https://doi.org/10.5392/JKCA.2017.17.02.096

대화 패턴 연구를 통한 스마트TV 음성 상호작용 모델의 탐구

Examination of a Voice Interaction Model for Smart TV through Conversation Patterns

최진해
LG전자

Jinhae Choi(mail.jinchoi@gmail.com)

요약

최근의 스마트 기기들은 사용자의 의도와 사용 맥락을 반영하는 지능형 에이전트의 형태로 발전하고 있으며, 기능을 더 쉽고 편리하게 활용할 수 있는 사용자 경험 설계가 경쟁력의 핵심이 되고 있다. 본 연구는 인간중심의 내추럴 인터랙션이 최적의 스마트TV 경험에 필수적이라는 전제하에 TV에 특화된 음성 인터랙션 방식을 탐구하고자 하였다. 사용자가 자연스러운 행태로 TV를 조작하는 모델을 구축하기 위하여 스마트TV의 주요 기능을 지능형 에이전트에 명령하도록 하였고 대화 패턴을 수집하였다. 수집된 문장은 CfA 모델에 대입하여 기능 실행을 위한 반응 별로 분류하였다. 분류된 5가지 대화 패턴은 스마트TV가 실행하는 기능 특성에 따라 '기능 실행'과 '정보 검색'으로 나눌 수 있었다. 사용자와 TV간의 음성 상호작용에서 모호한 요청의 경우 재확인을 위한 CfC1이 발생하고, 복합 의도나 조건부 요청에 대한 대응이 필요한 경우는 CfC2가 발생한다는 부분도 확인하였다. 본 연구의 결론은 스마트TV에서의 음성 UI 설계에서 Simple Request Type이 가장 효율적 모델이라는 점과 대화형 인터랙션은 가능한 사용자의 모호한 요청을 구체화하기 위한 단계에서만 활용되는 것이 적합하다는 것이다.

■ 중심어 : | 스마트TV | 언어행동 | 대화 패턴 | 대화 행위 모델 | 음성인식 | 음성 UI UX |

Abstract

As new smart devices are evolved into the intelligent agent who can reflect user intention and use context, user experience design for easy and convenient usability becomes a core competitive edge. Under the assumption that human centered natural interaction is necessary for the optimal smart TV experience, this study explores the types of voice interaction which are peculiar to TV watching context. In order to build a model for the users to naturally interact with Smart TV, conversation patterns were collected by requesting key features of Smart TV to intelligent agent. Collected sentences were applied to CfA model and classified by responses to activate features. The classified conversation patterns were divided into feature activation and information search. This study has identified that CfC1 occurred when voice interaction between Smart TV and users was vague and CfC2 occurred when the requests were complex or conditional. In conclusion, Simple Request Type is the most efficient model and voice interaction is more appropriate to use to clarify users' vague requests.

■ keyword : | Smart TV | Language Action | Conversation Pattern | Conversation for Action Model | Voice Recognition |

접수일자 : 2016년 10월 12일
수정일자 : 2016년 12월 06일
심사완료일 : 2016년 12월 22일
교신저자 : 최진해, e-mail : mail.jinchoi@gmail.com

I. 서 론

본 연구는 대화 패턴의 수집과 분류를 통해 인간 중심적 내추럴 음성 인터랙션을 스마트TV에 적용하여 자연스러운 사용 행태의 UI(User Interface)를 설계하기 위한 탐색적 연구이다. 스마트기기에 상용화된 최근의 음성 인터랙션 사례를 보면, 2011년 애플의 시리(Siri)가 단순 정보 제공과 검색 기능을 탑재하였고, 2014년에는 아마존의 에코(Echo)가 음악을 재생하거나 사용자의 질문에 답할 수 있는 가정용 음성인식 비서의 경험을 소개하였으며 이후, 당사에서 판매하는 상품을 주문까지 할 수 있는 커머스의 형태로 발전하였다. 또한 2016년에는 구글에서 구글홈(Google Home)을 소개하였는데 가정에 있는 다양한 스마트기기를 연결하여 음성으로 조정하거나 그 동안 축적해왔던 빅데이터를 활용하여 사용자에 최적화된 정보를 적시에 제공하는 인공지능형 시나리오를 소개하였다. 이는 단순한 음성 입출력 인식 기술의 정확도 차원을 넘어, 지능형 에이전트가 사용 맥락과 사용자 니즈를 이해하여 대화의 중단이나 포기 없이 복잡한 스마트 기능들을 음성으로 간단하게 조작할 수 있게 하는 궁극적인 UX 디자인의 청사진이라 할 수 있다. 이러한 산업계의 음성 인터랙션 발전 추세와 수요를 고려해볼 때, IPTV, Connected TV, VOD 스트리밍 서비스 등을 망라한 스마트TV 시스템 환경에서 음성 인터랙션은 더욱 중요한 역할을 맡을 가능성이 높다[1-3].

