ISSN 2005-9299

Proceedings

of the 2024 Autumn Conference of the Korean Society of Speech Sciences

2024. 10. 19.

Hankuk University of Foreign Studies

The Korean Society of Speech Sciences

ISSN 2005-9299

2024 한국음성학회 가을 학술대회
발표 논문집

주제: 언어의 경계: 인간언어와 기계언어의 음성과학
일시: 2024년 10월 19일(토)
장소: 한국외국어대학교 일반대학원 1층 BRICs 문화관
주최: 사단법인 한국음성학회
주관: 사단법인 한국음성학회, 한국외국어대학교
후원: (주)리드스피커코리아, (주)셀바스AI, (주)카카오뱅크

한 국 음 성 학 회
The Korean Society of Speech Sciences

2024년 한국음성학회 가을 학술대회

모 시 는 글

유난했던 올해의 여름 무더위에 많이 힘드셨지요? 그래도 어김없이 가을이 왔습니다. 한국음성학회는 10월 19일, 토요일에 한국외국어대학교 서울캠퍼스에서 2024년 가을 학술대회를 개최하게 되었습니다. 주로 11월 말경에 개최되었던 예년보다 올해는 조금 일찍 준비하였습니다.

가을 학술대회는 우리 학회의 고유 연구 영역인 음성학, 음성공학, 말장애 분야의 최신 연구 동향을 공유하고, 함께 더 나은 미래를 모색하고, 학회의 한해 활동을 마무리하면서 총회를 개최하는 자리이기도 합니다. 최근 다양한 생성형 AI 기술의 발전으로 인해 음성 자료에 대한 연구와 활용 및 교육 방법에도 많은 변화가 있었습니다. 이러한 변화에 발맞추어 이번 학술대회의 주제를 '언어의 경계: 인간언어와 기계언어의 음성과학'으로 정하고 관련된 세 편의 전문가 초청 강연을 모시게 되었으며 두 세션의 병렬 구두 발표 일정에서 14편의 연구가, 그리고 10편의 흥미 있는 포스터 발표가 소개될 것입니다. 흔쾌히 초청 강연을 수락해 주신 분들과 소중한 연구 성과를 발표해 주시는 회원 여러분께 진심으로 감사의 말씀을 전합니다. 또한, 이번 학술대회를 위해 성심성의를 다해 준비해 주신 박상희 학술위원장님을 비롯한 여러 준비위원님과 여러 가지 어려운 문제를 해결해 주신 정현성 수석부회장님, 그리고 회장단 및 모든 관계자분께도 깊은 감사의 인사를 드립니다. 여러분의 노고와 열정 덕분에 이 행사를 준비할 수 있었습니다. 마지막으로 늘 변함없이 후원해 주고 계시는 ㈜리드스피커코리아, ㈜셀바스AI, 그리고 이번에 새로 후원에 참여 해주시는 ㈜카카오뱅크에도 특별한 감사를 표합니다.

이번 학술대회에 참여하신 분들을 진심으로 환영합니다. 함께 풍부한 지식과 유익한 경험을 나누는 이 특별한 시간을 즐겨주시길 바랍니다. 감사합니다.

2024년 10월 19일

한국음성학회 회장 장 태 엽

2024년 한국음성학회 가을 학술대회

조 직 위 원 회

학술대회장	회장	장태엽(한국외대)
조직위원회	위원장	정현성(수석부회장, 한국교원대)
	위 원	김재옥(강남대)
		박상희(대구사이버대)
		신동진(전주대)
		오은혜(건국대)
		윤영선(한남대)
		윤원희(계명대)
		이주경(서울시립대)
학술위원회	위원장	박상희(대구사이버대)
	위 원	공은정(한국항공대)
		김유경(순천제일대)
		김지연(우송대)
		박기영(한국전자통신연구원)
		오재혁(건국대)
		유하진(서울시립대)
		이현정(인천대)
편집위원회	위원장	김유경(순천제일대)
	위 원	배미영(한국외대)

