디오 부스 내에서 이루어졌고, 외장 사운드 카드, 헤드셋, 그리고 전압식의 터치스크린이 장착된 태블릿 노트북을 사용하였다. 수집 환경은 다음의 <표 2>와 같다.

<표 2> 데이터 수집을 위한 시스템 환경

장소	음향 스튜디오 부스 (230cm x 230cm)	
시스템 사양	노트북 (FUJITSU LifeBook P1510)	
	CPU	Intel(R) Pentium(R) M Processor 1.20GHz
	RAM	512MB
	모니터	노트북 내의 터치모니터(전압식)
	운영체제	Windows XP Home Edition
오디오	외장 사운드카드 (Sound Blaster 24Bit), 헤드셋(Sennheiser)	
수집 툴 개발언어	Visual C++ 5.0	

3. 멀티모달 사용 유형 분석

3.1. 멀티모달 사용 유형 수집

멀티모달 인터페이스의 사용 유형을 분석하기 위하여 총 54명의 사용자로부터 TV 가이드 시나리오에 기반 한 3가지 작업에 대해 한번 씩 수행한 결과를 수집하였다. 따라서 수집된 데이터는 54명(사용자 수) * 3회(작업 당 1회) = 162 문장이다.

3.2. 음성 및 터치 동작 사용 유형 분석

음성과 터치 동작을 동시에 사용할 수 있는 경우, 앞의 2장에 제시한 3가지 작업에 대하여 사용자가 멀티모달 인터페이스를 어떠한 형태로 사용하는지 수집된 데이터를 분류한 결과는 <표 3>과 같다.

<표 3> 멀티모달 인터페이스 사용 유형

유형	예제
P_1 (Speech-Only)	“8시 뉴스 예약해”
P_2 (Redundancy)	“8시 뉴스(+터치 동작)예약해”
P_3 (Complement)	“이거(+터치 동작) 예약해” “ \emptyset (+터치 동작) 예약해”
P_4 (Combination)	“8시 뉴스랑 이거(+터치 동작) 랑 이거(+터치 동작) 예약해”

- (1) 유형 P_1 (Speech-Only) : 음성과 터치 동작을 함께 사용할 수 있음에도 불구하고, 음성만을 사용한 경우이다.
- (2) 유형 P_2 (Redundancy) : 음성 발화에서 객체 명을 발화하면서 동시에 터치 동작을 함께 취한 경우로 동일한 정보에 대해 두 모달리티를 이용하여 입력되는 경우이다.
- (3) 유형 P_3 (Complement) : 객체 명 대신 “이거”와 같은 지시어를 포함하여 발화하면서 터치 동작을 취한 경우로 터치 동작이 음성에 대한 보충정보를 제공하는 보상성을 지닌 경우이다. 여기에는 또한 객체에 대한 발화 없이 터치 동작만으로 객체를 지시하고 음성으로는 술어만 발화한 경우도 포함한다.
- (4) 유형 P_4 (Combination) : P_1 , P_2 , P_3 유형이 모두 섞여서 사용된 경우를 의미한다. 이러한 P_4 유형은 TV 시나리오의 3가지 작업 중, 두 번째인 T2의 경우에만 나타났다.

이와 같이 수집된 데이터를 분석하여 나타난 유형들에 대한 대상자의 분포도는 <표 4>와 같다.

<표 4> 유형에 따른 작업별 분석결과

유형	T1		T2		T3	
	인원수	분포율(%)	인원수	분포율(%)	인원수	분포율(%)
P_1 (Speech-Only)	9	16.6	3	5.55	37	68.5
P_2 (Redundancy)	21	38.9	13	24.1	10	18.5
P_3 (Complement)	24	44.5	35	64.8	7	13.0
P_4 (Combination))	0	0	3	5.55	0	0

하나의 프로그램을 예약하도록 하는 T1에서는 객체에 대한 보충정보를 터치 동작이 제공하는 P_3 유형이 24명(44.5%)으로 가장 빈도가 높았고, 그 다음으로는 객체 명에 대한 발화와 함께 터치 동작을 사용하는 P_2 유형으로 21명(38.9%)이었

다. 이와 동일하게 하나 이상의 프로그램을 예약하도록 하는 T2에서도 35명(64.8%)으로 P₃ 유형이 가장 많았으며, 그 다음으로는 P₂ 유형이 13명(24.1%)으로 높았다. 마지막으로 채널 변경에서는 음성 발화만을 사용하는 P₁ 유형이 37명(68.5%)으로 나타났다.