하지만 스마트TV는 스마트폰과 달리 개인용 디바이스가 아니기 때문에 맥락 기반 서비스(Context-based Service) 보다는 콘텐츠 기반의 서비스가 중심이 된다는 점에서 TV만의 특수한 사용 행태가 반드시 반영될 필요가 있다[4][5]. 구체적으로 스마트TV의 음성 인터랙션을 연구하기 위해서는 다음의 몇 가지 TV 특성들을 명확히 전제할 필요가 있다. 첫째, TV는 콘텐츠 소비를 위한 기기이다. TV는 전원을 켜는 동시에 방송이 재생되기 때문에 사용자는 에이전트와의 인터랙션에 집중하기 힘들 뿐 아니라 TV 입장에서도 콘텐츠의 음향이 노이즈로 작용하여 음성인식에 어려움이 따른다. 또한 사용자는 TV 앞에서 수동적인 Lean-back 성향을 보이기 때문에 스마트폰과 달리 빈번한 인터랙션이 일어나지 않는다. 둘째, TV는 공동 사용 기기이다. 스마트폰이 개인정보와 컨텍스트에 바탕을 둔 다양한 서비스에 강점을 보이는 것과 달리 로그인을 통한 개인화 서비스에는 적합하지 않아 핵심 영역에 차이가 있다.

이에 본 연구에서는 스마트 TV 특성을 고려한 가운데 대화의 패턴을 수집하였으며 어휘, 형태소별로 구분하였다. 대화 패턴은 요청(Request)과 반응(Response)에 따라 기기가 실행해야 하는 기능별로 분류하였다. 이 과정에서 대화가 끊길 확률이 높은 지점이 어디이고 TV는 어떤 반응을 통해 사용자 의도를 파악할 수 있는지 CfA (Conversation for Action) 모델에 대입하여 탐색해 보았다.

II. 이론적 모델

대화형 인터랙션의 이론적인 근거는 언어/행위 관점(Language/Action Perspective) 연구가 가장 대표적이다[6][7]. Flores와 Winograd에 의해 시작된 이 연구는 언어적 의사소통의 과정을 정보 시스템 설계에 도입하기 위하여 CfA(Conversation for Action) 모델을 적용하였다[8]. 이 모델에 따르면 언어는 인간의 모든 협업 행위에서 가장 근본적인 요인이며, 이러한 언어-행위 관점이 모든 CSCW (Computer-Supported Cooperative Work) 시스템 개발에서 매우 중요한 역할을 수행한다. 언어-행위 관점에서 초점을 두는 것은 언어의 의미와 사용이 실제 업무를 수행하는 형태, 즉 대화의 구조이다. 대화 구조의 기본 요소는 요청(Request)와 응답(Response)이다. 대화 참여자 중 한편이 상대편에게 요청을 하면, 상대편은 일련의 차후 행위를 예상하고 수락, 거부, 또는 수정제안의 세 가지 형태로 응답하게 된다. 이 과정은 순환적으로 진행되면서 대화 참여자간에 상호이해가 형성되며 이에 기반하여 의미 있는 협업 행위가 발생하게 된다는 것이다. Winograd[8]는 CfA 모델을 대화형 인터랙션의 개념적 기본 구조로 제시하였지만 본 연구의 목적은 스마트TV라는 특정기기에 적용 가능한 경험 모델 도출을 목

적으로 하고 있다. 따라서 가상의 지능형 에이전트를 통해 실제 스마트TV를 조작하는 상황을 자연스럽게 연출하면서 사람과 기기 간에 이루어지는 대화의 특성을 관찰하고 시사점을 도출하기 위한 실험이 필요하다. 이에 역할 수행(Role playing) 기법[9][10]을 적용하여 한 사람이 인간리모컨-대화로 사용자의 의도를 듣고, 대신 TV를 조작해주는 지능형 에이전트의 역할을 수행하도록 하고 실험 참여자는 친구나 가족을 대하듯이 음성으로 편하게 인간리모컨을 조작하도록 진행하였다.