한국음성학회

The Korean Society of Speech Sciences

INFORMATION

2024 한국음성학회 가을 학술대회 일정표

1) 전체 일정표

일정	발 표 및 내 용		
09:00~09:50	등록		
09:50~10:00	개회식 (사회: 신동진, 전주대)		
10:00~10:50	특강 1. 운율구조의언어보편성과 음성문법 : 어두 모음 성문화 현상을 중심으로 조태홍(한양대) [특강자료로] (사회: 이주경, 서울시립대)		
10:50~11:00	휴식		
11:00~12:15	구두 발표 I		
	분과 1 음성학 및 음운론	분과 2 음성공학	분과 3 말장애 및 음성의학
12:15~13:00	점심 식사		
13:00~14:00	포스터 발표		
14:00~14:50	특강 2. 장애음성의 기록과 분석:기기에서 프로그램까지 이승진(한림대) [특강자료로] (사회: 이영미, 이화여대)		
14:50~15:00	휴식		
15:00~16:15	구두 발표 II		
	분과 4 음성학 및 음운론	분과 5 음성공학	
16:15~16:25	휴식		
16:25~17:15	특강 3. 음성 딥페이크 탐지 현황 유하진(서울시립대) [특강자료로] (사회: 김선만, 한신대)		
17:15~17:30	스펙트로그램 reading 우수 발표 시상 및 폐회식 (사회: 박상희, 대구사이버대)		
17:30~18:00	정기총회 사회(신동진, 전주대)		

2) 발표 내용

● 구두 발표 I

	좌장: 윤원희(계명대)			
	시간	구분	제목	발표자
<분과 1> 음성학 및 음운론	11:00~11:25	PH1	국립국어원의 “일상대화 말뭉치” 설계 및 구축 [논문으로]	이용주, 차원철(나라지식정보), 이광수(스마트미디어테크), 강미영, 서샛별 (국립국어원)
	11:25~11:50	PH2	재위 기간 중 영국 엘리자베스 2세 국왕의 단순 모음 변화 [논문으로]	정현성(한국교원대학교)
	11:50~12:15	PH3	모음 음향 특징을 활용한 충남-전북 방언 구획 연구 [논문으로]	이주영, 김선희, 정민화 (서울대학교)
	좌장: 박기영(ETRI)			
	시간	구분	제목	발표자
<분과 2> 음성 공학	11:00~11:25	SE1	비원어민 한국어 발음평가를 위한 자기 지도 학습 기반 한국어 음소인식 [논문으로]	경나, 나희정, 박지은, 박정식(한국외국어대학교)
	11:25~11:50	SE2	미지의 공격에 강인한 음성 딥페이크 탐지 모델 [논문으로]	손지수, 김승빈, 허정우, 신현서, 임찬영, 구교원, 유하진(서울시립대학교)
	11:50~12:15	SE3	효율적인 음성 인식을 위한 조기 종료 기법 [논문으로]	윤지원(중앙대학교)
	좌장: 박진(가톨릭관동대)			
	시간	구분	제목	발표자
<분과 3> 말장애 및 음성 의학	11:00~11:25	SD1	시선추적 기반 인공와우이식 아동의 청각 단어재인 특성 [논문으로]	이영미(이화여자대학교), 윤해나, 최병윤 (서울대학교)
	11:25~11:50	SD2	인공와우이식 유아의 자발화 음성 분석에 대한 종단 연구: 예비 보고 [논문으로]	이승진(한림대학교), 이영미(이화여자대학교)
	11:50~12:15	SD3	시선추적장비를 활용한 일반 아동의 단어재인 양상 [논문으로]	김지연, 이나경, 이영미 (이화여자대학교)

● 구두 발표 II

	좌장: 이현정(인천대)			
	시간	구분	제목	발표자
<분과 4> 음성학 및 음운론	15:00~15:25	PH4	The Relationship Between Vowel Epenthesis and Stress-Focus Dynamics in L2 Speech Perception [논문으로]	Goun Lee (Dankook Univ.), Dong-Jin Shin (Jeonju Univ.)
	15:25~15:50	PH5	한국어 비격식체 종결어미와 억양의 상호작용 [논문으로]	김기유(계명대학교)
	15:50~16:15	PH6	Refining Cuteness: Fine-tuned ChatGPT evaluations of Korean aegyo [논문으로]	Rok Sim (Univ. of South Carolina), Drew Crosby (Korea Univ.)
	좌장: 박형민(서강대)			
	시간	구분	제목	발표자
<분과 5> 음성 공학	15:00~15:25	SE4	Apple 스마트폰 음성 녹음 파일의 기종, OS 추측 및 위변조 식별을 위한 메타데이터 기반 결정 트리 [논문으로]	문수호, 윤나영, 이지원(성신여자대학교), 김경화(대검찰청), 홍기형(성신여자대학교)
	15:25~15:50	SE5	웨어러블 청각 장치 적용을 위한 종단간 음성 디노이징 신경망 연구 [논문으로]	김선만(한신대학교)