작업의 특성에 따라 모달리티의 사용 형태에 따른 선호도가 다르다는 것을 알 수 있다. T3과 같은 작업에서는 음성과 터치 동작을 함께 사용하는 멀티모달 형태의 입력방법 보다는, 간단하게 음성만을 사용하여 지시하는 것을 선호한다는 것을 알 수 있다. 이것은 채널명이나 프로그램명 등의 이미 사람들이 익숙해져 있는 용어나 객체명이 포함되는 경우에는 터치 동작이 가능함에도 불구하고 음성만을 이용하는 것을 선호한다는 것을 보인다. 그러나 T1, T2와 같은 작업에서는 채널 명, 프로그램 명, 그리고 예약이라는 특수한 상황 때문에 해당 날짜 및 시간 정보까지 발화를 해야 하므로, 음성만으로 일일이 정보를 발화해야하는 P₁ 유형 보다는 음성과 터치 동작을 조합하여 사용하는 P₂나 P₃과 같은 멀티모달 형태의 유형을 더욱 선호한다는 것을 알 수 있다. 뿐만 아니라 T1과 T2의 경우, 대상 객체수가 늘어날수록 대상 객체에 대하여 객체 명을 발화하는 P₂의 경우보다 간단히 지시어나, 발화 없이 터치 동작을 수행하는 P₃ 유형을 더 선호한다는 것을 알 수 있다.

P₃ 유형은 객체를 가리키는 방법에 있어서 터치 동작과 함께 지시어를 발화하거나 객체 명에 대한 발화 없이 터치 동작만으로 객체를 가리키는 경우를 포함한다. 이와 같은 P₃ 유형을 세부적으로 분류한 결과는 <표 5>와 같다.

<표 5> P₃ (Complement) 세부 유형

세부 유형	예 제
P ₃₋₁ (Deictic)	“이거(+터치 동작) 예약해”
P ₃₋₂ (Pointng-Only)	터치 동작 + “예약해”

P₃₋₁(Deictic) 유형은 객체를 가리킬 때, 지시어 발화와 터치 동작을 함께 사용한 경우를 의미하고, P₃₋₂(Pointing-Only) 유형은 객체에 대하여 발화 없이 터치 동작만을 사용하고 동사를 발화한 경우를 의미한다. 분석 대상 54명의 데이터 중에서 이와 같은 두 가지 세부 유형을 포함하는 P₃ 유형은 T1의 작업에서는 24명, T2의 작업에서는 35명, T3의 작업에서는 7명이 사용하였다. P₃ 유형을 사용한 사용자에 대하여 P₃₋₁, P₃₋₂의 세부 유형으로 분류한 결과는 다음의 표와 같다.

<표 6> P_3 (Complement) 세부분석 결과

세부 유형	T1(24명)		T2(35명)		T3(7명)	
	인원수	분포율(%)	인원수	분포율(%)	인원수	분포율(%)
P_{3-1} (Deictic)	10	41.7	14	40	6	85.7
P_{3-2} (Pointng-Only)	14	58.3	21	60	1	14.3

T1에 대해서는 24명 중 14명(58.3%)이 P_{3-2} 유형을, T2의 작업에 대해서는 35명 중 21명(60%)이 P_{3-2} 유형을 선호하였다. 이와 달리 T3의 작업에서는 7명 중 6명(85.7%)이 P_{3-1} 유형을 선호하였다. 이것은 멀티모달 인터페이스를 설계할 때, P_3 유형 중 객체에 대한 음성 발화 없이 터치 동작만으로 객체를 지칭하는 P_{3-2} 의 경우를 특히 고려해야 한다는 것을 알려준다. 다른 유형과 달리 P_{3-2} 유형에서는 음성에서는 어떠한 정보도 없이 터치 동작만으로 대상 객체를 인지해야 한다.