III. 1차 연구: 대화 패턴 모델 탐색

3.1 스마트TV에 대한 요청의 분류

스마트TV 컨트롤 연구[11]에서 정의한 주요 기능 즉, 채널 재핑(Zapping), 단축 메뉴, 콘텐츠 검색, 로그인, 실시간 정보를 위한 빠른 검색, 메모, 북마크, SNS, 인터넷 사용 시 글자입력, 줌/스크롤, 커서 이동, 특화 기능 대응 인터페이스, 특화 콘텐츠 실행 등을 남녀 총 20명(남성 12명, 여성 8명)에게 제시하였고 가상의 지능형 에이전트가 탑재된 음성인식 TV에 대화 형태로 자유롭게 기능 수행을 요청하도록 하였다. 최대한 실제 사용 환경과 유사하게 진행할 것을 안내했기 때문에 참가자들은 위에서 제시한 기능만을 수행하지 않았고 갑자기 다른 기능을 요청하는 상황도 관찰되었다. 한 사람당 총 90분이 소요되었고 모든 과정은 분석을 위해 녹취하였다. 그 결과 요청한 내용은 TV전원 켜기부터 끄기까지의 총 22종이 수집되었고 다시 이 요청들은 기기가 실행해야 하는 기능의 전환을 기준으로 5가지 군집으로 분류되었다[그림 1]. 전원 켜기, 채널변경, 전원 끄기, 볼륨 크게/작게, 음소거 등 한 번의 음성 명령으로 원하는 목적 달성이 가능한 경우를 [Direct Command], 명확히 원하는 컨텐츠의 제목을 알고 있거나 목록을 불러내 그 중에 선택하면 실행되는 것을 [Menu tree 탐색], 사용자의 기호에 따라 원하는 컨텐츠를 추천해 주는 경우를 [추천 유도], 컨텐츠의 인기도나 배우에 대한 정보와 같이 시청 중 부가 정보를 요청하는 경우를 [시청 중 외부 정보 요청], 빨리 감기/되돌리기/일시정지 등 콘텐츠 자체를 컨트롤하거나 TV화질 조절 등 시청 컨디션을 조정하는 경우를 [시청 컨디션 조정]으로 분류하였다. 이 각 분류에 따른 대화의 특성을 보면 [Direct command] 와 [시청 컨디션 조정] 요청의 경우 리모컨 상에 존재하는 버튼을 말로 대신 하거나, '다음' 과 같은 단순한 명사구 형태의 명령을 반복하는 일방적 명령의 성격이었다. 이 경우 가상의 지능형 에이전트는 단순히 명령된 동작을 수행하면 요청이 완료되었다. 반면, [Menu Tree 탐색], [추천 유도], [외부 정보 요청] 의 경우 '~해줘' 형태의 명확한 요청형 어미를 사용하거나 '~가 무엇 / 누구 / 언제 / 어디지?', '저것, 이것, 저 사람'과 같이 의문사나 지시어 등 문장 구성요소를 포함하는 경향을 보였다. 이때 가상의 지능형 에이전트는 '네 알겠습니다', '잠시만요', '찾아볼까요?', '다시 한 번 말씀해 주세요' 등 현재의 요청에 대하여 명확히 알아들었는지의 여부와 기능 수행을 명시적으로 알려 줬을 때 참가자들은 자연스러운 대화 패턴으로 이해하였다. 분류 결과를 살펴보면 [Direct Command]나 [시청 컨디션 조정]과 같이 사용자가 결과에 대해 명확히 알고 있으며 명시적인 지시가 가능한 단순 조작 명령 방식의 대화와 [Menu Tree 탐색], [추천 유도], [시청 중 외부 정보 요청]과 같이 어떤 결과가 나올지 예측하지 못한 상태에서 개방형 질문을 던지는 2가지 대화 패턴으로 다시 분류될 수 있다.

A.	TV 전원 켜기	Direct Command
B.	채널을 차례로 돌리기	
C.	특정 채널로 바로 이동하기	
D.	특정 외부입력기기를 선택하여 실행하기	
E.	볼륨 변경하기	
F.	음소거 하기	
G.	음소거 해제하기	
H.	TV 전원 끄기	
I.	특정 채널을 선택하여 콘텐츠를 검색해 보기	Menu Tree 탐색
J.	특정 조건의 콘텐츠 리스트를 불러내 고르기	
K.	특정 콘텐츠 바로 찾아보기	
L.	채널 정보 보기	
M.	시청중 콘텐츠 정보 확인하기	
N.	콘텐츠 추천 받기	추천 유도
O.	콘텐츠 관련 외부정보 얻기	시청 중 외부정보 요청
P.	콘텐츠에 관한 다른 사람의 의견 묻기	
Q.	시청 중 콘텐츠 의미 관련 정보 검색하기	
R.	SNS 하기	
S.	콘텐츠 구간 점프하기 (빨리 스캐닝 하기)	시청 컨디션 조정
T.	콘텐츠 잠시 멈추기	
U.	콘텐츠를 다른 매체로 전송하기	
V.	TV 설정 하기	

그림 1. 스마트TV 음성 상호작용 요청 분류

3.2 분석 프레임 워크

앞에서 이론적 모델로 대표적 대화 패턴 모형인 Winograd의 CfA(Conversation for Action)모델[7][8]에 대입해 봄으로써 분석을 진행하였다[그림 2]. 원래 그가 제안한 CfA모델은 컴퓨터 시스템에서 명령에 대한 처리 및 반응을 위한 UI 설계를 위해 고안된 것으로, 대화의 요청이 1번부터 시작하여 5번의 목적 달성이 되기까지 중간에 발생할 수 있는 반응들을 파악하고 어떻게 5번까지 효율적으로 도달할 수 있는지를 파악하는 것이다. 그의 CfA모델은 일반 명령에 대한 거부나 임의적 미수행(예: 시스템의 Withdraw, Renege) 등의 경우가 포함되어 있다. 하지만 본 연구에서는 지능형 에이전트가 탑재된 스마트기기에서 발생하지 않아야 할 경우로 간주하였다. 따라서 7, 8, 9번의 경우는 고려하지 않았다.

사용자의 음성 요청에 대한 분석 방법은 녹취한 데이터를 품사 별로 구분하고 어떤 형태소(태깅)가 CfA의 어느 단계에서 순차 흐름에 영향을 주는지 파악하는 방법으로 진행하였다. 이때 순차 흐름을 파악하는 것이 주요 목적이기 때문에 문법적인 접사나 어순, 전사된 말뭉치의 억양은 고려하지 않았다. 형태소 분석은 사람의 말뭉치를 어절 별로 분해한 것인데, 구분을 위한 태깅 작업은 국어 특수자료 구축 기준에 따라 진행하였다[9]. “지난주 방영한 무한도전 틀어줘”라는 요청의 경우 ‘지난주/일반부사(MAG)+방영/일반명사(NNG)+하/동사파생접미사(XSV)+무한도전/고유명사(NNP)+틀다/(동사)VV+주다/동사파생접미사(XSV)’의 형태로 태깅이 가능하고 가상의 에이전트는 지난주/일반부사(MAG), 무한도전/고유명사(NNP), 틀다/동사(VV)를 이해하고 실행에 옮길 수 있다. 한편 “무료 영화 리스트 보여줘”라는 요청의 경우 ‘무료/일반명사(NNG)+영화/일반명사(NNG)+리스트/일반명사(NNG)+보다/동사(VV)+어/연결어미(EC)+주다/동사(VV)+어/종결어미(EF)’로 태깅되지만 명확한 고유명사(NNP)가 없기 때문에 액션 영화를 선호하는지 코미디 영화를 선호하는지 에이전트는 재확인 절차가(Counter)가 필요하다. 또한 ‘무한도전/대명사(NNP)’을 ‘무도’라고 요청하는 경우도 이에 해당하는데 에이전트가 명확하게 이해하기 위한 단계가 2번과 3번에 발생 한다.