● 포스터 발표

좌장: 공은정(한국항공대), 김유경(순천제일대), 윤지원(중앙대)			
시간	구분	발표 제목	발표자
13:00 ~ 14:00	P01	Acoustic cue weighting in lexical stress perception of real and nonce words by Korean learners of English [논문으로]	Eunkyung Sung (Cyber Hankuk Univ. of Foreign Studies), Seulgi Oh (Hankuk Univ. of Foreign Studies)
	P02	Prospective English teachers' cue-weightings in L2 stop perception [논문으로]	Hyunjung Lee (Incheon National Univ.), Eun Jong Kong (Korea Aerospace Univ.)
	P03	일상에서의 자가 인지기능 평가 문항개발을 위한 예비연구 [논문으로]	박상희, 이옥분 (대구사이버대학교)
	P04	인공와우이식 영유아 어머니의 아동지향어 변화에 관한 종단연구 [논문으로]	이누리(이화여자대학교), 박희선(김포연세언어치료교육원), 이영미(이화여자대학교)
	P05	생후 12개월 이전에 인공와우이식을 받은 선천성 농 영유아의 발성 발달 [논문으로]	이나경(이화여자대학교), 전예슬(보아스목소리언어연구소), 이영미(이화여자대학교)
	P06	멀티모달 접근법을 통한 L2 영어 말하기 평가 모델 개발 [논문으로]	오세현, 김선희, 정민화 (서울대학교)
	P07	인공지능 기반 말더듬 반복과 연장에 대한 자동 식별: 합성곱층신경망(CNN)을 이용하여 [논문으로]	박진, 이창균 (가톨릭관동대학교)
	P08	BSRNN과 HiFi-GAN을 이용한 디클리핑 방법 [논문으로]	최승운, 문승현, 최승호 (서울과학기술대학교)
	P09	후처리 필터를 이용한 HiFi-GAN 생성 음성의 음질개선 [논문으로]	손세현, 전민서, 황규찬, 최승호(서울과학기술대학교)
	P10	잡음 모델링을 이용한 잡음 인지 음성 향상 기법 [논문으로]	김장연, 신의협(서강대학교), 김경화(대검찰청), 박형민(서강대학교)

차 례

특 강

특강1 운율구조의언어보편성과 음성문법: 어두 모음 성문화 현상을 중심으로

조태홍(한양대) ····· 2

특강2 장애음성의 기록과 분석:기기에서 프로그램까지

이승진(한림대) ····· 32

특강3 음성 딥페이크 탐지 현황

유하진(서울시립대) ····· 78

구두 발표

PH 1 Design and construction of the National Institute of Korean Language's "Daily Conversation Corpus"

Yongju Lee, Wonchul Cha, Kwangsu Lee, Miyoung Kang, Saetbyol Seo ····· 94

PH 2 Monophthongal vowel changes of Queen Elizabeth II during her reign

Hyunsong Chung ····· 95

PH 3 Dialect Division Study on Chungcheongnam-do and Jeollabuk-do Dialects using Vowel Acoustic Features

Jooyoung Lee, Sunhee Kim, Minhwa Chung ····· 97

PH 4 The Relationship Between Vowel Epenthesis and Stress-Focus Dynamics in L2 Speech Perception

Goun Lee, Dong-Jin Shin ····· 100

PH 5 한국어 비격식체 종결어미와 억양의 상호작용

김기유 ····· 101

PH 6 Refining Cuteness: Fine-tuned ChatGPT evaluations of Korean aegyo

Rok Sim, Drew Crosby ····· 102

SE 1 Self-supervised Learning Based Korean Phoneme Recognition for Non-native Korean Pronunciation Evaluation

Na Geng, Hee-jung Na, Ji-eun Park, Jeong-Sik Park ····· 104

SE 2 Robust Audio Deepfake Detection Model Against Unknown Attacks

Jisoo Son, Seung-bin Kim, Jungwoo Heo, Hyun-seo Shin, Chan-yeong Lim, Kyo-won Koo, Ha-Jin Yu ····· 106

