Listeners' Boredom: How Long Can the Voice Information of a Smart Speaker be?

Huhn Kim[1*], Jaeyeong Ko[2], Seungwan Kim[2], Hoyeon Hwang[2]

[1]Department of Mechanical System Design Engineering, Professor, Seoul National University of Science & Technology, Seoul, Korea

[2]Department of Mechanical System Design Engineering, Student, Seoul National University of Science & Technology, Seoul, Korea

Abstract

Background If the length of voice information provided by a smart speaker is too long, the user may feel bored. On the contrary, even if the length of voice information is too short, the user may feel that the amount of information is insufficient. Therefore, in the case of long voice information provided by a smart speaker such as encyclopedia information, a guideline is needed for the length such that the user feels that there is a sufficient information amount without being bored.

Methods An experiment was conducted to hit a bell when the participant felt enough information or boredom while listening to long sentences with a different number of words per sentence and speech speeds. We also analyzed the length of voice information at the time of hitting the bell.

Results As a result of the experiment, the content of information, average length of sentences, and speed of speech did not affect the time when the user hit the bell. The length of voice information was about 300 Korean characters for participants that felt sufficient information without being bored. The length of voice information was about 450 Korean characters for participants that felt bored.

Conclusions The results of this study can be used as a guideline for determining the length of information such as encyclopedias, news and weather provided by smart speakers.

Keywords Smart Speaker, Voice UI, Conversational Design, TTS Length

This work has been conducted with the support of SK Telecom and the "Project for Nurturing Advanced Design Professionals" initiated by the Ministry of Trade, Industry and Energy of the Republic of Korea.

*Corresponding author: Huhn Kim (huhnkim@seoultech.ac.kr)

Citation: Kim, H., Ko, J., Kim, S., & Hwang, H. (2010). Listeners' Boredom: How Long Can the Voice Information of a Smart Speaker be?. *Archives of Design Research, 33*(1), 151-163.

http://dx.doi.org/10.15187/adr.2020.02.33.1.151

Received : Sep. 04. 2019 **; Reviewed :** Nov. 21. 2019 ; **Accepted :** Nov. 22. 2020
pISSN 1226-8046 **eISSN** 2288-2987

1. 연구의 배경 및 목적

인공지능과 음성인식 기술의 발전으로 2011년 애플 아이폰의 음성 비서인 쉬리가 등장하였고, 2015년에는 아마존에서 첫 스마트스피커인 알렉사를 출시하였다. 그 후 제조사나 통신사들은 다양한 스마트스피커를 개발하여 경쟁적으로 출시하고 있다. SKT의 누구, 네이버의 클로바, 카카오의 미니, KT의 기가지니, 구글의 구글홈, 그리고 애플의 홈팟 등이 대표적이다. 마켓연구보고서(Market Research Report, 2018)에 따르면, 전 세계 스마트스피커의 시장 규모는 2017년에 22억 달러였으며, 매년 38.7%의 성장률을 보일 것이라 한다. 현재 스마트스피커 시장의 경쟁이 거센 이유는 향후 스마트스피커가 네트워크로 연결되어 있는 가정 및 사무실 내의 여러 전자기기(예. TV, 스마트폰, 냉장고 등)를 컨트롤하는 스마트홈의 허브가 될 것으로 기대되기 때문이다. 즉, 스마트스피커가 스마트홈의 핵심 플랫폼이 될 것이므로 이를 선점하기 위한 경쟁이 거센 것이다. 그러나 던(Dunn, 2017)과 킨셀라(Kinsella, 2019)에 따르면, 아직 스마트스피커의 주된 사용 용도는 단순한 질문, 날씨, 뉴스 등의 정보를 얻거나 음악 등의 콘텐츠를 즐기거나 알람/타이머를 설정하는 것으로 한정되어 있다고 한다. 하지만 문요한, 김기준, 신동희(Moon, Kim & Shin, 2016), 루리아, 호프만, 주커만(Luria, Hoffman & Zuckerman, 2017), 호이(Hoy, 2017), 그리고 김준한, 김유정, 김병준, 윤수경, 김민준, 이정석(Kim, Kim, Kim, Yun, Kim & Lee, 2018)은 스마트스피커와 관련된 연구 및 개발이 활발하게 진행되고 있어, 그 사용 범위는 점차 확대될 것으로 예측하였다.

스마트스피커 VUI (Voice User Interface) 디자인의 주요 이슈들 중 하나는 제공할 음성정보의 적절한 양을 결정하는 것이다. 펄(Pearl, 2016)은 사람 사이의 성공적인 대화를 위한 협동 원리 네 가지를 품질, 양, 관련성, 매너로 제시하였는데, 그 중 정보의 양에 있어서는 “필요한 만큼의 정보를 모두 얘기하되, 너무 과하게 얘기하지 말라”고 하였다. 배혜진(Bae, 2018)에 따르면, “VUI와 대화 시 너무 많은 대화 내용은 다소 무겁게 느껴지며 숙련된 서비스에 대해서는 더 짧게 대답하면 좋겠다”는 사용자 의견이 많았다고 한다. 이러한 이슈는 뉴스, 날씨, 백과사전과 같이 정보의 양이 다소 긴 서비스에서 주로 발생한다. 사용자들은 시각 인터페이스를 이용할 때 관심 없는 부분은 건너뛰고 관련된 정보에 빠르게 주의력을 집중한다. 그러나 펄(Pearl, 2016)에 의하면 음성정보는 시각 정보와 달리 사용자가 원하는 부분만 선택적으로 획득할 수 없고, 처음부터 끝까지 선형적으로 들을 수밖에 없다. 따라서 정보의 양이 너무 많아지면 사용자가 지루함을 느낄 수 있다. 반대로 제공되는 정보의 양이 너무 적어도 사용자가 정보량을 미흡하다고 느껴 서비스에 대한 만족감이 떨어질 수 있다. 하지만 스마트스피커가 제공하는 음성정보의 적절한 양에 대한 디자인 가이드라인이나 기존 연구는 존재하지 않았다.