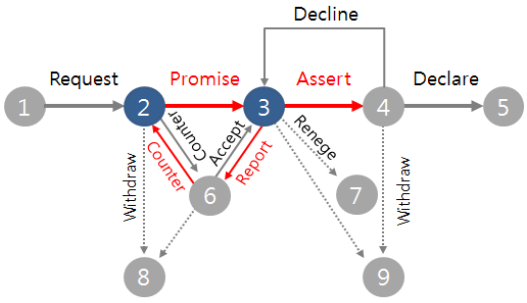


그림 2. 스마트TV 지능형 에이전트를 위한 CfA (Conversation for Action) 모델

20명의 참가자가 요청한 문장은 총 336회였으며 형태소에 따른 분류는 일반명사(NNG) 19.8%, 동사(VV) 15.7%, 일반부사(MAG) 14.1%, 종결어미(EF) 12.5%, 대명사(NP) 6.1%, 명사(NN) 5.2%, 접속부사(MAJ) 4.3%, 보조사(JX) 4.3%, 부사(MA) 4.1%, 주격조사(JKS) 4.1%, 감탄사(IC) 3.4% 순으로 빈도가 나타났다. 이를 다시 품사 기준으로 묶어보면 동사/어미가 34.6%, 명사/대명사가 24.8%, 수식어인 부사가 22.5%로 전체의 81.9%를 차지한다. 따라서 스마트TV 조작을 음성으로 요청할 경우 지능형 에이전트는 동사/어미, 명사/대명사, 부사를 CfA의 2번 단계에서 가장 많이 이해하는 것이 필요하다고 할 수 있다. 만약 그 요청이 바로 실행으로 옮겨질 수 있을 경우에는 요청이 종료되지만 그렇지 않을 경우에는 6번, 혹은 3번 단계에서 동사, 명사, 부사를 재확인하여 이해하는 것이 기능 실행에 불가결한 것을 알 수 있다.

3.3 스마트TV 특화 CfA 패턴

수집된 336회의 사용자의 요청들을 CfA모델에 대입해 보면 [그림 3]과 같이 다섯 가지 대표 패턴으로 정의할 수 있다[14-16]. 그 첫 번째는 [Simple Request Type]으로 사용자는 원하는 바를 단도직입적으로 요청하고 TV는 즉각적인 결과를 화면을 통해 보여주는 경우이다. 이 패턴은 가장 많은 총 186회(55.5%)의 요청이 있었는데 스마트TV 대화형 인터페이스에서 즉각적 실행이 가장 많다는 것은 중요한 시사점이다. 그 다음

은 총 56회(16.6%)의 요청이 있었던 [Promise Type]이다. 이는 TV를 시청하는 동안 배우나 소품의 정보습득을 목적으로 하는 경우로 지능형 에이전트가 사용자를 대신해서 정보검색을 수행하도록 요청한 경우이다. [Decline Loop Type]은 볼륨이나 채널을 바꾸는 경우에 해당하는데 총 47회(13.8%)의 빈도수를 차지하였다. 리모컨 버튼을 반복적으로 누르는 기존의 행위를 음성으로 대신한 결과, 적당한 결과 값을 얻을 때까지 요청을 반복할 수밖에 없는 경우이다. [Assert Declare Type]은 총 28회(8.3%) 나타났는데 [Promise Type]과 유사하게 정보습득을 목적으로 하지만 사용자가 정보를 습득한 이후에 대화를 능동적으로 종결하는 경우이다. 마지막으로 [Assert Type]은 총 19회(5.5%)로 가장 빈도수가 적었다. [Assert Declare Type]과 [Assert Type]의 빈도수가 적은 것은 가상의 지능형 에이전트를 사용하는 실험의 한계일 수도 있지만 기계와 대화할 때는 복잡한 기능 실행을 단축하기 위하여 음성 요청을 하는 경우가 많기 때문으로 해석할 수 있다.