P_4 유형은 T2에 대해서만 나타났다. 54명 중 3명에게서 나타났으며, <표 7>과 같이 각각 3명이 다른 유형을 보였다. 이것은 세 가지 작업으로 제한된 시나리오를 확장하고, 수집된 대상자를 확대하면 혼합된 유형이 다양하게 나타날 것이라 예상된다. 따라서 멀티모달 인터페이스를 설계할 때에 이와 같은 혼합 유형에 대하여 다양하게 고려해야 할 필요성이 있다.

<표 7> P_4 (Combination) 세부 유형

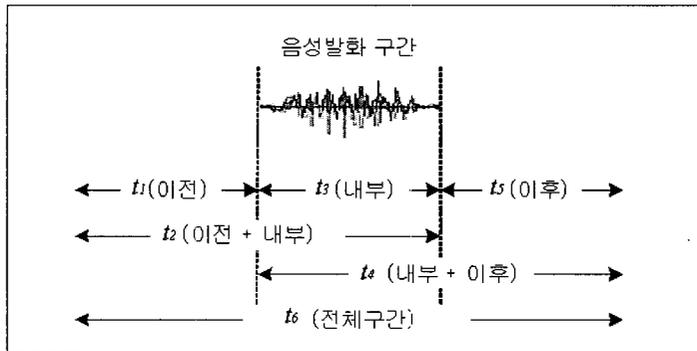
유형	예제
$P_{4-1}(P_2+P_{3-1})$	“이거(+터치 동작)랑 이거(+터치 동작)랑 그것이 알고 싶다(+터치 동작) 녹화해줘”
$P_{4-2}(P_1+P_{3-1})$	“생방송 투데이랑 이거(+터치 동작)랑 이거(+터치 동작) 예약할 수 있어?”
$P_{4-3}(P_1+P_{3-2})$	“동물농장” + 터치 동작 + 터치 동작 + “녹화해주세요”

3.3. 음성 및 터치 동작 시간관계 분석

음성 및 터치 동작의 시간관계를 분석할 때는 문장 단위로 분석했던 것과는 달리 문장 내에 포함된 대상 객체를 기준으로 분석하였다. 사용자가 “8시 뉴스(+터치 동작)랑, 이거(+터치 동작), ㅇ(+터치 동작) 예약해” 라고 음성과 터치 동작을 함께 사용하여 지시했을 경우에는 대상 객체가 3개이다. 프로그램명과 터치 동작을 함께 사용한 “8시 뉴스”, 지시어와 터치 동작을 함께 사용한 “이거(+터치 동

작), 그리고 음성 없이 터치 동작으로만 이루어진 마지막의 터치 동작이 대상 객체가 된다. 따라서 음성 및 터치 동작의 시간관계 분석에 있어서는 수집된 사용자 데이터 중에서 터치 동작이 일어나지 않는 $P_1(\text{Speech-Only})$ 유형은 제외하였다.

음성 및 터치 동작의 시간관계는 음성 발화구간을 기준으로 사용자가 대상 객체에 관련되는 터치 동작을 어느 시점에 했고 얼마나 지속했는지에 따라 <그림 4>와 같이 $t_1 \sim t_6$ 의 여섯 가지로 분류하였다.