본 연구에서는 백과사전 정보를 기준으로 하여 스마트스피커에 적절한 음성정보의 길이에 대한 연구를 수행하였다. 제공 정보에 대한 사용자의 관심도와 스피커의 음성이 동일하다고 할 때 사용자가 느끼는 음성정보 길이의 적절함은 두 가지 인자에 의해 영향을 받을 수 있다. 첫째, 문장당 글자수(문장길이)이다. 문장당 글자수가 많다는 것은 긴 문장의 음성이므로 사용자는 그 음성정보를 이해하기 어려울 수 있다. 들리는 정보를 이해하기 어렵다면 사용자는 쉽게 지루함을 느낄 것이다. 벤카타지리(Venkatagiri, 1994)는 문장길이가 합성음성의 이해도에 미치는 영향을 조사하였다. 연구 결과, 평균 11 단어 길이의 문장은 평균 5 단어 길이의 문장만큼 이해하기 쉬웠는데, 이는 어느 정도 긴 문장까지도 이해력을 해치지 않고 사용 가능함을 의미한다.

둘째, 음성의 발화 속도이다. 들리는 말이 빠르다면 사용자는 동일한 시간 내에 더 많은 양의 정보를 수용할 수 있을 것이다. 시몬즈, 메이어, 퀸란, 그리고 헌트(Simonds, Meyer, Quinlan & Hunt, 2006)에 의하면, 일반적으로 빠른 속도의 연설은 유능하고 지능적이며 신뢰성이 높다고 인식하는 경향이 있다고 한다. 물론 지나치게 빠른 말은 음성정보에 대한 사용자 이해를 오히려 방해하여 반대의 효과를 줄 수도 있다. 반면, 너무 느린 말은 지루하게 느껴질 수 있다. 폴포드와 장(Fulford & Zhang, 1993)에 따르면, 일상적인 대화에서 사람들은 125~150 WPM(Words Per Minute)으로 말한다고 한다. 파울케(Foulke, 1968)는 125에서 250 WPM까지의 발화속도는 들리는 음성의 이해력에 영향을 미치지 않으며, 그 이상 말이 빨라지면 이해에 악영향을 미침을 보였다. 한편,

로스, 미드랜드, 푹스, 파우지, 엥거트, 던컨, 바우한, 베르넷, 피터, 버넷, 그리고 메이(Ross, Midtland, Fuchs, Pauzie, Engert, Duncan, Vaughan, Vernet, Peters, Burnett & May, 1996)는 자동차 내비게이션의 음성안내는 분당 140 단어 혹은 그보다 짧은 속도로 제시되어야 함을 가이드라인으로 제시하였다. 서튼, 킹, 혹스, 그리고 베이켈맨(Sutton, King, Hux & Beukelman, 1995)는 150에서 200 WPM의 발화속도는 나이에 무관하게 편안하게 들을 수 있음을 보였다. 레이놀드와 기븐스(Reynolds & Givens, 2001)에 의하면, 150~200 WPM 수준의 발화속도를 가진 자연음성과 합성음성 사이에는 사람들의 응답 지연에 유의한 차이가 없었다고 한다.

앞서 살펴본 바와 같이 음성정보의 적절한 길이는 발화되는 음성의 문장길이와 발화속도에 영향을 받을 수 있다. 따라서 본 연구에서는 문장길이와 발화속도에 따른 적절한 음성정보의 길이를 도출하기 위한 실험을 수행하였다. 실험은 문장길이와 발화속도가 다른 충분히 긴 문장의 음성을 들으면서 실험참여자가 충분한 정보를 얻었다고 생각되는 시점 및 지루함이 느껴지는 시점에 의사를 표시하는 방법으로 진행되었다.

2. 적절한 TTS 길이에 대한 사전 연구

김헌, 박환수, 백미선, 이소향(Kim, Park, Baek & Lee, 2019)은 국내 스마트스피커 제조사의 백과사전에서 제공하고 있는 음성정보를 대상으로 TTS(Text-To-Speech) 길이에 따른 사용자가 느끼는 정보량과 지루함을 평가하였다. 실험참여자들은 스피커에서 나오는 음성정보를 들으며 충분한 정보를 얻었다고 판단될 때 주어진 종을 한번 치도록 하였다(충분한 정보량 - 첫 번째 종). 그리고 지루함이 느껴져 더 이상 듣고 싶지 않다고 판단되면 다시 한번 종을 치도록 하였다(지루한 정보량 - 두 번째 종). 그러나 실험 결과, 한번 이상 종을 친 경우의 비율이 20%에 불과했으며, 나머지 80%는 한 번도 종을 치지 않았다. 이 실험에 사용한 발화문의 길이는 한글 평균 200자, 최대 300자로 구성되어 있었다. 그러므로 적절한 TTS 길이는 200자 보다는 길 것이라는 사실을 확인할 수 있었을 뿐, 보다 명확한 결과를 도출하는 데는 한계가 있었다.