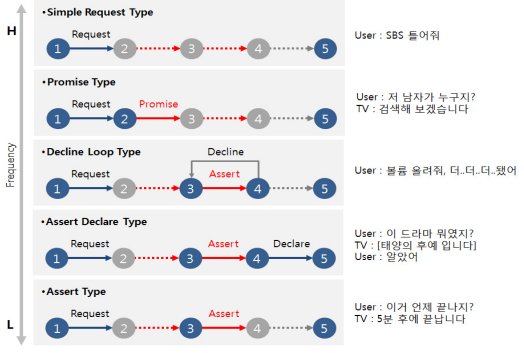


그림 3. 사용자와 지능형 에이전트간의 대화 패턴

3.4 CfC (Conversation for Clarification)

1차 연구에서 실시한 대화 패턴 모델 분류의 특징은 TV 화면에서 콘텐츠가 나오고 있는 중에도 음성 피드백이 '인간리모컨의 입' 이라는 별도의 채널을 통해 제공되었기 때문에 원활한 대화형 인터랙션이 가능했다는 것이다. 하지만 보다 실제 상황을 고려한 이후 연구를 위해서는 지능형 에이전트와 TV 콘텐츠가 동일한 음성채널을 공유해야 하는 상황에 대한 고민이 필요하다. 대화형 인터랙션에 관한 또 다른 상황은 사용자의 발화를 지능형 에이전트가 명확하게 이해하지 못해 목표를 달성하는 데 시간이 지체되는 상황에 대한 것이었다. 대화패턴의 분석을 통해 이 부분에 관한 심도 깊은 고찰이 진행되었는데, TV가 사용자의 요청에 대해 적절한 행동을 할 수 없어 지체되는 경우를 다음과 같이 세 가지로 구분할 수 있었다. 첫째는 사용자의 명령어가 적절치 않거나 소음으로 인해 요청 자체를 명확히 이해하지 못하는 경우이다. 둘째, 요청은 명확하게 이해하였으나 결과를 구체화할 정보가 조금 더 필요한 경우이다. 셋째, 요청을 명확히 이해하였고 결과를 구체화할 정보가 충분하나 해당 서비스를 지원하지 않는 경우이다[그림 4].

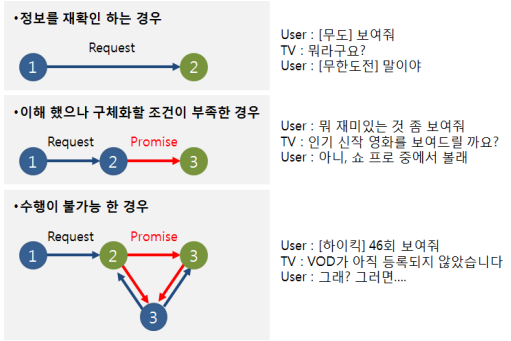


그림 4. 대화 지체의 사례

지체가 일어나는 부분은 녹색 2번과 녹색 3번 단계로 여기가 바로 사람과 사람사이의 대화와 유사한 형식의 대화형 인터랙션이 필요한 영역이라고 볼 수 있다. 이 영역에서는 지능형 에이전트가 사용자에게 명령을 반복해 주기를 요청하거나 부가적인 정보를 부탁 또는 유도하는 등의 행동이 일어나는데 Winograd는 이와 같은 성격의 대화 행위를 명료화 대화(Conversation for Clarification)라고 구체화한 바 있다[7]. 요청 자체를 명확히 하는 녹색 2번 단계에서의 CfC를 'CfC1', 결과 구체화를 위해 정보를 추가 수집하는 녹색 3번 단계에서의 CfC를 'CfC2'라 정의할 때, CfC1와 CfC2를 얼마나 부드럽게 처리하느냐에 따라 사용자가 느끼는 대화의 자연스러움이 달라질 것이라는 점을 유추할 수 있다.

CfC1의 경우 사용자로 하여금 시스템이 이해할 수 있는 음성 명령어를 사용하도록 유도함으로써 기존 리모컨보다 신속한 조작이 가능하다는 데에 의미가 있다. 예를 들면 “프로그램 순서 나온 표 보여줘” 와 같은 명령이 있을 때 “네, 편성표를 보여드리겠습니다” 라는 응답을 통해 사용자가 ‘편성표’라는 정확한 단어를 학습하도록 유도하여 점점 신속하고 정확한 피드백을 제공할 수 있다. CfC2의 경우는 짧아질수록 좋은 CfC1과는 달리 특정한 목적이 없이 볼만한 콘텐츠를 찾아가는 과정에서 의미를 찾을 수 있다. 즉, 사용자에게 의사결정을 도와주는 비서와 같은 역할을 수행하면서 대화 인터랙션의 장점을 극대화할 수 있을 것으로 보인다. 예를 들면, “뭐 재미있는거 없나?” 라는 명령에 대해 “주말 동안 놓친 쇼프로와 드라마가 있습니다. 리스트를 보여드릴까요?” 와 같은 대화를 통해 볼만한 콘텐츠를 함께 구체화 할 수 있다.