SE 3 Early Exit Technique for Efficient Speech Recognition

Ji Won Yoon ····· 108

SE 4 Metadata-Based Decision Tree for Inferring Device Models, OS Versions, and Detecting Forgery in Apple Smartphone Voice Recordings

Suho Moon, Nayoung Yoon, Jiwon Lee, Kyungwha Kim, Kihyung Hong ····· 109

SE 5 End-to-End Speech Denoising Neural Network for Wearable Hearing Device Applications

Seon Man Kim ····· 111

SD 1 Real-time Spoken Word Recognition Processing in Young Children with Cochlear Implants:

Youngmee Lee, Haena Yoon, Byung Yoon Choi ····· 112

SD 2 A Longitudinal Study on the Voice Analysis of Spontaneous Speech in Children with Cochlear Implants: A Preliminary Report

Seung Jin Lee, Youngmee Lee ····· 113

SD 3 Real-time Spoken Word Processing in Children with Typical Hearing

Jiyeon Kim, Nakyoung Lee, Youngmee Lee ····· 115

포스터 발표

P01 Acoustic cue weighting in lexical stress perception of real and nonce words by Korean learners of English

Eunkyung Sung, Seulgi Oh ····· 117

P02 Prospective English teachers' cue-weightings in L2 stop perception

Hyunjung Lee, Eun Jong Kong ····· 120

P03 A Preliminary Study for the Development of Self-Assessment Tool for Cognitive Function in Daily Life

Sanghee Park, Okbun Lee ····· 121

P04 A longitudinal study on the changes in Korean mother's child-directed speech to young children with cochlear implants

Nuri Lee, Heesun Park, Youngmee Lee ····· 123

P05 Vocal Development in Congenitally Deaf Children Receiving Cochlear Implants in the First Year of Life

Nakyoung Lee, Yesol Jeon, Youngmee Lee ····· 125

P06 멀티모달 접근법을 통한 L2 영어 말하기 평가 모델 개발

오세현, 김선희, 정민화 ····· 127

P07 AI-based Automatic Detection of Repetitions and Prolongations in Stuttered Speech: Using a Convolutional Neural Networks (CNN)

Jin Park, Chang-Gyun Lee ····· 128

P08 A Declipping Method Using BSRNN and HiFi-GAN

Seung-Un Choi, Seung-Hyun Moon, Seung-Ho Choi ····· 130

P09 Audio Quality Enhancement of HiFi-GAN Generated Speech Using Post-Processing Filters

Sehyun Sohn, Minseo Jeon, Gyuchan Hwang, Seungho Choi ····· 131

P10 잡음 모델링을 이용한 잡음 인지 음성 향상 기법

김장연, 신의협, 김경화, 박형민 ····· 132

2024 한국음성학회 가을 학술대회

특 강

▮ 음성언어에 나타나는 운율구조의 언어보편성

조태홍(한양대)

▮ 장애음성의 기록과 분석:기기에서 프로그램까지

이승진(한림대)

▮ 음성 딥페이크 탐지 현황

유하진(서울시립대)

특강 1

▉음성언어에 나타나는 운율구조의 언어보편성

조태홍(한양대)

2024년 한국음성학회 가을 학술대회 (2024.10.19)

Phonetic Universals of Prosodic Structure and Phonetic Grammar: With a Focus on the Glottalization of Word-Initial Vowels

운율구조의 언어보편성과 음성문법: 어두 모음 성문화 현상을 중심으로

2024.10.19

Taehong Cho
Hanyang University
tcho@hanyang.ac.kr

한양 음성 • 언어인지과학 연구소
HIPCS
HANYANG INSTITUTE FOR PHONETICS & COGNITIVE SCIENCES OF LANGUAGE

음성 DB • 인지과학 융합연구센터
Convergence Research Center for Articulatory DB and Cognitive Sciences