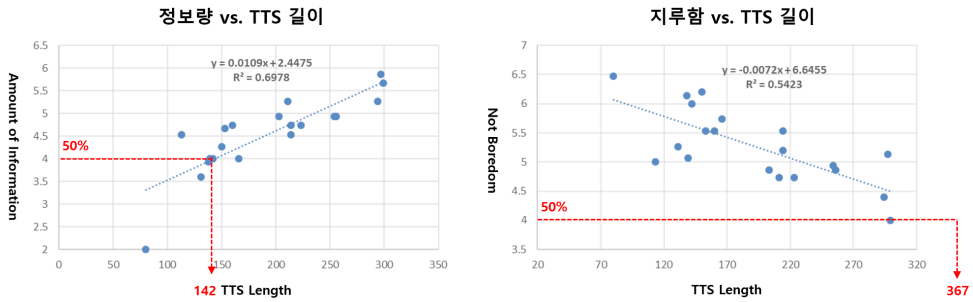

Figure 1 Regression analyses between TTS length and amount of information and boredom(김헌, 박환수, 백미선, 이소향 (Kim, Park, Baek & Lee, 2019)

김헌, 박환수, 백미선, 이소향(Kim, Park, Baek & Lee, 2019)은 Figure 1과 같이 TTS 길이와 지루함, TTS 길이와 정보량 간의 회귀분석을 수행하였는데, 50%의 사람들이 만족하는 수준을 기준으로 적절한 TTS 길이는 140자에서 360자 사이 정도일 것으로 예측하였다. 하지만 보다 정확한 결과를 위해서는 더 긴 TTS를 가지고 추가 실험을 수행할 필요성이 있었다. 본 연구는 위 연구의 후속 연구로서 1,000자의 긴 TTS 길이를 가지고 그에 따른 사용자의 반응을 조사하였다.

3. 실험

3. 1. 실험 목적

본 실험의 목적은 아래 세 가지 가설을 검증하면서 사람들이 지루하지 않으면서도 충분한 수준의 정보량을 느끼는 적절한 TTS 길이를 알아내는 것이다.

- 가설1: 사용자가 충분하다고 느끼는 음성정보의 길이는 내용에 대한 관심도에 따라 달라진다.
- 가설2: 문장길이가 길어지면, 사용자가 충분하다고 느끼는 음성정보의 길이는 짧아진다.
- 가설3: 발화속도가 빨라지면, 사용자가 충분하다고 느끼는 음성정보의 길이는 길어진다.

3. 2. 실험 환경 및 시스템

본 연구에서는 실험을 위해 백과사전에서 검색 빈도가 높은 여섯 가지 정보유형(책-나의 라임 오렌지나무, 여행-보라카이섬, 영화-인터스텔라, 회사-삼성전자, 인물-마이클 조던, 국가-베트남)의 발화문을 위키피디아와 네이버 사전을 참고하여 작성하였다. 각 발화문의 길이는 약 1,000자(약 3분 분량) 수준으로 충분히 길게 구성하였다. 이때, 가설2를 검증하고자 문장의 길이를 문장당 약 30자와 문장당 약 60자의 두 종류로 발화문을 구성하였다. 이 두 수준은 기존 스마트스피커들의 문장당 글자수를 벤치마킹하여 결정하였다(네이버 클로바: 54.83자/문장, SKT 누구: 30.05자/문장, 카카오 미니: 37.65자/문장). Table 1은 이렇게 구성된 두 가지 수준의 문장길이를 가진 발화문의 예를 보여준다.

작성된 발화문 텍스트들은 TTS 제작 소프트웨어를 이용하여 동일한 여성 목소리의 TTS 파일로 변환하였다. 이때, 가설3을 검증하고자 TTS 제작 시 음성의 빠르기를 5.3자/초, 5.6자/초, 5.9자/초의 세 수준으로 각각 제작하였다. 이 수준 역시 기존 스피커들의 발화 속도를 분석하여 결정하였다(네이버 클로바: 5.5자/초, SKT 누구: 5.59자/초, 카카오 미니: 5.27자/초, 구글홈: 5.99자/초). 이렇게 제작된 TTS는 노트북으로 재생하여 블루투스로 연결된 스마트스피커 누구 캔들(NUGU Candle)을 통해 실험참여자에게 들려주었다(Figure 2).

Table 1 Examples of the short and long sentences

문장길이	음성정보 내용
Short (문장당 약 30자)	보라카이 섬은 필리핀의 섬으로 필리핀 관광국과 아클란 주 정부에서 관리합니다. / 세계적인 휴양지로 손꼽히는 보라카이는 필리핀의 중서부 파나이 섬 북서쪽에 떠 있는 섬입니다. / 마지막 남은 천국이라 불릴 만큼 때 묻지 않은 자연을 지닌 휴양지에요. / 이곳에는 길이 칠 킬로미터에 달하는 길고 넓은 화이트 비치가 있답니다. / 그리고 야자수 숲이 어우러진 서른 두 개의 크고 작은 독특한 매력을 지닌 비치가 있습니다. / 보라카이에서는 자연 경관과 조화를 이룬 건축물을 짓기 위해 코코넛 나무 크기 이상의 건물을 지을 수 없습니다. / 또한, 파도가 밀려오는 지점에서 삼백 미터 이내에도 건물을 지을 수 없다고 하네요. (생략)
Long (문장당 약 60자)	보라카이 섬은 필리핀의 섬으로 필리핀 관광국과 아클란 주 정부에서 관리하며, 필리핀의 중서부 파나이 섬 북서쪽에 떠 있는 섬입니다. / 필리핀 최고의 휴양지로써 마지막 남은 천국이라 불릴 만큼 때 묻지 않은 자연을 지닌 휴양지에요. / 이곳에는 길이 칠 킬로미터에 달하는 길고 넓은 화이트 비치와 야자수 숲이 어우러진 서른 두 개의 크고 작은 독특한 매력을 지닌 비치가 있습니다. / 보라카이에서는 자연 경관과 조화를 이룬 건축물을 짓기 위해 코코넛 나무 크기 이상의 건물을 지을 수 없으며 파도가 밀려오는 지점에서 삼백미터 이내에도 건물을 지을 수 없다고 하네요. (생략)