IV. 2차 연구: CfC 음성 대화 모델 적용

4.1 태스크 별 대화 모델

CfC2 단계에 해당하는 정보 요청 방법과 에이전트가 대응하는 흐름을 고려하여 3가지 가설적인 모델을 [그림 5]와 같이 구분해 보았다. 첫 번째는 ‘최적 안 수행(DO)’ 형으로 가령 사용자가 “종료”라고 말하면 에이전트는 즉각 이해하고 “TV를 종료합니다”라는 피드백과 함께 기능 실행에 옮기는 것이다. 이는 사용자 요청에 대하여 동사/어미, 명사/대명사, 부사를 명확하게 이해한다는 가정하에 가장 효율적인 모델이다. 또한 기계와 많은 대화를 하는 것이 어색한 인간의 특성상, 한 번에 기능실행이 된다는 측면에서 가장 효과적이기도 하다. 하지만 이때 에이전트는 마치 곁에서 모든 상황을 이해하고 있는 집사처럼 간단한 넛지(Nudge)만으로 사용자를 이해할 수 있어야 하기 때문에 설계에서는 사용자의 컨텍스트에 대한 많은 정보가 필요한 모델이라 할 수 있다. 두 번째는 ‘선택지 제시(Which)’형으로 TV가 복수의 선택지를 제시하고 이 중 하나를 사용자가 선택하도록 하는 방식이다. 사용자의 요청에 대하여 이해가 명확하지 않을 때인데, 가령 ‘종료’에 대한 요청에 대하여 사용자에게 TV를 종료할 것인지 셋톱박스를 종료할 것인지 등 몇 가지 조건을 제시하고 기능을 실행하는 모델이다. 에이전트 입장에서는 사용자를 학습할 수 있는 좋은 방법이지만 즉각적이지 못한 에이전트의 대응은 불편하게 느껴질 수 있다. 마지막으로 ‘상세화 질문(What)’ 형은 사용자에게 무엇을 할지 답을 직접 물어보는 모델이다. “종료”라고 요청했을 경우 “무엇을 종료할까요?”라고 되물어 보는 것이다. 명확하게 사용자의 의도를 파악하고 다음으로 진행할지 말지에 대한 판단을 사용자에게 물어보는 방식이다.

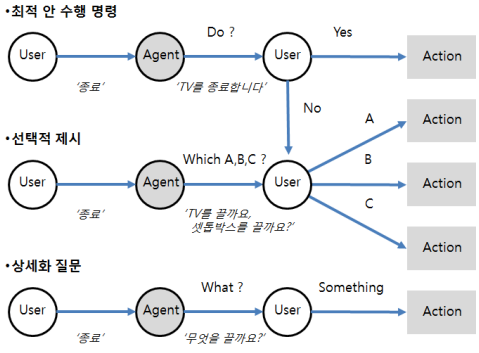


그림 5. CfC2의 3가지 가설적 모델

4.2 모델 적용 확인

스마트TV의 음성 UI를 설계하기 위해서 CfC2의 3가지 가설적 모델에 대해 실제 상황에서는 어떻게 요청하고 어떻게 반응해야 하는지 확인하기 위하여 실제 TV와 사람간의 대화에 적용해 볼 필요가 있다. 이를 위해 피실험자가 TV에 명령을 내리고 원격지에서 그 말을 듣고 조작을 하여 실제 음성 인식 TV를 사용하는 경험을 제공하는 오즈의 마법사 시뮬레이션 기법[17]을 실시하였다. 스마트TV나 Siri 등의 사용 경험이 없어 아직 음성 인터페이스에 대한 멘탈모델이 형성되지 않은 20~30대 남·녀 총 8명(남성 4명, 여성 4명)이 실험에 참여하였다. 진행자가 참가자에게 정해진 상황 8가지를 제시하고 참가자는 그 상황에 맞는 요청을 음성으로 TV에게 전달하도록 하였다. TV를 조종하는 원격지의 오퍼레이터는 사용자에게 요청 받은 사항을 마치 TV

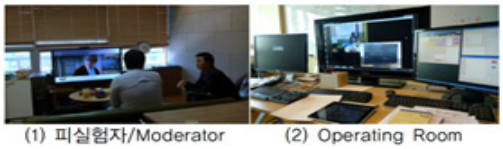
가 하는 것처럼 TTS(Text To Speech)로 대답하고 결과를 보여주는 방식으로 진행되었다. 대표적인 태스크는 다음[표 1]과 같으며 한 사람 당 60분이 소요되었다.

표 1. 모델 적용을 위한 태스크

	Selecting	Consuming	Managing
조작	채널 변경 콘텐츠 탐색 검색, 추천		볼륨 조절 재생 속도 변경 재생 위치 변경
정보	콘텐츠 선택을 위한 정보 요청	콘텐츠 관련 정보 요청	

4.3 프로토타입 구성

모델 적용을 위한 프로토타입은 크게 두 파트로 나눠 진행하였다. A파트는 피실험자가 마주하는 TV와 관찰 장비에 관련된 부분으로 1) 스마트TV, 2) 관찰용 웹캠, 3) 마이크, 4) 원격으로 TV를 조작하기 위해 Arduino를 이용해 제작한 IR 리모컨, 5) TTS를 재생하기 위한 스피커, 6) VOD에 대응하기 위한 IPTV 셋톱박스로 구성되었다. B파트는 오퍼레이터(Wizard)가 사용자의 요청에 맞게 TV를 조종하기 위한 부분으로 1) 사용자의 요청 사항에 대한 피드백을 주기 위해 검색 결과를 GUI로 표현해주는 소프트웨어, 2) 원격 리모컨을 조종하기 위한 소프트웨어, 3) 사용자의 요청에 피드백을 주기 위한 TTS 소프트웨어, 4) IPTV를 조종하기 위한 타블렛과 리모컨 앱, 영상을 녹화하기 위한 소프트웨어를 포함하였다[그림 6]. TV가 사용자의 음성 요청에 대응하는 방식은 TTS를 이용해 기계 음성으로만 대답하여 현실감 있게 하였다.



