JSLHD 2022. Vol. 31, No. 4. 081-090
ISSN 1226-587X / eISSN 2671-7158
Journal of Speech-Language & Hearing Disorders 언어치료연구

ORIGINAL ARTICLE
https://doi.org/10.15724/jslhd.2022.31.4.081
http://jslhd.org/

A Comparison of Acoustic and Perceptual Measures of a Female Synthesized Voice Versus a Real Voice: A Multi-Dimensional Analysis

Soonha Kwon[1], Yesol Jeon[1], Sung Ah Yoo[1], Yoorim Oh[1], Youngmee Lee[2*]

[1] *Dept. of Communication Disorders, Graduate School, Ewha Womans University, Master's Student*
[2] *Dept. of Communication Disorders, Graduate School, Ewha Womans University, Professor*

Purpose : This study aimed to examine the difference between a female synthesized voice and a real voice based on multi-dimensional analysis.
Methods : Sixteen female adults aged between 23 and 35 years old participated in this study. They read three sentences with their speech being recorded. In addition, 16 young female voices were synthesized using the voice synthesis program Naver Clova Dubbing and Typecast. The data for all 32 voices were measured in terms of 4 dimensions: speech rate, prosody, articulation, and auditory-perceptual evaluation.
Results : First, the portion of synthesized voice pause was significantly longer than that of the real voice. Second, there were no significant differences between synthesized and real voice pitches. Third, the vowel duration in the CVC contexts was significantly longer than that of the real voice, and the vowel space area of the synthesized voice was significantly larger than that of the real voice. Finally, listeners perceived the real voice as being softer than the synthesized voice. They also perceived the synthesized voice as being faster and clearer than the real voice.
Conclusions : The current study found that female synthesized voices are different from female real voices in terms of speech rate, articulation, and auditory perception. Compared to the human voice, the synthesized voice represented a young female voice with a faster speech rate and clearer articulation. Through experimental evaluations, we have confirmed that TTS (text-to-speech) techniques successfully synthesize a real voice.

Keywords : Synthesized voice, real voice, acoustic evaluation, auditory-perceptual evaluation, multi-dimensional analysis

Correspondence : Youngmee Lee, PhD
E-mail : youngmee@ewha.ac.kr

Received : July 08, 2022
Revision revised : October 01, 2022
Accepted : October 31, 2022

ORCID
Soonha Kwon
https://orcid.org/0000-0003-0199-8389
Yesol Jeon
https://orcid.org/0000-0001-8988-3222
Sung Ah Yoo
https://orcid.org/0000-0001-6810-8745
Yoorim Oh
https://orcid.org/0000-0001-7104-7787
Youngmee Lee
https://orcid.org/0000-0003-1809-5944

Ⅰ. 서 론

음성합성(text-to-speech: TTS) 기술은 음성신호(speech signal)의 자동생성(automatic generation) 시스템으로서(Jung, 1994), 초기에는 주로 구어 의사소통에 어려움이 있는 사람들이 TTS 기술을 사용하였지만 점차 실생활 곳곳으로 응용 범위가 넓어지고 있는 상황이다(Yu, 2019). 예를 들면, ARS 전화 음성 서비스를 비롯하여 안내 방송, 전자 우편, 내비게이션, 가전, AI 스피커, 음성 번역기, 로봇, 전자책, 스크린 리더, 무인 종합 정보 안내 시스템, 교육, 오락 등 우리 삶 전반의 영역에서 다양하게 활용되고 있다(Voicewear, 2019). 이처럼 TTS의 활용이 광범위해짐에 따라, 일반인들에게 합성음성이 얼마나 자연스럽게 인식되는가에 대한 이슈가 대두되고 있다. 합성음성은 전하고자 하는 메시지를 정확하고 자연스러운 발음으로 산출되는 것이 중요하며, 언어학적 분석을 바탕으로 알맞은 억양(intonation)과 장단(duration)으로 음성이 산출되는 것도 중요하다(Jung, 1994). 또한, 음성 검사 시 사용하는 요소인 음의 높낮이(pitch), 음의 세기(intensity), 음성 품질(voice quality)과 유동성(flexibility) 등이 적절히 조화를 이루는지를 분석함으로써 좋은 음성인지 평가할 수 있다(Kim & Kwon, 2014).

합성음성과 관련된 선행연구(Gown et al., 2021; Park et al., 2018; Yu, 2019)에서는 주로 말 명료도(speech intelligibility)와 자연스러움(naturalness) 향상을 위한 공학적 차원의 분석이 진행되어 왔다. 그러나 합성음성을 언어학적 차원에서 분석한 국내 논문은 아직 없는 실정이다. 합성음성을 의사소통과 언어학적 측면에서 분석하기 위해서는 합성음성의 발화에 대한 시간, 운율, 조음 차원의 평가와 일반 청자의 청지각적 인식이 반영된 다차원적인 음성 프로파일 분석과 해석이 필요하다. 이를 통해 시간적 차원에서 합성음성의 속도가 적절한지, 운율적 차원에서 합성음성의 음도는 어떠한지, 조음 차원에서 말의 명료성 정도와 모음산출의 길이가 자연스러운지, 또, 청지각적으로 합성음성이 어떻게 지각되고 호감 있는 음성으로 느껴지는지 살펴볼 수 있다.

시간적 차원에서 음성을 분석한 Lee와 Hong(2013)은 합성음성의 속도에 따라 청자의 메시지 이해 정도가 달라질 수 있다고 보고하면서, 합성음성의 말속도는 청자가 정보를 정확하게 인지하고 습득하는 데 영향을 미치는 요소임을 설명했다. Lee 등(2014)의 연구에서는 합성음성의 처리 과정을 소개하며 발성 시 기본 주파수, 음소의 지속시간, 음량, 휴지구간 길이 등에 의해 합성음성의 자연스러움이 결정된다고 하였다. 특히, 음소의 지속시간 및 휴지구간의 길이가 합성음성의 명료도와 자연성을 결정하는 중요한 요소라고 설명하였다.

운율적 차원에서 음성의 음도를 분석하여, 운율의 적절성을 평가할 수 있다. Nam과 Choi(2008)의 연구에서는 정상인 및 장애인들의 음성 특성을 평가하는데 음도(pitch)가 중요한 요인 중 하나라고 주장하였다. 음도는 음의 자극에 대한 청자의 반응으로, 화자가 음도를 조절함으로써 음색을 제어할 수 있으며, 그 이유는 화자 별로 성문의 닫힌 구간의 간격이 다르기 때문이라고 설명하였다. 이에 닫힌 구간에 대한 특성을 이용하여 음도를 조절함으로써 화자의 음색이 달라질 수 있다고 하였다(Choi et al., 1998).