Figure 2 Experimental environment

3. 3. 실험 참여자

실험에는 20-30대 남녀 23명(남성 11명, 여성 12명; 평균 24.4세)이 참여하였다. 그들은 모두 대학생 및 대학원생이었으며, 그중 여덟 명은 스마트스피커를 사용해본 경험을 가지고 있었다.

3. 4. 실험 태스크 및 절차

각 실험참여자들은 문장길이 2가지(30자/문장, 60자/문장) × 발화속도 3가지(5.3자/초, 5.6자/초, 5.9자/초)의 총 6종류의 음성을 임의의 순서로 스피커를 통해 듣고 반응하는 태스크를 수행하였다. 한 참여자가 듣는 6종류 음성의 내용 자체는 6개의 정보유형(책, 여행, 영화, 회사, 인물, 국가)으로 매번 달라지게 구성하였다.

Figure 3은 실험참여자들의 태스크를 보여주는데, 이 실험 태스크는 김헌, 박환수, 백미선, 이소향(Kim, Park, Baek & Lee, 2019)과 거의 동일하였다. 참여자들은 스피커에서 나오는 약 1,000자 길이의 음성을 들으면서 충분한 정보를 얻었다고 판단될 때 주어진 종을 한번 치도록 요구되었다(충분한 정보량 - 첫 번째 종). 그리고 지루함이 느껴져 더 이상 음성정보를 듣고 싶지 않다고 판단되면 다시 한번 종을 치도록 하였다(지루한 정보량 - 두 번째 종). 참여자가 두 번째 종을 치면 TTS 재생은 중단되었다. 실험의 반응값은 첫 번째 및 두 번째 종을 치기까지 들은 음성 TTS의 길이였다. 또한, 발화를 모두 들은 후 실험참여자들은 해당 음성정보에 대한 관심도를 7점 척도로 평가하였는데 이는 가설1을 검증하기 위함이었다.

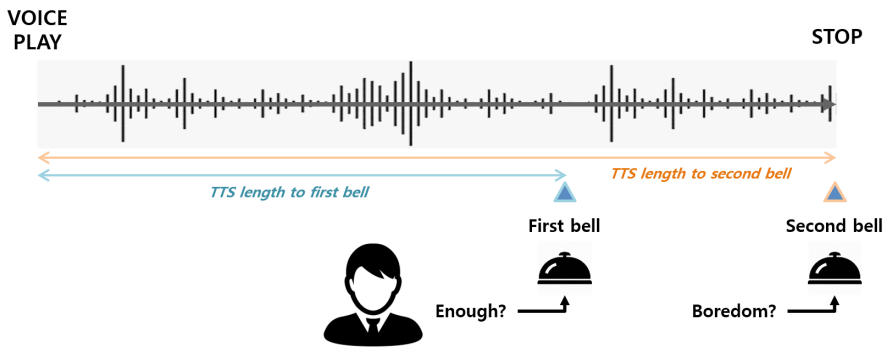

Figure 3 Experimental task and measures

4. 실험 결과

4. 1. 첫 번째 종을 치기까지의 TTS 길이

Table 2는 첫 번째 종을 치기까지의 TTS 길이에 대한 분산분석 결과이다(유의하지 않은 교호작용은 풀링 처리함). 본 연구에서 수립한 가설인 문장길이(가설2)나 발화속도(가설3)에 따라서는 TTS 길이에 유의한 차이가 존재하지 않았다. 또한, 음성정보의 여섯 유형 간에도 TTS 길이에 유의한 차이는 없었다. 하지만 남성이 여성보다 유의하게 종을 더 늦게 쳤다(남성 411.3±232.6자 vs. 여성 312.4±147.4자). 이는 실험에 사용된 음성정보의 내용이 여성보다는 남성에게 더 관심이 많을 내용이었기 때문으로 보인다. 이는 성별×관심도 교호작용이 존재했음에 의해서도 설명이 된다. 남성은 6이나 7점의 높은 관심도를 부여한 음성정보에서 첫 번째 종을 치기까지의 TTS 길이가 여성보다 더 길었다. 또한, Figure 4에서 보여주듯이 내용에 대한 관심도가 높을수록 유의하게 종을 더 늦게 치는 경향을 보였다(가설1). 관심도와 첫 번째 및 두 번째 종 TTS 길이 사이에는 약한 상관관계가 존재하였다(Pearson 상관계수 r = 0.429, 0.451). 즉, 당연한 말이지만 관심이 많은 내용은 참여자들이 지루해하지 않고 충분히 길게 들었다는 것이다. Figure 4를 보면 관심도가 보통 수준(4점)일 때 첫 번째 종은 대략 300자, 두 번째 종은 대략 500자일 때 친 것으로 보인다.