(1) 피실험자/Moderator (2) Operating Room

그림 6. 프로토타입 구성

4.4 CfC에 대한 분석

참가자 8명이 음성으로 요청한 13건을 분류해 본 결과, 요청의 목적과 요청의 방식에 따라 네 가지 경우로 나눠볼 수 있었다[표 2]. 특히 네 가지 경우 중 사용자가 정보를 요청하기 위한 목적으로 모호한 질문을 할 때 CfC1이 발생하고 의도보다 적은 정보를 담아 요청을 하거나 복잡적인 의도를 담아 요청을 할 때 모자란 정보를 알아내기 위해 CfC2 가 발생한다는 점을 확인할 수 있었다.

표 2. CfC1과 CfC2의 구분

	조작 목적	정보 목적
지시	구체적	모호함 (다의적) → CfC1 발생
의도	단순 (요청=의도)	복합 (요청 < 의도) → CfC2 발생

<CfC1 예시>
- 사용자 : 유재석이랑 김종국이 같이 나온 예능프로 뭐가 있더라?
- TV : 런닝맨 말씀이신가요?
- 사용자 : 응 그거 보여줘

<CfC2 예시>
- 사용자 : 지금 하는 야구경기 없어? (없으면 야구에 대한 다른 걸 제안 해줘)
- TV : 지금은 방송 중인 야구 경기가 없습니다. 이후에는 KBS에서 11시에 방송합니다. 예약할까요?
- 사용자 : 응 그거 예약해줘

CfC2의 경우 최초 설계한 [그림 5]의 3가지 대응 방식에 대해 각각의 만족도를 조사해 보았다. 조사는 개인별 심층인터뷰로 진행 했으며 그 결과 동일한 조건일 경우 '최적 안 수행 명령'과 '선택적 제시' 형태가 '상세화 질문' 보다 만족도가 8명 모두 높았다. 이는 '최적 안 수행 명령'과 '선택적 제시'의 대응 방식이 '상세화 질문' 방식보다 선택 비용이 상대적으로 더 낮기 때문으로 해석 할 수 있다. '최적 안 수행 명령'과 '선택적 제시' 간의 비교 결과는 8명 중 7명이 '최적 안 수행 명령'을 선호한 것으로 나타났다. 이는 사용자가 명확한 목적을 간략히 전달하고 바로 수행 할 수 있는 방식이 복수의

선택지에서 하나를 고르는 객관식 방식보다 선택 비용이 더 낮기 때문으로 해석할 수 있다. 이 결과를 통해 사람들은 TV 앞에서는 선택 비용을 최소화 할 수 있는 방향으로 인터랙션 하는 것을 선호한다는 것을 알 수 있다[표 3].

표 3. CfC2 유형별 특성

질문 유형	응답	선택 비용
최적 안 수행 확인	Yes / No	낮음
선택지 제시	A, B, C	높음
상세화 질문	Something	매우 높음

하지만 선택 비용에 따라 응답 방식을 결정하는 설계에는 한 가지 고려할 사항이 있다. 선택 비용이 가장 낮은 '최적 안 수행 확인' 유형의 요청 시, 에이전트가 이해를 한 번에 못하거나 대안을 제시했을 때도 사용자가 거부하면 다시 처음부터 요청을 해야 하는 중복이 발생한다. 이 경우 '선택지 제시'나 '상세화 질문' 유형을 선택했을 때 보다 결과적으로 더 많은 비용과 시간이 소요된다. 따라서 가장 좋은 설계 방향은 선택 확률을 계산해서 일정 수준 이상일 경우 '최적 안 수행 확인' 유형으로 제시하고 확률이 낮을 경우엔 다른 유형을 따르는 것이다.

V. 결 론

본 연구에서는 가상의 지능형 에이전트와의 사용자 간의 대화 인터랙션 패턴을 수집하여 CfA 모델에 대입하였다. 그 결과 조작 목적 중심으로 Simple Request Type, Decline Loop Type과 정보 습득 목적 중심으로는 Assert Type, Assert Declare Type, Promise Type으로 분류되었다. 사용자와 스마트TV 간의 음성 인터랙션은 간단한 명령을 즉각적으로 실행하는 Simple Request Type이 가장 빈번하게 일어나지만 사용자의 의도가 명확하지 않는 경우는 지연이 발생하게 된다. 이 과정에서 사용자의 의도를 명확히 파악하기 위한 CfC가 발생하는데 모호한 요청에 대한 대응이 필요할 경우 CfC1이 발생하고 복합 의도나 조건부 요청에 대한 대응이 필요할 경우는 CfC2가 발생한다. 이러한 대화 패턴의 수집과 분류를 통하여 얻어진 탐색의 시사점은 다음과 같다.