조음 차원에서 음성을 분석할 때의 요소로, F_2 기울기, 모음길이, 모음면적 등이 있다. Lee(2012)의 연구에서는 이중모음에서의 제2 포먼트 기울기를 측정함으로써 성도 모양에 큰 변화 즉, 이중모음을 산출하는 동안의 혀 위치 변경에 따른 전체적인 말 명료도를 예측할 수 있다고 설명하였다. 또한 Kim 등(2011)의 연구에서 F_2 기울기가 완만할수록 청지각적으로 말 명료도가 저하되었다고 설명하였다. 그리고 말소리의 분석요소로 모음면적과 모음길이비율이 있다. 모음공간면적은 개구도와 혀의 전, 후 위치를 나타내는 음형대 주파수 값을 수치화시킨 것으로, 모음공간면적과 명료도는 밀접한 관계가 있다(Ferguson & Kewley-Port, 2007). 그리고 모음길이는 발화 시 한 음소가 다른 음소에게 영향을 주는 동시조음의 영향에 의해 결정되는데 동시조음의 영향을 받는지를 측정함으로써 말소리가 자연스럽게 발음되는지를 알 수 있다(Kong et al., 1997).

본 연구에서는 시간, 운율, 조음 차원과 함께 청지각적 차원에서 합성음성을 실제음성과 비교해보고, 그 특성을 분석해보고자 하였다. 음성의 사용은 주로 의사소통의 상황에서 이루어지므로, 음성을 객관적인 수치로 분석하는 것 외에도 청지각적으로 청자에게 어떻게 인식되는지도 중요하다. Lee 등(2016)의 연구에서는 말소리의 자연스러움 정도에 대한 평가요소를 운율, 말속도, 말소리 전달 명료도, 전체적인 평가 4개 항목으로 평가하였고, Kwon(2015)의 연구에서는 목소리의 호감도를 평가하였다. 그래서 본 연구에서도 객관적인 지표와 함께 실제로 청자의 입장에서 합성음성을 어떻게 인식하는지를 호감도 및 자연스러움의 평가항목을 반영하여 설문지를 통해 알아보고자 하였다.

합성음성은 기술 발전에 따라 점차 명료성과 자연성이 높아지고 있고, 그 활용의 범위도 넓어지고 있다. 그럼에도 합성음성이 사용되는 상황은 자연스러운 의사소통보다는 주로 내비게이션, 키오스크 등의 일방적인 정보전달 상황인 경우가 많다. 이러한 상황을 고려할 때, 합성음성이 다양한 의사소통 목적에 따라 사용될 수 있을지에 대한 음향학적인 분석에 기반을 둔 연구가 필요하다. 이에 따라, 본 연구에서는 시간, 운율, 조음 측면에서의 음향음성학적 분석 및 청지각적 평가를 통해 합성음성과 실제음성 간에 어떠한 차이가 있는지를 비교 분석해보고자 하였다. 이를 통해, TTS 기술이 실제음성을 구현하는데 어떠한 측면을 보완해야 할지 살펴보고자 한다. 나아가 AAC 기기의 합성음성, AI를 활용한 의사소통 중재, 합성음성 챗봇을 활용한 언어 검사 등 교육과 언어병리학 분야에서 다양하게 활용될 수 있는 방법을 모색하는데, 본 연구의 결과가 기초자료로 제공될 수 있을 것이다.

본 연구의 구체적인 연구 문제는 다음과 같다.

첫째, 합성음성과 실제음성 간에 시간 차원의 음성 특성(전체 말속도, 휴지시간, 휴지비율)에 유의한 차이가 있는가?

둘째, 합성음성과 실제음성 간에 운율 차원의 음성 특성(피치의 평균, 중앙, 범위)에 유의한 차이가 있는가?

셋째, 합성음성과 실제음성 간에 조음 차원의 음성 특성(F_2 기울기, 모음길이, 모음면적)에 유의한 차이가 있는가?

넷째, 합성음성과 실제음성 간에 청지각적 평정점수(말속도, 운율, 명료도, 부드러움, 친절, 호감도)에 유의한 차이가 있는가?

II. 연구 방법

1. 연구 대상

1) 화자

본 연구에서는 실제음성 데이터는 20~30대 여성 16명을 대상으로 확보하였으며, 합성음성 데이터는 네이버 클로바 더빙과 타입캐스트 프로그램의 TTS 기술을 이용하여 확보하였다.

실제음성 화자의 선별 기준은 다음과 같다. (1)한국어를 모국어로 사용하는 20~30대 여성이며, (2)서울, 경기, 인천에 거주하고, (3)자가 보고를 통해 목소리와 관련된 질환이 없었던 자로 선정하였다. 최종 선정된 여성 화자의 평균 연령은 26.38세(SD=3.59)였으며, 연령범위는 23~35세였다. 합성음성 화자는 네이버 클로바 더빙(https://clovadubbing.naver.com)과 타입캐스트(https://typecast.ai/ko)의 데이터베이스를 이용하였다. 이때 연구자는 수집된 실제음성과 동일한 연령대, 성별, 언어에 해당하는 합성음성을 수집하기 위해서, 성별은 여성, 연령대는 청년, 언어는 한국어로, 스타일은 기본, 뉴스/리포터, 구연동화, 내레이션, 게임으로 선택하였다. 이후, 연구자들은 합성된 음성에 대한 자연스러움에 대해서 평정하여 자연스러움이 높은 경우만으로 선별하였다.

2) 청자

청지각적 평가는 20대 남녀 15명을 대상으로 진행하였다. 청자 선별 기준은 다음과 같다. (1)한국어를 모국어로 사용하는 20대이며, (2)서울·경기·인천에 거주하고, (3)자가 보고를 통해 청력에 이상이 없는 자로 선정하였다. 청자의 평균 연

령은 26.13세(*SD*=2.29)였으며, 연령범위는 20~29세였다.

2. 검사 도구

1) 발화 과제

본 연구에 참여한 대상자는 가을 문단(Kim, 1996)에서 발췌한 3문장을 낭독하였다(Table 1). 가을 문단은 표준화된 문단 읽기 과제로 자모음이 일정한 비율을 갖추고 있다(Kim, 1996). 대상자는 목표 문장을 3번 반복해서 읽었으며, 반복할 때마다 5초간의 간격을 두었다.

Table 1. Sentence list

우리나라의 가을은 참으로 아름답다. 무엇보다도 산에 오를 땐 더욱 더 그 빼어난 아름다움이 느껴진다. 숲속에 누워서 하늘을 바라보라.

2) 청지각적 평가 과제

실제음성과 합성음성의 자연스러움의 정도를 알아보기 위해 청지각적 평가를 하였다. 발화 과제에 대한 32개의 실제음성과 합성음성을 무작위로 배열하여 검사자료를 PPT 형태로 만들어 청자에게 제공하였다. 청지각적 설문은 각 음원 당 총 6가지 항목(속도, 운율, 명료도, 부드러움, 친절, 호감도)이며, 7점 척도로 평가하였다(Table 2). Lee 등(2016)의 연구에서 운율, 말속도, 명료도(speech intelligibility), 전체적인 평가(overall rating) 항목을 발췌하였고, Kwon(2015)의 연구에서는 여성 화자의 목소리를 평가하는 4가지 호감도 항목 중 부드러움(softness), 친절함(kindness)을 참고하여 총 6가지 항목으로 구성하였다.