<그림 4> 터치 동작의 시간관계

- (1) t_1 (이전): 대상 객체에 대한 발화 이전에 관련 터치 동작이 나타난 경우
- (2) t_2 (이전+내부): 발화 이전부터 발화 중간까지 관련 터치 동작이 지속된 경우
- (3) t_3 (내부): 대상 객체에 대한 발화구간 내에 관련 터치 동작이 나타난 경우
- (4) t_4 (내부+이후): 발화시점으로부터 발화 이후까지 관련 터치 동작이 지속된 경우
- (5) t_5 (이후): 대상 객체에 대한 발화구간 이후에 관련 터치 동작이 나타난 경우
- (6) t_6 (전체구간): 관련 터치 동작이 발화 이전부터 발화 이후까지 지속된 경우

3 가지 작업의 멀티모달 사용 유형에서 P_1 , 즉 음성만을 사용한 경우를 제외한 시간관계 분석 결과는 <표 8>과 같다. T2에서 분석 대상이 142개인 것은 한 문장에 3개의 대상 객체가 포함될 수 있기 때문이다. T3의 경우 일부 사용자가 채널명과 프로그램명을 함께 사용한 경우가 관측되었다.

<표 8> 작업별 시간관계 분석

시간 관계	T1 (45명, 45개)		T2 (51명, 142개)		T3 (17명, 21개)	
	관측수	분포율(%)	관측수	분포율(%)	관측수	분포율(%)
t_1 (이전)	24	53.3	65	45.8	8	38.1
t_2 (이전+내부)	1	2.22	4	2.82	4	19
t_3 (내부)	7	15.6	42	29.6	3	14.3
t_4 (내부+이후)	7	15.6	17	12.0	4	19
t_5 (이후)	5	11.1	14	9.86	2	9.52
t_6 (전체구간)	1	2.22	0	0	0	0

<표 8>에 따르면 T1, T2, T3의 세 가지 작업 모두 각각 24개(53.3%), 65개(45.8%), 8개(38.1%)로 대상 객체에 대한 터치 동작이 음성 발화 이전에 나타나는 t_1 구간이 가장 많이 나타났음을 알 수 있다. <표 8>은 P_{3-2} 의 유형까지 포함한 결과이다. 이 경우는 대상 객체 자체에 대한 발화 없이 터치 동작이 음성 발화를 대신한 경우로 대상 객체에 대한 터치 동작을 하면서 “예약해”라고 발화한 경우이다. 따라서 P_{3-2} 에서는 터치 동작이 음성 발화 이전에 나타나게 된다.

음성 및 터치 동작의 시간관계 분석에 대해서는 유형 분석에서처럼 작업을 기준으로 하는 분류대신, 대상 객체에 대한 발화의 유·무에 따른 분류를 하였다. ‘객체 발화가 있는 경우’는 P_2 유형 및 P_3 의 세부 유형인 P_{3-1} 유형을 포함하며, 대상 객체에 대하여 발화 없이 터치 동작으로만 객체 정보가 입력되는 경우는 ‘객체 발화가 없는 경우’로 P_3 의 세부 유형인 P_{3-2} 유형을 의미한다. 또한, 사용 유형 P_4 는 대상 객체에 대한 부분을 P_2 , P_{3-1} , P_{3-2} 로 다시 분류하여 분석에 포함하였다.

<표 9> 대상 객체에 대한 발화 유·무에 따른 분류

분석 기준	예 문	포함되는 사용 유형
객체 발화가 있는 경우	“8시 뉴스(+터치 동작) 예약해”, “이거(+터치 동작) 예약해”	P_2 , P_{3-1}
객체 발화가 없는 경우	터치 동작 + “예약해”	P_{3-2}

<표 8>의 208개의 분석 대상 객체를 기반으로 음성 발화의 유·무에 따른 터치 동작의 시간관계는 <표 10>과 같다.