첫째, 스마트TV에서 음성 UI를 적용할 경우, [Simple Request Type]화 되는 것이 가장 좋다. 스마트TV라 하여도 영상 소비가 주목적이고 음성 UI를 사용한다 하여도 콘텐츠 탐색이나 정보 검색이 목적이기 때문에 가장 간단하고 효율적인 방법이 우선시 되어야 한다. 둘째, 대화형 인터랙션은 명령어 입력(Command Input)보다는 사용자의 의도를 구체화하여 소비할 콘텐츠를 결정해가는 탐색 단계에 활용되는 것이 적합하다. 이 과정에서 시스템이 이해하지 못했을 경우 발생하게 되는 CfC1, 이해는 했으나 구체화할 조건이 부족하여 발생하는 CfC2 단계에서의 처리가 내추럴 음성 인터랙션 경험에 중요하다는 것을 알 수 있었다. 사용자의 모호한 의도를 파악하기 위한 CfC2의 3가지 방식에 대한 선호도를 조사한 결과, '최적 안 수행 확인' 방식을 선호한다는 것은 사용자가 가능한 기기와 대화를 기피한다는 것을 알 수 있는 시사점이다. 사용자는 본인의 의도를 모두 '말'에 담아서 요청하지 않기 때문에 TV는 부족한 정보를 채워나가면서 사용자 의도에 도달해야 한다. 이때 모호한 요청을 극복하기 위해서 TV가 스스로 시청 패턴이나 선호도 등 사용자 의도에 가깝게 콘텐츠를 추천해 준다면 사용자는 낮은 선택 비용으로 원하는 목적을 달성할 수 있고 이 과정에서 만족감을 줄 수 있을 것이다. 향후 스마트TV의 큐레이션이 사용자의 의도에 가깝게 접근하기 위해서는 스마트폰이나 IoT 기기 등을 활용하는 지능형 에이전트를 구축해야 할 것이다. 본 연구를 통해 얻어진 스마트TV에서의 음성 입력 경험 모델을 UI 개발에 적용하여 사용자의 만족감을 높일 수 있는 방안에 대한 구체적인 실용적 성과로 연계되어야 할 것으로 보인다.

참 고 문 헌

[1] 김종진, 김준석, 김정희, 이현아, 서희철, "음성언어기술 기반 대화형 질의응답 시스템 개발 동향,"

전자공학회지, 제41권, 제3호, pp.77-91, 2014.
[2] 문영찬, 김선정, 김진, 고영웅, “홈오토메이션 환경에서 스마트 TV 통합 인터페이스 설계 및 구현,” 한국정보기술학회지, 제11권, 제2호, pp.87-94, 2013.
[3] 이혜민, 김승인, “음성인식 기반의 모바일 지능형 개인비서 서비스 사용성 비교 - Samsung S 보이스와 Apple 시리를 중심으로,” 디지털디자인학연구, 제14권, 제1호, pp.231-240, 2014.
[4] 박전규, 정훈, 정의석, 강병옥, 박기영, 오유리, 이윤근, “스마트TV를 위한 음성인식 서비스 시스템의 구현,” 대한전자공학회 하계종합학술대회, pp.1856-1857, 2013.
[5] 변대호, “감정표현어를 이용한 스마트TV의 사용자경험 평가,” 한국콘텐츠학회논문지, 제15권, 제5호, pp.132-141, 2005.
[6] D. Te'eni, “The language-action perspective as a basis for communication support systems,” Communications of the ACM, 제49권, 제5호, pp.65-70.
[7] T. Winograd and F. Flores, *Understanding Computers and Cognition: A New Foundation for Design,* Addison-Wesley, 1986.
[8] T. Winograd, “A Language/Action Perspective on the Design of Cooperative Work,” Human-Computer Interaction, Vol.3, No.1, pp.3-30, 1987.
[9] G. Seland, "System designer assessments of role play as a design method: a qualitative study," NordiCHI '06 Proceedings of the 4th Nordic conference on Human-computer interaction: Changing roles, pp.222-231.
[10] K. T. Simsarian, "Take it to the next stage: the roles of role playing in the design process," CHI '03 Extended Abstracts on Human Factors in Computing Systems, pp.1012-1013, 2003.
[11] 조성일, 홍사윤, 홍지영, 양경인, 장구앙, 최진해, "Operation Command 에 따른 스마트 TV 입력방식 연구," 한국HCI학회 학술대회, pp.631-634, 2012.
[12] J. H. Choi and J. Y. Hong, “The natural way of gestures for interacting with smart TV,” Journal of the Ergonomics Society of Korea, Vol.31, No.4, pp.567-575, 2012.
[13] 임용기, *21세기 세종 계획: 국어 특수자료 구축 연구보고서*, 국립국어원, 2005,
[14] 양경인, 여준희, 이현철, 장세훈, 최지호, 최진해, “인간 대화패턴 연구를 통한 CFA 경험 모델 도출,” 2013 한국 HCI학회 학술대회, pp.36-39, 2013.
[15] A. Dix, J. Finlay, G. D. Abowd, and R. Beale, *Human-Conputer Interaction, 3rd Ed.*, Pearson, 2004.
[16] D. Twitchell, M. Adkins, J. Nunamaker, and J. Burgoon, “Using speech act theory to model conversations for automated classification and retrieval,” Proceedings of the 9th International Working Conference on the Language-Action Perspective on Communication Modelling, pp.121-129, 2004.
[17] Scott R. Klemmer, Anoop K. Sinha, Jack Chen, James A. Landay, Nadeem Aboobaker, and Annie Wang, "Suede: a Wizard of Oz prototyping tool for speech user interfaces," UIST '00, Proceedings of the 13th annual ACM symposium on User interface software and technology, pp.1-10, 2000.

저 자 소 개

최 진 해(Jinhae Choi) 정회원

- 2008년 3월 : Chiba University 인간환경디자인과학(공학박사)
- 2013년 6월 ~ 현재 : LG전자 MC연구소 UX실장

<관심분야> : UX Design, User Interface, Human Centered Design, HCI, Design System