Table 2 The ANOVA result of TTS length to the moment the first bell is pressed ($^{*}p<.05$, $^{**}p<.01$)

인자	자유도	F-값	p-값
성별	1	8.15	0.005**
정보유형	5	0.94	0.460
발화속도	2	1.03	0.362
문장길이	1	2.74	0.101
관심도	6	4.56	0.000**
성별×문장길이	1	4.06	0.047*
성별×관심도	6	2.50	0.028*
정보유형×발화속도	10	2.93	0.003**
관심도×발화속도	12	2.35	0.011*
오차	93		
총계	137		

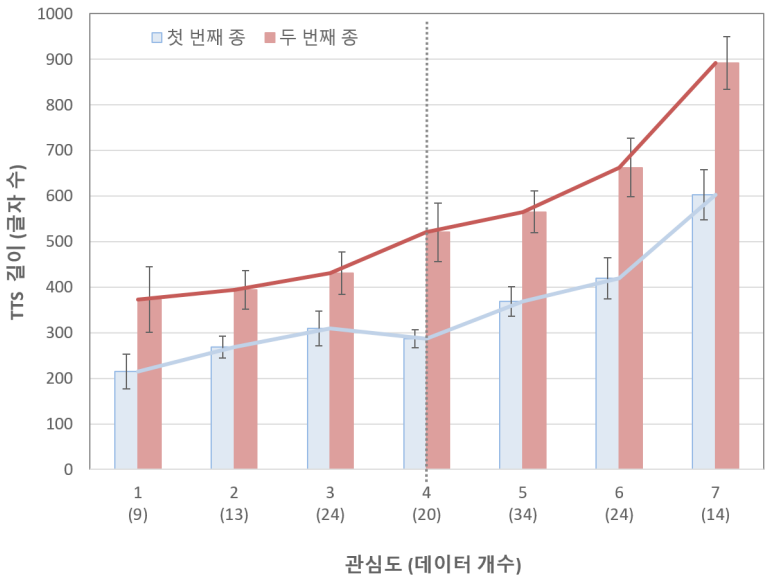

Figure 4 TTS lengths in terms of participant's interest

Table 2를 보면 성별×문장길이 교호작용이 존재하였다. 여성은 문장길이에 무관했으나 남성의 경우 긴 문장길이의 발화문에서 좀 더 늦게 종을 치는 경향을 보였다. 또한 정보유형×발화속도와 관심도×발화속도 교호작용이 유의한 것은 정보 내용이나 관심도에 따라 적합한 말의 빠르기가 다를 수 있음을 얘기해준다. 예를 들면, 1번 발화문은 발화속도 5.6자, 2번 발화문은 발화속도 5.3자, 그리고 6번 발화문은 발화속도 5.9자일 때 가장 늦게 종을 쳤다. 하지만 본 실험결과만으로는 관심도의 고저에 따른 적합한 발화속도의 뚜렷한 패턴을 발견하기는 어려웠다.

Figure 5는 첫 번째 종을 치기까지의 TTS 길이 데이터를 히스토그램으로 그린 것이다. TTS 길이 300자에서 첫 번째 종을 친 참여자가 가장 많았으며, 평균(표준편차)은 359.7(198.5)자였고 95% 신뢰구간은 (326, 393)자였다. 참여자에 따라 다소 차이가 있긴 했으나 주제에 무관하게 대략 300~400자 정도면 충분한 정보를 얻었다고 판단한 것으로 보인다.

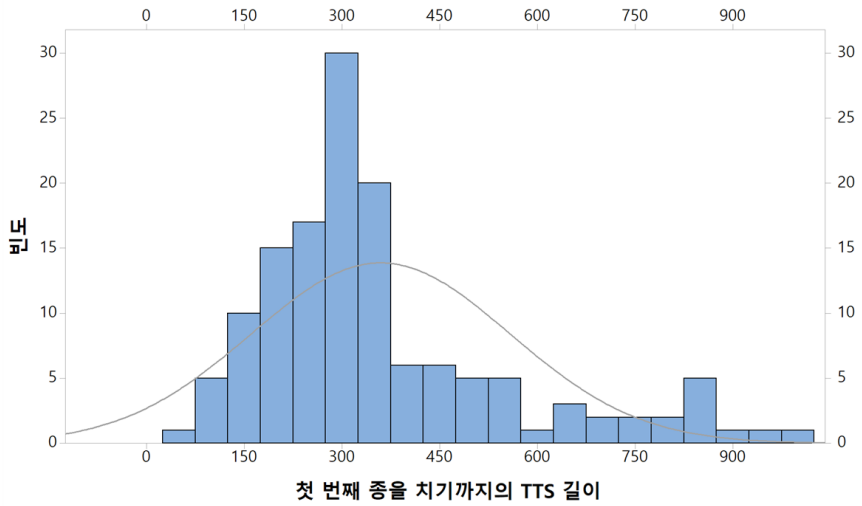

Figure 5 Histogram of TTS lengths to the first bell

4. 2. 두 번째 종을 치기까지의 TTS 길이

두 번째 종을 치기까지의 TTS 길이는 첫 번째 종을 치기까지의 TTS 길이와 높은 상관관계를 보였다(Pearson 상관계수 r = 0.771). Table 3은 두 번째 종을 치기까지의 TTS 길이에 대한 분산분석 결과이다. 첫 번째 종까지의 TTS 길이와 다른 점은 발화되는 정보유형에 따라 두 번째 종을 치기까지의 TTS 길이에 유의한 차이가 존재한 점이다. 책, 여행 관련 정보가 회사나 인물 정보보다 더 긴 TTS 길이를 보였다. 이는 정보의 내용에 대한 실험참여자의 관심도가 지루해지기까지 듣는 TTS의 길이에 영향을 미쳤기 때문일 것이다.