3. 실험 설계

1) 화자 발화 수집

실제음성 발화 수집은 조용한 공간에서 진행되었으며, 참가자의 입과 핀 마이크(ECM-CS10, Sony Inc, Osaka, Japan) 사이의 거리는 10cm를 유지하였다. 녹음기는 Roland의 EDIROL R-05HR(Roland, Inc., Osaka, Japan)을 사용하였으며, 녹음 시 표본 추출률은 44.1kHz, 양자화 16bit로 하였다. 연구자는 대상자에게 발화 과제 문장을 최대한 명료하고 정확하게 읽어달라고 요청하였다. 대상자는 목표 문장을 3번 반복해서 읽었으며, 반복할 때마다 5초간의 간격을 두었다. 연구자들은 대상자별로 수집된 발화 3개를 명료성을 기준으로 평정하여, 가장 높게 평정된 발화를 최종 음성 샘플로 선정하였다. 합성음성의 경우, 합성음성을 제공하는 네이버 클로바 더빙과 타입캐스트에서 연구 과제에 해당되는 문장을 적은 뒤, 선정된 16명의 합성음성 화자 별로 음성을 추출하였다.

2) 청지각적 평가 설문

청자에게 음성자료로 제시하기 위하여 연구 과제에 대한 32개의 실제음성과 합성음성을 무작위로 배열하여 PPT 형태로 만들었으며, 설문지의 양식은 구글폼(https://docs.google.com/forms) 설문 양식을 사용하였다. 청자는 음성자료를 듣고 각 음원 당 총 6가지 항목(속도, 운율, 명료도, 부드러움, 친절, 호감도)을 7점 척도로 설문지에 평가하였다.

3) 음향음성 분석

16개의 실제음성과 16개의 합성음성을 합한 총 32개의 발화를 Praat(ver. 6.1.51)으로 분석하였다. 발화에 대한 시간, 조음, 운율, 청지각적 평정에 대한 구체적인 방법은 아래에 기술하였다.

(1) 시간차원 분석

전체 말 속도(overall speech rate)는 초당 음절 수(Shin, 2018; Yu, 2019)로, 제시된 문단 발화의 휴지시간이 포함된 문단 전체 발화시간을 음절수로 나누어 측정하였다. 평균 휴지시간(average duration of pause)은 총 휴지시간을 휴지 빈도수로 나누어 산출하였으며, 이때의 휴지시간은 생리적, 조음 실현을 위해 발생한 묵음을 제외한 0.1ms 이상의 묵음 구간으로 측정하였다(Lee, 2020; Yoo & Shin, 2020, Figure 1). 휴지비율

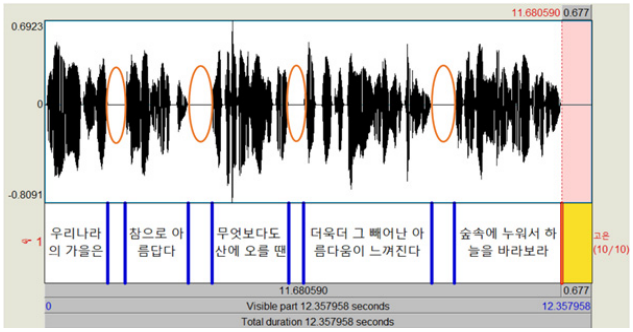

Figure 1. Example of pause analysis using Praat

Table 2. Perceptual evaluation questionnaire

	Negative	Definitely disagree	Mostly disagree	Slightly disagree	Neutral	Slightly agree	Mostly agree	Definitely agree	Positive
Speech rate	Very slow	①	②	③	④	⑤	⑥	⑦	Very fast
Prosody	Less of rhythm	①	②	③	④	⑤	⑥	⑦	Full of rhythm
Speech intelligibility	Very unclear	①	②	③	④	⑤	⑥	⑦	Very clear
Softness	Very hard	①	②	③	④	⑤	⑥	⑦	Very soft
Kindness	Very unkind	①	②	③	④	⑤	⑥	⑦	Very kind
Preference	Low preference	①	②	③	④	⑤	⑥	⑦	High preference

(portion of pause)은 전체 발화시간에서 휴지가 차지하는 비율로 전체 휴지시간에서 전체 발화시간을 나눈 후에 100을 곱하여 산출하였다.

(2) 운율차원 분석

음의 높낮이에 대해 피치의 평균/중앙/범위(pitch level, pitch median, pitch span)를 사용하였으며 해당 값들은 문단 전체를 선택하여 Praat의 voice report를 통해 측정하였다.

(3) 조음차원 분석

F_2 기울기(second formant slope: F_2 slope)는 제시된 문단의 이중모음(/ㅕ/, /ㅟ/)에 대하여, 20ms 구간 기준으로 20Hz 이상 차이가 날 때의 F_2 기울기를 절대값으로 측정하였다(Byeon, 2009; Cho et al., 2021; Choi, 2011; Han et al., 2013; Kim et al., 2009; Lee, 2012, Figure 2).

모음길이(vowel duration)는 포먼트가 일정한 구간(또는 파형이 규칙적인 구간)을 기준으로 4개의 모음(/a/, /i/, /u/, /ae/)을 V, CV, CVC 상황에서 각각 측정하였다. 모음면적(vowel space area)은 다음과 같은 공식을 사용하여 모음사각도를 측정하였다(Han, 2013; Higgins & Hodge, 2002; Park & Hugh, 2014).

$$\text{모음면적} = .5(F_1/i/ \times F_2/u/ - F_1/u/ \times F_2/i/) + .5(F_1/u/ \times F_2/a/ - F_1/a/ \times F_2/u/) + .5(F_1/a/ \times F_2/æ/ - F_1/æ/ \times F_2/a/) + .5(F_1/æ/ \times F_2/i/ - F_1/i/ \times F_2/æ/)$$

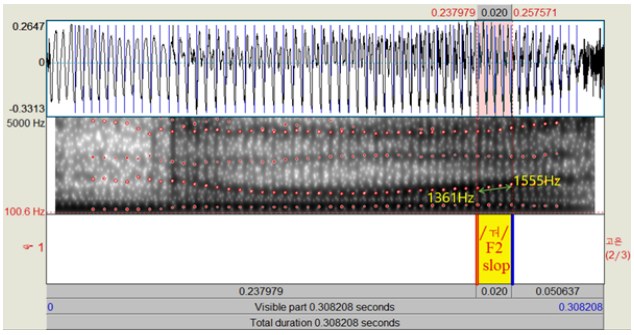

Figure 2. Example of F_2 slope analysis using Praat

(4) 청지각적 차원 분석

본 연구에서는 16명의 청자가 실제음성과 합성음성을 듣고 6가지 항목(속도, 운율, 명료도, 부드러움, 친절, 호감도)에 따라 1~7점으로 평정하였으며, 항목별로 평균 점수를 산출하였다.