<표 10> 대상 객체에 대한 발화 유·무에 따른 시간관계 분석 결과

시간 관계	객체 발화가 있는 경우 (129개)		객체 발화가 없는 경우 (79개)	
	관측 수	분포율(%)	관측 수	분포율(%)
t1(이전)	25	19.38	73	92.4
t2(이전+내부)	6	4.65	3	3.79
t3(내부)	48	37.21	3	3.79
t4(내부+이후)	28	21.7	0	0
t5(이후)	21	16.28	0	0
t6(전체구간)	1	0.77	0	0

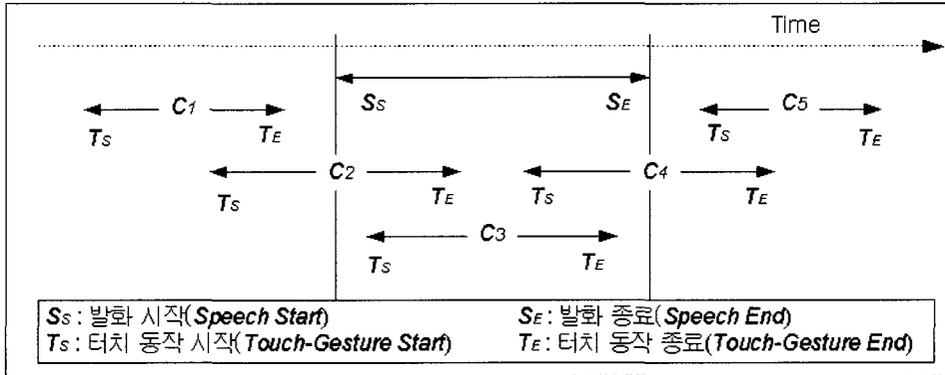
<표 10>에 의하면 ‘객체 발화가 있는 경우’에는 48개(37.21%)로 발화구간 이내에 대상 객체에 대한 터치 동작이 일어나는 t_3 가 가장 높았으며, 그 다음으로는 t_4 가 28개(21.7%)로 높았다. ‘객체 발화가 없는 경우’의 경우에는 음성 발화 이전에 터치 동작이 나타나는 t_1 의 경우가 73개(92.4%)로 높았으며, 동일하게 t_2 , t_3 가 3개(3.79%)씩 나타났다.

‘객체 발화가 없는 경우’에는 “예약해”라고 음성 발화 이후에 관련 터치 동작을 시작하는 사용자가 한 명도 없었기 때문에 t_4 , t_5 , t_6 는 0%로 나타났다. 이것은 데이터 수집을 위한 멀티모달 인터페이스가 우리말을 기준으로 설계되었고, 한국어 사용자를 수집 대상으로 삼았기 때문에 일반적으로 행위를 제시하는 동사(Verb)가 문장의 제일 뒤에 오는 한국어의 음성 언어 습관에 따른 결과이다.

3.4. 연관된 터치 동작 발생시간

음성과 터치 동작이 함께 사용된 경우, 음성 발화구간을 기준으로 연관되는 터치 동작의 발생시간을 분석하였다.

음성 정보와 연관된 터치 동작 인식 결과는 결합(fusion)을 하여야만 사용자의 완전한 의도를 이해할 수 있다. 이를 위해서는 특정 음성 발화와 연관되는 터치 동작이 어떤 것인지를 알아야 한다. 3.3절에서 기술한 바와 같이 터치 동작의 발생 시간이 음성 발화의 시작 전, 후에 걸쳐 다양하게 나타나고 있다. 사용자의 입력 시간을 명확하게 알 수 있는 PTT(Push-To-Talk)버튼 등을 사용 할 수 없는 경우에는 음성 발화와 결합할 대상인 터치 동작이 어떤 것인지를 발화구간을 기준으로 판단할 수밖에 없다.



<그림 5> 음성 발화구간 기준 고려할 터치 동작

<그림 5>는 음성 발화구간 ($S_S - S_E$)을 기준으로 하였을 때 연관된 터치 동작의 발생을 나타낸 것($C_1 \sim C_5$)이다. 발화구간 내에서 터치 동작이 시작하는 C_3 과 C_4 의 경우는 당연히 음성 발화와 연관되는 터치 동작으로 추정할 수 있다. 하지만 음성 발화 시작 이전에 터치 동작이 발생하는 C_1 , C_2 ($T_S < S_S$ 의 경우), 그리고 음성 발화 종료 이후에 터치 동작이 발생하는 C_5 ($S_E < T_S$ 의 경우)에 대해서는 어느 시점까지의 동작을 음성발화와 연관된 것으로 추정할 것인지 알기 위하여 다음과 같은 방법으로 분석하였다.