두 번째 종을 치기까지의 TTS 길이에서도 성별×문장길이 교호작용이 존재하였다. 여성은 문장길이가 짧을 때, 반면 남성은 문장길이가 길 때 더 오래 음성정보를 들었다. 관심도×발화속도 교호작용도 존재했는데, 실험참여자의 음성정보에 대한 관심도가 1점이거나 7점으로 내용에 대한 선호가 명확한 경우에 발화속도가 빠를수록 더 오래 듣는 경향을 보였다.

Table 2 The ANOVA result of TTS length to the moment the second bell is pressed ($^{*}p<.05$, $^{**}p<.01$)

인자	자유도	F-값	p-값
성별	1	6.99	0.009**
정보유형	5	2.54	0.032*
발화속도	2	0.15	0.863
문장길이	1	0.81	0.370
관심도	6	5.62	0.000**
성별×문장길이	1	6.59	0.012*
관심도×발화속도	12	1.87	0.046*
오차	109		
총계	137		

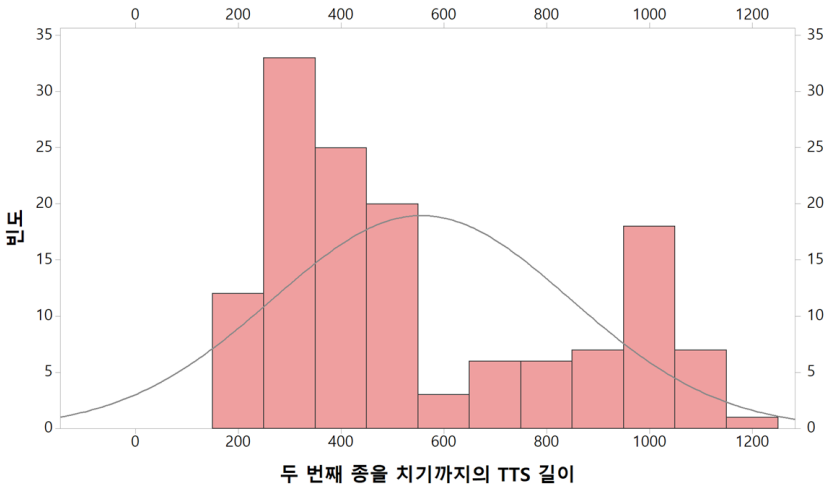

Figure 6 Histogram of TTS lengths to the second bell

두 번째 종의 경우에도 TTS 길이 300자에서 가장 높은 빈도로 울렸으며, 평균(표준편차)은 556.6(290.3)자, 95% 신뢰구간은 (507, 605)자였다. Figure 6의 두 번째 종까지의 TTS 길이를 보면 특이하게도 쌍봉 형태의 그래프를 보여준다. 이는 Table 3의 분산분석 결과에서도 알 수 있듯이 제공되는 정보유형에 따라 참여자가 느낀 지루함 정도가 달라졌기 때문이다. 또한, 참여자의 해당 주제에 대한 관심도에 따라서도 지루함을 빠르게 느낀 사람도 있었던 반면, 오래 견딘 사람들은 1,000자 정도까지도 음성정보를 듣고 있었기 때문이다.

4. 3. TTS 길이에 따른 종을 친 비율

Figure 7은 Figure 5의 히스토그램을 누적하여 그린 누적분포도이다. 스노드그래스, 레비버거, 그리고 헤이든(Snodgrass, Levy-Berger & Haydon, 1985)과 페드램(Pedram, 2016)에 따르면, 정신물리학에서의 절대식역(Absolute threshold)은 탐지확률이 50%인 자극 강도를 기준으로 정의한다. 본 연구에서는 50%의 사용자가 정보량이 충분하다고 판단한 수준으로 볼 수 있다. Figure 7에서 50%의 빈도로 첫 번째 종을 치는 지점은 305자였으며, 70%의 빈도로 첫 번째 종을 치는 지점은 364자였다. 또한, 50%의 빈도로 두 번째 종을 치는 지점은 445자였으며, 70%의 빈도로 두 번째 종을 치는 지점은 725자였다. 즉, 스마트스피커에서 나오는 음성정보의 길이가 대략 300 (360)자면, 50% (70%)의 사용자는 정보가 충분함을 느꼈으며, 음성정보의 길이가 대략 450 (730)자가 되면 50% (70%)의 사용자는 지루함을 느꼈다고 볼 수 있다.