4. 신뢰도

본 연구에서는 평가자 내 신뢰도(intra-rater reliability)와 평가자 간 신뢰도(inter-rater reliability)를 산출하기 위해서, 전체 32개의 자료 중 20%에 해당하는 7개의 샘플을 선택하였으며, 신뢰도는 Pearson 상관계수로 산출하였다. 본 연구의 음향·음성분석을 실시한 평가자들은 모두 총 6시간 이상의 Praat 프로그램에 대한 훈련을 받았으며, 충분히 Praat 프로그램 사용에 대해서 숙련된 이후에 음성 분석을 실시하였다.

평가자 내 신뢰도 산출을 위해서 언어병리학 석사과정에 있는 본 연구의 제1 저자가 데이터 분석을 실시하였다. 그 결과, 모든 측정치에서 상관계수가 .911~1.000이었다(all p<.01). 검사자 간 신뢰도 측정을 위해 언어병리학 석사과정에 있는 대학원생 2인이 제2 평가자로 참여하여 독립적으로 측정치를 분석하였다. 그 결과, 모든 측정치에서의 상관계수가 .917~1.000이었다(all p<.01).

5. 결과 처리

본 연구에서는 IBM SPSS Statistics ver 28.0을 사용하여 통계분석을 실시하였다. 본 연구의 종속변수에 대해서 샤피로 윌크 검정(Shapiro-Wilk normality test)을 실시한 결과, 모든 종속변수가 정규분포를 보였다. 이에 따라, 시간, 운율, 조음 차원에서의 음향음성분석 및 청지각적 평가에서 합성음성과 실제음성의 파라미터 간에 유의한 차이가 있는지를 살펴보기 위해서 다변량분산분석(Multivariate Analysis of Variance: MANOVA)을 실시하였다.

Ⅲ. 연구 결과

1. 시간 차원의 분석

실제음성과 합성음성 간 전체말속도, 평균 휴지시간, 발화 당 휴지비율의 기술통계 결과는 Table 3과 같다. 다변량분산분석을 실시한 결과, 실제음성과 합성음성은 전체말속도($F_{(1, 30)}$=.822, p=.372), 평균 휴지시간($F_{(1, 30)}$=.255, p=.617)에서 유의한 차이가 나타나지 않았다. 휴지비율에서는 합성음성과 실제음성 간에 유의한 차이가 나타났으며, 합성음성의 휴지비율($F_{(1, 30)}$=4.238, p=.048)이 실제음성보다 유의미하게 길었다.

Table 3. Descriptive results of overall speech rate, average duration of pause, and portion of pauses in the synthesized voices and real voices

	Real voice	Synthesized voice
Overall speech rate	5.20 (.41)	5.35 (.53)
Average duration of pause	.47 (.21)	.50 (.08)
Portion of pause	10.22 (4.39)	13.01 (3.17)

Note. Values are presented as mean (*SD*).

2. 운율차원의 분석

실제음성과 합성음성 간 피치의 평균, 피치의 중앙, 피치의

범위의 기술통계 결과는 Table 4와 같다. 다변량분산분석을 실시한 결과, 실제음성과 합성음성은 피치 평균($F_{(1, 30)}$=2.773, p=.106), 피치 중앙값($F_{(1, 30)}$=2.334, p=.137), 피치 범위($F_{(1, 30)}$=4.114, p=.052)에서 유의한 차이를 보이지 않았다.

Table 4. Descriptive results of pitch-related variables in the synthesized voices and real voices

	Real voice	Synthesized voice
Pitch average	217.79 (17.02)	232.78 (31.73)
Pitch median	217.18 (21.07)	232.00 (32.58)
Pitch range	165.77 (44.08)	205.09 (63.81)

Note. Values are presented as mean (*SD*).

3. 조음차원의 분석

실제음성과 합성음성 간 F_2 기울기, V 모음길이 평균, CV 모음길이 평균, CVC 모음길이 평균, 모음면적의 기술통계 결과는 Table 5와 같다. 다변량분산분석을 실시한 결과, 실제음성과 합성음성은 F_2 기울기 평균($F_{(1, 30)}$=.621, p=.437), V 모음길이 평균($F_{(1, 30)}$=.066, p=.799), CV 모음길이 평균($F_{(1, 30)}$=2.258, p=.143)에서 유의한 차이가 없었다. CVC 모음길이 평균과 모음면적에서는 실제음성과 합성음성 간에 유의미한 차이가 나타났다. 합성음성의 CVC 모음길이 평균($F_{(1, 30)}$=4.423, p=.044)이 실제음성보다 유의미하게 길었고, 합성음성의 모음면적($F_{(1, 30)}$=6.770, p=.014)이 실제음성보다 유의하게 컸다.

Table 5. Descriptive results of F_2 slope, vowel duration in the V context, vowel duration in the CV context, vowel duration in the CVC context, and vowel space area in the synthesized voices and real voices

	Real voice	Synthesized voice
F_2 slope	7.54 (1.82)	8.00 (1.40)
Vowel duration in the V context	58.76 (9.85)	59.69 (10.73)
Vowel duration in the CV context	57.65 (10.18)	76.54 (49.25)
Vowel duration in the CVC context	42.85 (7.14)	49.02 (9.31)
Vowel space area	240973.49 (122536.32)	389227.43 (192174.36)

Note. Values are presented as mean (*SD*).

4. 청지각적 차원의 분석

실제음성과 합성음성 간 청지각적 차원의 특성(속도, 운율, 명료도, 부드러움, 친절, 호감도)을 살펴본 결과는 Table 6과 같다. 다변량분산분석 결과, 운율($F_{(1, 30)}$=3.516, p=.071), 친절($F_{(1, 30)}$=.731, p=.399), 호감도($F_{(1, 30)}$=3.922, p=.057)에서는 유의한 차이가 나타나지 않았지만, 속도($F_{(1, 30)}$=6.882, p=.014), 명료도($F_{(1, 30)}$=19.116, p=.002), 부드러움($F_{(1, 30)}$=12.125, p=.002)에서 유의한 차이가 나타났다.

Table 6. Descriptive results of speech rate, prosodic, speech intelligibility, softness, kindness, preference in the synthesized voices and real voices

	Real voice	Synthesized voice
Speech rate	3.85 (.70)	4.76 (1.20)
Prosody	4.17 (.59)	4.67 (.89)
Speech intelligibility	4.61 (.44)	5.28 (.42)
Softness	4.36 (.62)	3.64 (.56)
Kindness	4.15 (.62)	3.97 (.62)
Preference	4.13 (.53)	3.73 (.62)

Note. Values are presented as mean (*SD*).

Ⅳ. 논의 및 결론

본 연구는 20~30대 여성의 실제음성과 합성음성을 시간, 운율, 조음, 청지각적 차원의 분석을 통해서 실제음성과 합성음성의 특성을 살펴보고자 하였다. 그 결과, 시간차원에서는 합성음성의 휴지비율이 실제음성보다 유의하게 길었으며, 운율차원에서는 합성음성과 실제음성의 음도 간 유의미한 차이가 없었다. 조음차원에서는 합성음성의 CVC 음절구조에서의 모음길이의 평균이 실제음성보다 유의하게 길었으며, 합성음성의 모음면적이 실제음성보다 유의미하게 컸다. 청지각적 차원에서는 청자들이 합성음성의 말속도가 실제음성보다 유의하게 빠르고 명료하다고 지각하였으며, 실제음성이 합성음성보다 유의하게 부드럽다고 지각하였다. 연구결과에 대한 논의는 다음과 같다.