$T_S < S_S$ 의 경우는 $T_S - S_S$ 로, $S_E < T_S$ 의 경우는 $T_S - S_E$ 로 연관된 터치 동작의 발생 시간을 음성 발화구간에 대비하여 상대적으로 구하였다.

T1의 작업에 대하여 ‘객체 발화가 있는 경우’와 ‘객체 발화가 없는 경우’를 함께 고려하였을 때 평균적으로 음성 구간의 -1.03/+1.17(초) 시간구간 내에 관련 터치 동작이 시작하였다. ‘객체 발화가 있는 경우’의 경우만 고려하면 평균적으로 음성 구간의 -1.06/+1.17(초) 시간구간 내에서 대상 객체에 대한 터치 동작이 시작하였다. ‘객체 발화가 없는 경우’의 경우만 고려 할 때에는 음성 구간의 -1.0(초) 시간구간 내에 객체에 대한 터치 동작이 시작하는 것을 알 수 있다. 터치 동작만으로 객체 정보를 나타내는 것이기 때문에 이때의 음성 발화는 객체에 대한 기준이 아니므로 $T_S - S_E$ 에 대한 정보는 산출하지 않았다. T1에 대한 최대 연관 구간은 음성 구간의 -3.35/+3.45(초) 시간구간에서 연관된 터치 동작이 발생하였다. T2와 T3 작업에 대해서도 <표 12>을 통해 동일하게 해석할 수 있다.

<표 12> 음성 발화와 연관된 터치 동작 발생시간 분석 (단위: 초)

		Total Average		객체 발화가 있는 경우						객체 발화가 없는 경우	
				전체 평균값		P_2 (Redundancy)		P_{3-1} (Deictic)		P_{3-2} (Pointing-Only)	
		평균값	최대값	평균값	최대값	평균값	최대값	평균값	최대값	평균값	최대값
T1	$T_S - S_S$	-1.03	-3.35	-1.06	-3.35	-1.27	-3.35	-0.81	-1.60	-1.0	-2.46
	$T_S - S_E$	+1.17	+3.45	+1.17	+3.45	+1.41	+3.45	+0.23	+0.23	0	0
T2	$T_S - S_S$	-1.86	-7.51	-1.06	-2.19	-0.47	-1.02	-1.55	-2.19	-2.02	-7.51
	$T_S - S_E$	+0.71	+2.66	+0.71	+2.66	+0.92	+2.66	+0.09	+0.25	0	0
T3	$T_S - S_S$	-0.57	-2.38	-0.59	-2.38	-1.79	-2.38	-0.35	-1.23	-0.37	-0.37
	$T_S - S_E$	+0.47	+0.92	+0.47	+0.92	+0.92	+0.92	+0.02	+0.02	0	0

3.5. 분석 결과

지금까지 음성 및 터치 동작의 멀티모달 사용 유형, 시간관계, 음성과 관련 있는 터치 동작의 발생시간의 분석 결과를 정리하면 다음과 같다.