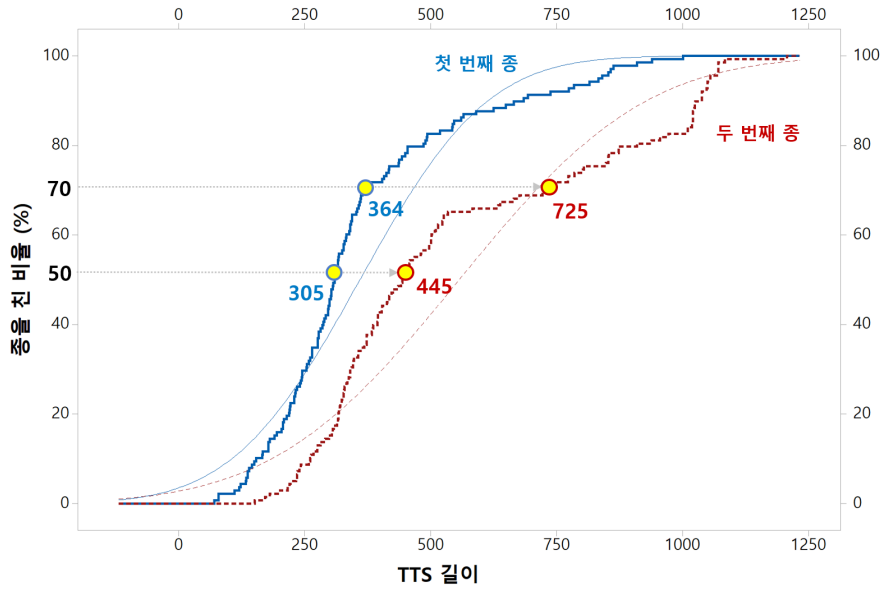

Figure 7 Cumulative distribution of TTS lengths to the first and second bells

5. 결론 및 고찰

본 연구는 스마트스피커가 백과사전 정보와 같이 긴 음성정보를 들려주는 경우에 사용자가 지루하지 않으면서도 적절한 정보량을 느끼는 TTS 길이를 도출하는 실험을 수행하였다. 실험의 가설과 그 결과는 아래와 같았다.

- 음성정보의 적절한 길이: 50%의 사용자들이 정보의 충분함을 느끼는 음성정보의 길이는 약 300자 정도였으며, 지루함을 느끼는 음성정보의 길이는 약 450자였다.

- 사용자가 충분하다고 느끼는 음성정보의 길이는 내용에 대한 관심도에 따라 달라진다(가설1 채택). 정보에 대한 사용자의 관심도가 높을수록 사용자는 더 긴 음성정보를 수용하였다.

- 문장길이가 길어지면, 사용자가 충분하다고 느끼는 음성정보의 길이는 짧아진다(가설2 기각). 글자수는 사용자가 충분하거나 지루하다고 느끼는 음성정보의 길이에 영향을 미치지 않았다.

- 발화속도가 빨라지면, 사용자가 충분하다고 느끼는 음성정보의 길이는 길어진다(가설3 기각). 음성의 발화속도는 사용자가 충분하거나 지루하다고 느끼는 음성정보의 길이에 영향을 미치지 않았다.

하지만 본 연구의 실험 결과에는 다음과 같은 한계점이 있다. 첫째, 충분한 정보량을 얻었다고 생각되어 첫 번째 종을 치는 시점이 실험참여자들의 주관적 판단에만 의존했다는 점이다. 종을 친 시점에 참여자들이 실제로 음성정보의 내용을 얼마나 잘 기억했고, 이해했는지를 교차 검증할 필요성이 있었다. 둘째, 본 연구의 실험에 사용된 음성 TTS의 문장길이(문장당 글자수)와 발화속도는 국내 시장에 출시된 스마트스피커의 음성을 기준으로 제작하였다. 만일 실험에 사용된 수준 이상으로 문장당 글자수가 더 많았거나 발화속도가 더 빨랐다면 실험 결과는 달랐을 수도 있다. 하지만 그러한 수준의 음성은 문장을 지나치게 이해하기 어렵게 만들거나 알아듣기 힘든 빠르기의 음성이었을 것이다. 셋째, 본 연구의 실험에 참여한 사람들의 연령은 20~30대 젊은 층으로 한정되어 있었다. 윙필드, 맥코이, 피엘, 튼, 그리고 콕스(Wingfield, McCoy, Peelle, Tun & Cox, 2006)의 연구에 의하면, 들

리는 음성에 대한 이해도는 문장의 복잡성이나 발화속도뿐 아니라 연령과 청력손실의 영향에 의해서도 달라질 수 있다고 하였다. 즉, 고령층이나 어린이에 특화된 음성 서비스의 경우 그에 적합한 문장길이나 발화속도를 고려해야 할 필요성이 있을 것이다.

본 연구의 실험 결과에 따르면, 음성정보의 내용과 사용자의 관심도에 따라 사용자가 충분하다고 생각하는 정보량이나 지루함의 수준은 달라질 수 있으나 음성정보의 길이는 대략 300자 내외로 작성하되 최대 450자를 넘지 않는 것이 좋을 것으로 보인다. 이를 실험에 사용된 발화속도로 나눠서 시간으로 환산하면 약 50~90초에 해당한다. 이 가이드라인은 스마트스피커의 VUI 디자이너가 백과사전과 같은 긴 콘텐츠를 제작할 때 제공할 음성정보의 적절한 길이를 결정하는데 활용할 수 있다. 또한 본 연구의 실험이 다양한 주제의 백과사전 정보에 대해 이루어진 점을 고려한다면, 이 가이드라인은 백과사전 외의 다른 콘텐츠 유형(예. 뉴스, 날씨 등)에도 확대 적용 가능할 것으로 판단된다.