본 연구에서는 합성음성과 실제음성 간의 시간 차원에서의 특성을 살펴본 결과, 전체 말속도와 평균 휴지시간에서는 유의한 차이가 없었으나, 합성음성의 발화 당 휴지비율이 실제음성보다 유의하게 길었다. 이러한 결과는 여성 아나운서가 여성 TTS보다 휴지 비율이 길었다는 Yu(2019)의 연구 결과와 상반된 결과였다. 이러한 선행연구와 본 연구와의 결과 차이는 화자의 발화 특성과 관련이 있을 것으로 생각된다. 아나운서를 대상으로 발화를 수집한 Yu(2019)의 연구와는 달리 본 연구는 일반 여성을 대상으로 발화를 수집하였는데, 뉴스를 진행할 때 나타나는 아나운서의 발화 특성이 일반 여성의 발화 특성과 차이가 있었기 때문인 것으로 해석된다. 본 연구에서 주목할 점은 합성음성과 실제음성 간에 전체 말속도에서는 유의한 차이가 없었으나, 합성음성의 휴지비율은 실제음성보다 유의하게 길었다는 것이다. 이러한 결과는 합성음성의 조음 속도가 실제음성보다 빨랐다는 것을 간접적으로 시사한다. 본 연구의 청지각적 평가에서도 청자들은 합성음성의 말속도가 실제음성보다 유의하게 빠르다고 지각하였는데, 이러한 시간 차원의 음향음성 분석에서 합성음성의 휴지비율이 실제음성보다 유의하게 큰

것이 영향을 미쳤을 수 있을 것이다.

합성음성과 실제음성 간의 운율 차원에서의 특성을 살펴보면, 실제음성과 합성음성이 피치의 평균, 중앙, 범위 등에서 유의한 차이가 나지 않았다. 이러한 결과는 현재의 TTS 기술이 실제음성의 운율 특성을 상당히 반영하고 있다고 해석할 수 있다. 다시 말해서, TTS 기술 발전으로 인해서 합성음성이 기존 합성음성의 특징인 단조로운 음성에서 벗어나 실제음성과 유사한 운율형태를 보이는 것으로 해석할 수 있다.

조음 차원의 분석에서 살펴보면, V, CV 맥락에서는 실제음성과 합성음성 간 모음길이에 유의한 차이가 없었으나, CVC 맥락에서는 두 음성 간 모음길이에 유의한 차이가 있었다. 즉, CVC 맥락에서 합성음성의 모음길이가 실제음성보다 유의하게 길었다. 실제음성 산출 과정에서는 동시조음(coarticulation) 현상이 발생한다. 동시조음이란 두 개 이상의 음소가 같은 시간에 두 개 이상의 다른 음소들을 동시에 산출하기 위해 움직이는 것을 말한다(Ferrand, 2007). 여러 조음자를 순차적, 연속적으로 산출하는 상황에서는 동시조음 현상이 발생하며, 이 과정에서의 조음동작은 부드럽고 자연스러우며 빠르게 진행된다. 본 연구에서 CVC 맥락에서만 두 음성 간 모음길이에 차이가 발생한 것은 합성음성이 실제 조음 특성을 충분히 구현하지는 못한 것으로 해석할 수 있겠다. 반면에, 조음복잡성이 낮은 V, CV 맥락에서는 합성음성이 실제음성과 유사한 조음 특성을 구현하였다고 볼 수 있겠다. 실제음성에서는 화자의 조음과정에서 산출하는 음소들이 서로 의존적으로 영향을 미치기 때문에 음소별로 깔끔하게 나누어지지 않는 경향이 있지만(Park, 2015), 이러한 측면은 청지각적 평가에서 청자들이 합성음성보다 실제음성이 부드럽다고 지각하게 된 것과 관련될 수 있을 것이다. 그리고 본 연구에서는 합성음성의 모음면적이 실제음성보다 유의하게 컸다. 모음면적은 말 명료도와 상관관계가 있으며(Turner et al., 1995), 말 명료도에 대한 음향학적인 객관적인 평가에 많이 사용되고 있다(Han et al., 2013; Lee et al., 2010; Park & Hugh, 2014). 추가적으로 말 명료도에는 시간적 차원에서의 휴지비율이 영향을 미쳤을 것으로 예상된다. 합성음성의 휴지비율이 실제음성보다 유의하게 길게 나타났고, 이는 합성음성의 조음 속도가 실제음성보다 빨라서 청자들로 하여금 말의 내용을 처리하는 데 용이하였다. 이러한 결과는 청지각적 평가에서 청자들이 합성음성의 명료도가 실제음성보다 유의하게 높다고 지각한 것도 관련된 것으로 해석할 수 있다.

청지각적 평가 결과, 청자들은 합성음성의 말속도가 실제음성보다 유의하게 빠르고, 명료하다고 지각하였으며, 실제음성이 합성 음성보다 더 부드럽다고 지각하였다. 호감도 부분에서는 합성음성과 실제음성 간의 유의한 차이가 나타나지 않았는데, 이는 청지각적 평가의 대상자인 20대 남·녀가 합성음성에 익숙한 세대로서 합성음성에 대한 거부감이 없는 것과 관련되었을 것으로 생각된다. 현재 합성음성이 주로 사용되는 상황은 내비게이션, 스마트폰과 같은 명확한 정보 전달을 목적과 관련된 상황이므로, 합성음성에서 가장 중요한 요소는 명료성이라 할 수 있다. 그러나 화자는 정보 전달 외에도 타인과의 상호작용을 위한 목적으로 발화를 산출하므로 합성음성에서는 구현되지 못하는 정서적 표현, 정교한 억양 표현이 가능하였을 것이다. 이러한 이유로 인해서 실제음성의 부드러움 요소가 합성음성보다 유의하게 큰 것으로 평정되었을 것으로 생각된다.

본 연구의 결과에 대한 합성음성의 언어병리학적 해석과 제안은 다음과 같다. 첫째, 본 연구는 다차원적 분석을 통해 다양한 합성음성을 실제음성과 비교 분석하였다. 본 연구는 합성음성을 실제음성과 더 유사하게 구현하기 위한 다차원적 분석의 기본적인 토대를 마련하였다. 따라서 더 다양한 합성음성의 다차원적인 분석과 이를 토대로 더 정교한 TTS의 기술 발전이 가능할 것으로 생각된다. 둘째, 본 연구는 청지각적 평가를 통해 합성음성에 대한 청자들의 세부적인 인식의 차이를 밝혀내었다. 이를 통해 합성음성이 정보 전달 외에도 상호작용에도 초점을 둘 수 있도록, 음성의 부드러움을 강화하기 위한 초분절적인 음향음성 파라미터의 개선을 제안한다. 셋째, 본 연구는 합성음성과 실제음성 간의 조음과 관련된 음향음성 측정치 간의 차이를 발견하였다. 이에 본 연구 결과를 토대로 TTS 기술을 개발하는 회사들에게 공학적인 부분에서 조음 관련 파라미터를 고려할 것을 제안한다. 넷째, 본 연구의 결과를 통해 청자들은 합성음성를 실제음성과 비슷한 호감도를 가지며 더 명료하게 지각하였다. 이러한 결과는 의사소통장애 아동 및 성인의 치료에 합성음성이 사용될 가능성이 있음을 시사한다. TTS 기술의 저비용 고효율의 강점을 살려, 언어병리학 분야에서 다양한 교재 및 도구 제작에 유용하게 사용할 것을 제안한다.