- 1) 음성과 터치 동작을 함께 사용할 때, 객체를 지칭하는 유형은 음성 발화만 하는 경우(P_1 :Speech-Only), 객체 명 발화와 터치 동작을 함께 사용하는 경우(P_2 :Redundancy), 지시어(“이거”, “저거” 등)와 함께 터치 동작을 사용하는 경우(P_{3-1} :Deictic), 객체 명 발화 없이 터치 동작만으로 지칭하는 경우(P_{3-2} :Pointing-Only), 이와 같은 네 가지 방법이 혼합된 경우(P_4 :Combination)에 대한 유형이 나타났다.
- 2) 음성 및 터치 동작 사용 유형은, 작업 종류에 따라 선호하는 유형이 다를 수 있었다. 객체를 가리킬 때, 익숙한 용어가 포함되는 경우에는 터치 동작이 가능함에도 불구하고 음성만을 사용하는 P_1 유형, 또는 객체에 대한 음성 발화와 함께 터치 동작을 중복 사용하는 P_2 유형을 선호하였다. 반면, 음성만으로 말하기가 긴 경우에는 음성으로는 지시어를, 터치 동작으로는 객체를 가리키는 P_3 유형을 선호하였다.
- 3) 시간관계를 바탕으로 사용자 유형을 분석한 결과, 대상 객체에 대한 발화가 있을 때는 음성 발화구간 내에 터치 동작이 발생하는 경우가 가장 많았다.
- 4) 연속적인 음성과 터치 동작의 입력이 들어올 때, 음성 발화와 관련 있는 터치 동작이 무엇인지, 시간적으로 어떻게 나타나는가를 알기 위하여 음성 발화구간 기준으로 관련된 터치 동작의 발생시간을 분석하였다. 그 결과, 음성 발화구간 기준으로 -7.51/+3.45(초)의 구간(최대)에서 연관된 터치 동작이 발

생하는 경우가 관측됨으로써 예상보다 긴 구간의 시간에 대해 멀티모달 결합시 고려해야 함을 알 수 있었다.

- 5) 객체에 대한 음성 발화가 있는 경우와 없는 경우, 연관된 터치 동작의 발생 구간에서 차이가 있었다. <표 12>의 T1의 경우를 예를 들면, “8시뉴스(+터치 동작) 예약해”, “이거(+터치 동작) 예약해”처럼 객체에 대한 음성 발화가 있는 경우에는 음성 구간의 -3.35/+3.45(초) 시간구간 내에 관련 터치 동작이 시작하였다. 하지만 터치 동작과 함께 “예약해”처럼 객체에 대한 음성 발화가 없는 경우에는 음성 구간의 -2.46(초)부터 음성 발화 종료시점까지의 시간구간 내에 관련 터치 동작이 시작하므로, 이 경우에는 해당 음성구간 이전 및 음성구간에서 시작하는 터치 동작만 고려하면 된다는 것을 알 수 있다.

4. 결론 및 향후 연구 과제

본 논문에서는 효과적인 음성과 터치 동작의 결합을 위하여 멀티모달 사용 유형 분석을 하였다. 음성과 터치 동작을 함께 사용하는 유형을 수집하기 위해 멀티모달 사용 유형 수집 시스템을 만들고 54명의 사용자로부터 TV 가이드 영역의 3가지 작업을 바탕으로 사용자 유형을 수집하였다. 수집된 유형을 바탕으로 음성 및 터치 동작의 사용 유형, 시간관계, 음성 발화구간을 기준으로 연관되는 터치 동작의 발생시간을 분석하였다. 분석된 결과를 바탕으로 음성과 터치 동작의 결합(fusion) 알고리즘을 설계할 때에는 다음과 같은 요소를 고려해야 함을 알 수 있다.

첫째, 작업 종류에 따라 나타나는 사용 유형과 선호도가 다르므로 대상 시나리오에 대하여 멀티모달 사용 유형을 수집하고, 수집된 유형의 분류(3.2절) 및 시간관계 분석(3.3절)이 필요하다.

둘째, 대상 시나리오에 따른 사용자의 멀티모달 사용 유형에서 사용자의 음성 발화 유형을 별도 수집하고, 이에 기반을 둔 음성인식기의 개발이 음성기반 멀티모달 인식의 성능을 높일 수 있다.

셋째, 음성과 터치 동작의 결합을 위하여 음성 발화구간을 기준으로 연관된 터치 동작 발생 시간 구간을 찾는 작업이 선행되어야 한다. 본 논문에서는 산술평균과 최대치만을 분석하였으나, 음성 발화구간 기준으로 터치 동작의 발생시간 분포를 이용하면, 보다 효과적인 음성과 터치 동작의 결합이 가능해질 것이다.