References

1. Bae, H. (2018). *음성 대화형 인터페이스(VUI) 설계를 위한 가이드라인 연구: 의사소통 전략을 중심으로 [Guidelines for voice user interface design: based on communication strategy]* (Master's thesis). Available from Korean Education and Research Information Service.
2. Dunn, J. (2017, May, 31). People mainly use smart speakers for simple requests. *Business Insider*, Retrieved, May 2019, from https://www.businessinsider.com/how-people-use-smart-speakers-amazon-echo-chart-2017-5.
3. Foulke, E. (1968). Listening comprehension as a function of word rate. *Journal of Communication, 18*(3), 198-206.
4. Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: an introduction to voice assistants. *Medical reference services quarterly, 37*(1), 81-88.
5. Kim, J., Kim, Y., Kim, B., Yun, S., Kim, M., & Lee, J. (2018). Can a Machine Tend to Teenagers' Emotional Needs?: A Study with Conversational Agents. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (p. LBW018). ACM.
6. Kim, H., Park, H., Baek, M., & Lee, S. (2019, October). 스마트스피커의 백과사전 제공 음성정보의 적절한 길이[Proper length of voice information provided by encyclopedia in smart speakers]. In *Proceedings of the ESK Conference* (pp.52-55).
7. Kinsella, B. (2019). Smart speaker owners agree that questions, music, and weather are killer apps. What comes next?. *Voicebot*, Retrieved March, 2019, from https://voicebot.ai/2019/03/12/smart-speaker-owners-agree-that-questions-music-and-weather-are-killer-apps-what-comes-next/.
8. Luria, M., Hoffman, G., & Zuckerman, O. (2017). Comparing social robot, screen and voice interfaces for smart-home control. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 580-628). ACM.
9. Market Research Report (2018). Smart speakers market size, share & trends analysis report by virtual personal assistant (Amazon Alexa, Google Assistant, Apple Siri), Region, and segment forecasts 2018 - 2025.
10. Moon, Y., Kim, K. J., & Shin, D.H. (2016). Voices of the internet of things: An exploration of multiple voice effects in smart homes. In *International Conference on Distributed, Ambient, and Pervasive Interactions* (pp. 270-278). Springer, Cham.
11. Pearl, C. (2016). *Designing Voice User Interfaces: Principles of Conversational Experiences*. O'Reilly Media, Inc.
12. Pedram, S. A. (2016). H*aptic texture rendering and perception using coil array magnetic levitation haptic interface: Effects of torque feedback and probe type on roughness perception* (Doctoral dissertation), University of Hawai'i at Manoa.
13. Reynolds, M. E., & Givens, J. (2001). Presentation rate in comprehension of natural and synthesized speech. *Perceptual and motor skills, 92*, pp.958-968.

14. Ross, T., Midtland, K., Fuchs, M., Pauzié, A., Engert, A., Duncan, B., Vaughan, G., Vernet, M., Peters, H., Burnett, G. E., & May, A. J. (1996). HARDIE design guidelines handbook. *Human Factors Guidelines for Information Presentation by ATT Systems 530*.
15. Simonds, B. K., Meyer, K. R., Quinlan, M. M., & Hunt, S. K. (2006). Effects of instructor speech rate on student affective learning, recall, and perceptions of nonverbal immediacy, credibility, and clarity. *Communication Research Reports, 23*(3), pp.187–197.
16. Snodgrass, J. G., Levy–Berger, G., & Haydon, M. (1985). *Human experimental psychology* (Vol. 395). New York: Oxford University Press.
17. Sutton, B., King, J., Hux, K., & Beukelman, D. (1995). Younger and older adults' rate performance when listening to synthetic speech. *Augmentative and Alternative Communication, 11*(3), 147–153.
18. Venkatagiri, H. (1994). Effect of sentence length and exposure on the intelligibility of synthesized speech. *Augmentative and Alternative Communication, 10*(2), 96–104.
19. Wingfield, A., McCoy, S. L., Peelle, J. E., Tun, P. A., & Cox, C. L. (2006). Effects of adult aging and hearing loss on comprehension of rapid speech varying in syntactic complexity. *Journal of the American Academy of Audiology, 17*(7), pp.487–497.

청자의 지루함: 스마트스피커에 적합한 음성정보의 길이는?

김헌[1*], 고재영[2], 김승완[2], 황호연[2]

[1]서울과학기술대학교 기계시스템디자인공학과, 교수, 서울, 대한민국

[2]서울과학기술대학교 기계시스템디자인공학과, 학생, 서울, 대한민국

초록

연구배경 스마트스피커에 의해 제공되는 음성정보의 길이가 너무 길다면, 사용자는 지루해할 수 있다. 반대로 음성정보의 길이나 너무 짧다면, 사용자는 정보의 양이 불충분하다고 느낄 수 있다. 따라서 백과사전 정보와 같이 스마트스피커에서 제공하는 긴 음성정보의 경우 사용자가 지루해하지 않으면서 정보량은 충분하다고 느끼는 길이에 대한 가이드라인이 필요하다.

연구방법 본 연구에서는 실험참여자가 서로 다른 문장길이와 발화속도를 가진 긴 문장을 듣다가 충분한 정보량이나 지루함을 느꼈을 때 벨을 누르는 실험을 수행하였다. 그리고 벨이 눌린 시점들에서의 음성정보 길이를 분석하였다.

연구결과 실험 결과, 정보의 내용, 문장의 평균길이, 음성의 발화속도는 사용자가 벨을 누르는 시점에 영향을 미치지 않았다. 실험참여자가 지루해하지 않으면서도 충분한 정보량을 느끼는 음성정보의 적절한 길이는 한글 약 300자 내외였으며, 지루함을 느끼는 길이는 한글 약 450자 수준이었다.

결론 본 연구의 결과는 스마트스피커가 제공하는 백과사전, 뉴스, 날씨와 같은 긴 음성정보의 길이를 결정하기 위한 가이드라인으로 활용할 수 있다.

주제어 스마트스피커, 음성 인터페이스, 대화형 디자인, TTS길이

본 연구는 SK텔레콤과 산업통상자원부 R&D사업 '창조혁신형 디자인고급인력양성사업'의 지원으로 진행되었음

*교신저자 : 김헌 (huhnkim@seoultech.ac.kr)