본 연구에 대한 제한점 및 향후 연구 제언은 다음과 같다.

첫째, 본 연구의 합성음성 및 실제음성의 표본의 수에 제한이 있어 모든 합성음성에 대한 일반화에 제한이 있다. 따라서 추후 연구에서는 더 많은 표본의 다양한 합성음성으로 분석할 것을 제언한다. 둘째, 본 연구에서는 청지각적 대상이 20~30대로 제한하였다. 본 연구의 청지각적 평가의 대상인 20~30대는 다른 연령층과 비교하여 합성음성에 익숙하다. 따라서 추후 연구에서는 다양한 연령층을 대상으로 청지각적 평가를 진행할 것을 제언한다. 셋째, 본 연구의 음성 분석은 20~30대 여성만을 대상으로 진행했다는 제한이 있다. 합성음성에는 현재 다양한 연령과 성별이 있다. 따라서 추후 연구에서는 다양한 연령과 성별을 대상으로 음성을 분석할 것을 제언한다. 넷째, 본 연구는 실제음성과 합성음성의 일부만 수집하여, '같다-다르다'의 이분법적인 차원으로 결과를 도출한 것에 대한 제한이 있다. 따라서 추후 연구에서는 실제음성과 합성음성의 다양하고 변이적인 특성을 고려하여 자연스러움, 호감도, 명료도에 따른 실제음성과 합성음성의 음향적인 특성을 추가적으로 분석할 것을 제언한다. 다섯째, 본 연구의 청지각적 평가에서 6가지 항목 중 '부드러움', '친절', '호감도'의 3가지 항목은 모두 호감이라는 공통적인 속성을 가지고 있어 평가자들로 하여금 혼란을 느끼게 했을 수 있다. 향후 연구에서는 음성의 호감도 하위 항목에 대한 보다 더 구체적인 기준과 방법으로 연구를 진행할 것을 제언한다.

본 연구의 결과를 바탕으로, 보다 자연스러운 합성음성 구현을 위해 제언하는 바는 다음과 같다. 첫째, 합성음성을 실제음성과 유

사하게 구현하기 위해 여성 화자의 음성을 다양한 음도로 반영한다면 더 넓은 범주의 여성 화자의 목소리를 대변할 수 있을 것이다. 둘째, 합성음성의 경우에도 공적인 상황과 더불어 사적인 상황에 맞게 부드러운 음성으로의 개선이 필요하다.

마지막으로, 본 연구의 결과를 토대로 임상적 측면에서 기대되는 바는 다음과 같다. 첫째, 보완대체의사소통(augmentative and alternative communication: AAC) 기기의 부자연스러운 합성음성을 자연스러운 음성으로 개선함으로써 AAC 사용자들의 의사소통 질과 만족도가 향상될 수 있을 것으로 기대된다. 둘째, AI를 활용한 언어 중재에 도움이 될 것으로 기대된다. AI를 활용한 상호작용 중재에서 지적장애 학생은 의사소통 의도와 화용론적 특성이 향상되었으며(Kim & Jeong, 2020), 학령전기 아동들은 대화 차례 주고받기에서 향상을 보였다(Park & Yim, 2021). 또한, 주관적 기억장애(subjective memory impairment)은 AI를 활용한 의사소통 교류가 기억력, 이름 대기, 주의집중, 우울감에서 향상을 보였다(Park, 2020). 이에 AI를 활용한 의사소통 중재에서 AI의 자연스러운 합성음성 사용으로 의사소통 중재의 효과가 더 향상될 것으로 기대된다. 셋째, Choi 등(2021)의 연구는 아동의 자발화 수집을 위한 챗봇을 제안하였다. 이에 자연스러운 합성음성을 가진 챗봇을 통하여 아동의 자발화를 효율적으로 수집할 수 있을 것으로 기대된다.

Reference

Byeon, H. (2009). Comparing the acoustic character of diphthong production between flaccid dysarthria and spastic dysarthria. *Korean Journal of Communication Disorders, 15*(1), 66-78. uci:G704-000725.2010.15.1.007

Cho, Y. S., Pyo, H. Y., Han, J. S., & Lee, E. J. (2021). Acoustic features of diphthongs produced by children with speech sound disorders. *Phonetics and Speech Sciences, 13*(1), 65-72.

Choi, J. M., Lee, Y. K., & Kim, Y. S. (2021). Conversation data collection chatbot for language development assessment of children. *Korea Computer Congress 2021*, 319-321.

Choi, M. K. (2011). *F_2 slopes of Korean diphthongs: Comparison among younger, middle-aged, and older adults* (Master's thesis). Yonsei University, Seoul.

Choi, Y., Yeo, S., Kim, J., & Sung, K. (1998). Pitch modification based on a voice source model. *Speech Sciences, 3*, 132-147.

Ferrand, C. T. (2007). *Speech science: An integrated approach to theory and clinical practice* (2nd ed.). Boston: Pearson/Allyn & Bacon.

Ferguson, S. H., & Kewley-Port, D. (2007). Talker differences in clear and conversational speech: Acoustic characteristics or vowels. *Journal of Speech, Language and Hearing Research, 50*(5), 1241-1255. doi:10.1044/1092-4388(2007/087)

Gown, S., Maeng, J., Baek, Y., & Kim, Y. K. (2021). Comparison of speech synthesis deep learning models for personalized TTS. *Proceedings of Korean Institute of Information Technology Conference* (pp. 759-762), Jeju, Korea.

Han, J., Sung, J., Shim, H., & Lee, Y. (2013). Effects of speaking rate nanipulation and the severity of dysarthria on speech intelligibility and acoustic parameters in persons with cerebral palsy. *Journal of Speech-Language & Hearing Disorders, 22*(1), 35-54. doi:10.15724/jslhd.2013.22.1.003

Higgins, C. M., & Hodge, M. M. (2002). Vowel area and intelligibility in children with and without dysarthria. *Journal of Medical Speech-Language Pathology, 10*(4), 271-277.

Jung, J. (1994). Development of a signal processing algorithm to improve the naturalness of synthetic sound. *Korea Institute of Communication Sciences*, 1-108.