본 논문에서 기술한 멀티모달 인터페이스 시스템을 위한 사용 유형 분석은 향후 음성, 제스처, 영상, 감정 등의 멀티모달 결합 연구를 위한 기초 자료가 될 것이다. 앞으로는 사용 유형 수집의 카테고리 및 시나리오의 종류, 객체 수, 그리고 실험 대상자를 늘려, 멀티모달 사용 유형 분석 방법론을 보다 체계적으로 정립할 필요가 있다. 또한, 사용 유형의 분석 결과를 바탕으로 음성과 터치 동작의 의미

를 결합하는 알고리즘 개발을 지속적으로 수행하여야 한다.

참고문헌

- [1] 홍기형, “음성기반 멀티모달 인터페이스 표준”, *말소리*, 제 51호, pp. 117-135, 2004.
- [2] M. T. Vo, and A. Waivel, “A multimodal human-computer interface: combination of speech and gesture recognition”, *Proc. InterCHI*, pp. 231-249, 1993.
- [3] F. Flippo, A. Krebs, and I. Marsic, “A framework for rapid development of multimodal interface”, *Proc. ICMI*, pp. 109-116, 2003.
- [4] J. L. Flanagan, “Speech-centric multimodal interfaces”, *IEEE Signal Processing Magazine*, Vol. 21, No. 6, pp. 76-81, 2004.
- [5] J.-Y. Suh, K.-N. Lee, K.-H. Hong, and Y.-J. Lee, “Correcting Korean vowel speech recognition errors with limited lip features”, *Proc. ICSLP*, pp. 2529-2532, 2004.
- [6] R. Bolt, “Put-That-There: voice and gesutre at the graphic interface”, *Computer Graphics*, Vol. 14, No. 3, pp. 262-270, 1980.
- [7] D. Perzanowski, A. C. Schultz, W. Adams, E. Marsh, and M. Bugajska, “Building a multimodal human-robot interface”, *IEEE Intelligent Systems*, Vol. 16, No. 1, pp. 16-21, 2001.
- [8] M. T. Vo, and C. Wood, “Building an application framework for speech and pen input integration in multimodal learning interfaces”, *Proc. ICASSP*, pp. 3545-3548, 1996.
- [9] B. Xiao, C. Girand, and S. Oviatt, “Multimodal integration patterns in children”, *Proc. ICSLP*, pp. 16-20, 2002.
- [10] 김지영, 이경님, 홍기형 “심부름 영역에서의 음성과 터치제스처의 통합 인식”, *The 4th Technical Workshop*, Center for Intelligent Robotics, 2005.
- [11] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book* (for HTK Version 3.2), Entropic Cambridge Research Laboratory, 2002.
- [12] T. Anastasakos, and A. Gupta, “Integration patterns during multimodal interaction”, *Proc. ICSLP*, pp. 2293-2296, 2004.

접수일자: 2006년 6월 5일

게재결정: 2006년 6월 20일

▶ 김지영(Ji-young Kim)

주소: 136-742 서울특별시 성북구 동선동 3가 169-1 성신여자대학교

소속: 성신여자대학교 교육대학원 전자계산교육과

전화: 02) 928-9997

E-mail: kj2112love@sungshin.ac.kr

▶ 이경님(Kyong-Nim Lee)

주소: 121-742 서울시 마포구 신수동 1번지 서강대학교

소속: 서강대학교 대학원 컴퓨터학과

전화: 02)706-8954

E-mail: knlee@sogang.ac.kr

▶ 홍기형(Ki-Hyung Hong) : 교신저자

주소: 136-742 서울특별시 성북구 동선동 3가 169-1 성신여자대학교

소속: 성신여자대학교 미디어정보학부

전화: 02) 920-7525

E-mail: khhong@sungshin.ac.kr