Kim, D. I., & Jeong, E. H. (2020). The effect of interaction using artificial intelligence speakers on communication intention and pragmatic characteristics of students with intellectual disabilities. *The Study of Education for Hearing-Language Impairments, 11*(3), 91-113. doi:10.24009/ksehli.2020.11.3.005

Kim, H. H. (1996). *Perceptual, acoustical, and physiological tools in ataxic dysarthria management: A case report. In Proceedings of the KSPS Conference* (pp. 9-22). The Korean Society Of Phonetic Sciences and Speech Technology.

Kim, J., & Kwon, C. (2014). Qualitative classification of voice quality of normal speech and derivation of its correlation with speech features. *Phonetics and Speech Sciences, 6*(1), 71-76. doi:10.13064/KSSS.2014.6.1.071

Kim, Y., Kent, R. D., & Weismer, G. (2011). An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria. *Journal of Speech, Language, and Hearing Research, 54*(2), 417-429. doi:10.1044/1092-4388(2010/10-0020)

Kim, Y. J., Weismer, G., Kent, R. D., & Duffy, J. R. (2009). Statistical models of F2 slope in relation to severity of dysarthria. *Folia Phoniatrica et Logopaedica, 61*(6), 329-335. doi:10.1159/000252849

Kong, K. S., Yoon, K. H., & Kim, C. H. (1997). Coarticulated model for Hangul lip animation. *The Korean Institute of Information Scientists and Engineers, 24*(2 II), 603-606.

Kwon, S. (2015). An experimental study of favorable voice analysis and good impressions using paralinguistic construction elements. *Journal of Speech-Language & Hearing Disorders, 24*(1), 157-167. doi:10.15724/jslhd.2015.24.1.013

Lee, H., & Hong, K. (2013). A study of Korean TTS listening speed for the blind using a screen reader. *Phonetics and Speech Sciences, 5*(3), 63-69. doi:10.13064/KSSS.2013.5.3.063

Lee, H. J. (2012). *F_2 slopes of Korean diphthongs: Comparison between Parkinson's disease groups and normal groups* (Master's thesis). Yonsei University, Seoul.

Lee, J., Kang, E., & Lee, O. (2016). A study of vocabulary intelligibility and speech naturalness of children with hearing impairment. *Journal of Speech-Language & Hearing Disorders, 25*(4), 273-282. doi:10.15724/jslhd.2016.25.4.021

Lee, J., Park, K., & Lee, J. (2014). Current status of speech synthesis technology for the visually and speech impaired. *The Magazine*

of the IEIE, 41(3), 28-39.

Lee, O., Han, J., & Park, S. (2010). Speech intelligibility in syllables and vowel space according to dysarthric severity. *Phonetics and Speech Sciences, 2*(2), 85-92. uci:G704-SER000000671.2010.2.2.003

Lee, S. H. (2020). A study on the paralanguage through the examination of main message in persuasive speech. *The Journal of Linguistics Science, 93*, 147-181.

Nam, H., & Choi, Y. (2008). Distributions on F_0 and amplitude of persons with cerebral palsy in the reading task. *Malsori, 1*(66), 1-20. uci:G704-001087.2008.1.66.001

Park, C. H. (2020). *A study on the effect of AI speaker on the performance of language and psychological behavior test of adults with subjective memory impairment* (Master's thesis). Daegu University, Gyeongbuk.

Park, C. J., Kim, S. T., Kang, M. H., Hwang, M. J., & Kim, Y. S. (2018). Voice recognition performance improvement using TTS system in address recognition system. *Proceedings of Korea Computer Congress 2022* (pp. 551-553).

Park, H., & Hugh, M. (2014). Vowel space area and speech intelligibility of children with cochlear implants. *Phonetics and Speech Sciences, 6*(2), 89-96. doi:10.13064/KSSS.2014.6.2.089

Park, M. (2015). *Characteristics of coarticulation in childhood apraxia of speech* (Master's thesis). Yonsei University, Seoul.

Park, W., & Yim, D. (2021). Effects of using a communication maintenance strategy in the context of AI speaker and preschoolers' conversation and book reading interaction: Comparison of group differences on the levels of expressive language development. *Journal of Speech-Language & Hearing Disorders, 30*(2), 1-8. doi:10.15724/jslhd.2021.30.2.001

Shin, J. (2018). Breath and memory in speech based on quantitative analysis of breath groups and pause units in Korean. *Korean Linguistics, 79,* 91-116. doi:10.20405/kl.2018.05.79.91

Song, H. J., Jung, H. Y., & Park, J. G. (2015). A study of DNN training based on various pretraining approaches. *Proceedings of the 2015 Spring Conference of the Korean Society of Speech Sciences* (pp. 169-170).

Turner, G. S., Tjaden, K., & Weismer, G. (1995). The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. *Journal of Speech Hearing Reserch, 38*(5), 1001-1013. doi:10.1044/jshr.3805.1001

Yu, J. (2019). *Comparison of TTS (text-to-speech) devices' and human anouncers' paralinguistic characteristics and audience perceptions about paralinguistic characteristics* (Doctoral dissertation). Sungkyunkwan University, Seoul.

언어치료연구 2022년 제31권 제4호 081-090
ISSN 1226-587X / eISSN 2671-7158
JSLHD (Journal of Speech-Language & Hearing Disorders)

ORIGINAL ARTICLE
https://doi.org/10.15724/jslhd.2022.31.4.081
http://jslhd.org/

다차원적 음성 프로파일 분석을 통한 20-30대 한국 여성의 합성음성과 실제음성의 특성 비교

권순하[1], 전예솔[1], 유성아[1], 오유림[1], 이영미[2*]

[1] 이화여자대학교 일반대학원 언어병리학과 석사과정
[2] 이화여자대학교 일반대학원 언어병리학과 교수

목적: 본 논문은 여성의 합성음성과 실제음성을 다차원적으로 분석하여 그 차이를 알아내는 데 목적을 가진다.
방법: 23~35세의 성인 여성 16명이 본 연구에 참여하였고 대상자들은 3개의 문장을 읽고, 녹음하였다. 합성음성을 제공하는 네이버 클로바 더빙과 타입캐스트 프로그램에서 여성 청년 목소리에 해당하는 16개의 합성음성을 선정하였고 실제음성이 녹음한 것과 같은 문장으로 작성하여 녹음파일을 추출하였다. 그 후, 총 32개의 음성을 시간, 운율, 조음 측면, 그리고 청지각적 평정점수에 따라 분석하였다.
결과: 첫째, 시간차원에서 합성음성의 휴지비율이 실제음성에 비해 유의미하게 길었다. 둘째, 운율차원에서는 합성음성과 실제음성 간의 유의미한 차이가 없었다. 셋째, 조음차원을 살펴보았을 때, CVC 구조의 문장에서 합성음성의 모음 지속시간은 실제음성에 비해 유의미하게 길었으며, 모음면적에서도 합성음성이 실제음성보다 유의하게 크게 나타났다. 마지막으로, 청지각적 평가 결과 청자들은 합성음성에 비해 실제음성을 더 부드럽다고 지각하였고, 실제음성에 비해 합성음성을 더 빠르고 명료하게 인식하였다.
결론: 본 연구는 여성의 합성음성이 시간, 조음, 청지각적 차원에서 여성의 실제음성과 다른 특성을 보인다는 것을 발견하였다. 실제음성에 비해 합성음성은 더 빠른 말속도, 더 명료한 발음으로 여성 청년의 음성을 나타내었다. 본 연구를 통해 TTS 기술이 실제음성을 구현하는데 어떠한 측면을 보완해야 하는지 제시하고 나아가 언어병리학 분야에서 다양하게 활용될 수 있는 방법을 모색하는데 본 연구 결과가 기초자료로 제공될 수 있을 것이다.

검색어: 합성음성, 실제음성, 음향학적 평가, 청지각적 평가, 다차원적 분석

교신저자 : 이영미(이화여자대학교)
전자메일 : youngmee@ewha.ac.kr

게재신청일 : 2022. 07. 08
수정제출일 : 2022. 10. 01
게재확정일 : 2022. 10. 31

ORCID
권순하
https://orcid.org/0000-0003-0199-8389
전예솔
https://orcid.org/0000-0001-8988-3222
유성아
https://orcid.org/0000-0001-6810-8745
오유림
https://orcid.org/0000-0001-7104-7787
이영미
https://orcid.org/0000-0003-1809-5944

참 고 문 헌

공광식, 윤광희, 김창헌 (1997). 동시조음을 적용한 한글 발음 입술 애니메이션. **한국정보과학회 1997 학술발표논문집, 24**(2 II), 603-606.

권세영, 맹지연, 백예슬, 김영국 (2021). 개인화된 TTS 구현을 위한 음성합성 딥러닝 모델 비교. **한국정보기술학회 2021 추계 종합학술대회 논문집,** 759-762.

권순복 (2015). 준언어적 구성 요소를 통한 매력적인 목소리 분석과 호감도에 관한 실험 연구. **언어치료연구, 24**(1), 157-167.

김동인, 정은희 (2020). 인공지능 스피커를 활용한 상호작용이 지적장애 학생의 의사소통의도와 화용론적 특성에 미치는 영향. **한국청각언어장애교육연구, 11**(3), 91-113.

김정인, 권철홍 (2014). 정상 음성의 목소리 특성의 정성적 분류와 음성 특징과의 상관관계 도출. **말소리와 음성과학, 6**(1), 71-76.

김향희 (1996). 운동실조형 마비성구음장애에 적용되는 지각적, 음향학적, 생리학적 도구에 관하여: 환자 사례를 중심으로. **제2회 음성학 학술대회 자료집,** 9-22.

남현욱, 최양규 (2008). 읽기 과제에서 나타난 뇌성마비인의 기본 주파수 및 진폭의 분포 특성. **말소리, 1**(66), 1-20.

박민수 (2015). **아동기 말실행증 아동의 동시조음 특성.** 연세대학교 대학원 석사학위 논문.

박원정, 임동선 (2021). AI 스피커의 상호작용 유지 전략 사용 여부가 아동 발화 및 이야기 이해 수행에 미치는 영향: 표현 언어발달 수준에 따른 차이 비교 연구. **언어치료연구, 30**(2), 1-8.

박찬준, 김선태, 강민호, 황명진, 박병건, 김연수 (2018). 주소인식 시스템에서 TTS를 활용한 음성인식 성능 개선. **한국정보과학회 2018 학술발표 논문집,** 551-553.

박철형 (2020). **인공지능 스피커를 활용한 의사소통 교류가 주관적 기억장애 장년층의 언어 및 심리·행동검사 수행에 미치는 영향**. 대구대학교 대학원 석사학위 논문.

박혜미, 허명진 (2014). 인공와우이식 아동의 모음공간면적과 말명료도. **말소리와 음성과학, 6**(2), 89-96.

변해원 (2009). 이완형과 경직형 마비말장애 남성의 이중모음 /야, 위, 의/의 음향음성학적 특성. **언어청각장애연구, 15**(1), 66-78.

송화전, 정호영, 박전규 (2015). 다양한 Pretraining 방법에 따른 DNN 훈련 방법에 대한 고찰. **한국음성학회 2015 봄 학술대회 논문집,** 169-170.

신지영 (2018). 언어 수행에서의 호흡과 기억: 호흡 단위와 휴지 단위의 양적 분석 결과를 바탕으로. **한국어학, 79**, 91-116.

유도영, 신지영 (2020). 휴지 단위와 호흡 단위의 실현 양상 연구. **말소리와 음성과학, 12**(1), 19-31.

유지철 (2019). **뉴스 낭독에 나타난 TTS(음성합성기)와 아나운서의 준언어 특성 비교 및 준언어에 대한 수용자 인식에 관한 연구.** 성균관대학교 대학원 박사학위 논문.

이소현 (2020). 설득적 말하기의 핵심메시지 발화에 나타나는 준언어 연구: 속도와 휴지를 중심으로. **언어과학연구, 93**, 147-181.

이옥분, 한지연, 박상희 (2010). 마비말장애 심각도에 따른 음절단위 말명료도와 모음공간. **말소리와 음성과학, 2**(2), 85-92.

이종석, 박기태, 이준우 (2014). 시각 및 언어장애인을 위한 음성합성 기술의 현황. **대한전자공학회, 41**(3), 28-39.

이지윤, 강은희, 이옥분 (2016). 학령기 청각장애아동의 단어 명료도 및 말 자연스러움 특성 연구. **언어치료연구, 25**(4), 273-282.

이혜진 (2012). **이중모음에서의 제2 포먼트 기울기(F_2 slope): 파킨슨병 환자군과 정상군 간의 비교.** 연세대학교 대학원 석사학위 논문.

이희연, 홍기형 (2013). 스크린리더를 사용하는 시각장애인의 한국어 합성음 청취속도 연구. **말소리와 음성과학, 5**(3), 63-69.

정재호 (1994). **합성음의 자연성 향상을 위한 신호처리 알고리즘의 개발**. 한국통신학회 전기통신학술연구과제, 1-108.

정현성, 장태엽, 윤원희, 윤일승, 사재진 (2008). 한국인 영어 학습자의 발음 정확성 자동 측정방법에 대한 연구. **언어와 언어학, 42**, 165-196.

조윤수, 표화영, 한진순, 이은주 (2021). 말소리장애 아동이 산출한 이중모음의 음향학적 특성. **말소리와 음성과학, 13**(1), 65-72.

최민경 (2012). **이중모음에서의 제2 포먼트 기울기 지수: 청년층, 장년층, 노년층 간의 비교**. 연세대학교 대학원 석사학위 논문.

최정명, 이윤경, 김유섭 (2021). 아동 언어 발달 평가를 위한 대화 자료 수집 챗봇. **한국정보과학회 2021 한국컴퓨터종합학술대회 논문집**, 319-321.

최용진, 여수진, 김진영, 성광모 (1998). 음원 모델에 기초한 합성음의 피치 조절. **음성과학, 3**, 132-147.

한지후, 성지은, 심현섭, 이영미 (2013). 말속도 조절 및 중증도가 마비말장애 화자의 말명료도와 음향학적 파라미터에 미치는 영향. **언어치료연구, 22**(1), 35-54.