1

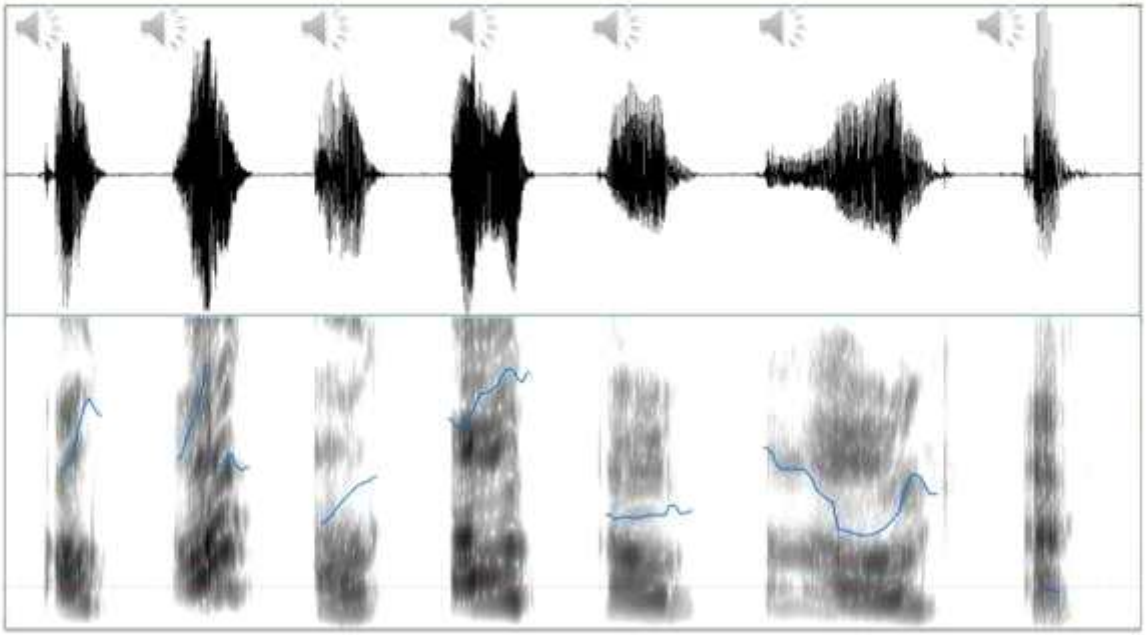
Overview

- Speech variation as an output of prosodic structure – assumed to be an integral component of speech production (and perception)
- Some preliminary results of studies on glottalization of word-initial vowels in English and Korean in relation to syntax-prosody interface
- Some theoretical implications for phonetic encoding of prosodic structure and other higher-order linguistic structures

3

speech variation 1

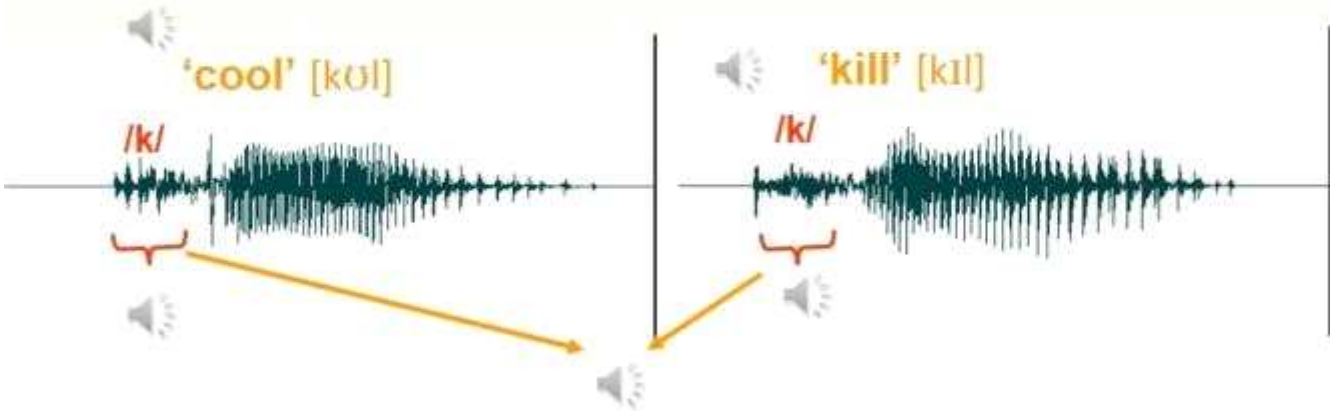
variation of 'Okay'
driven by paralinguistic reasons



4

speech variation 2

variation of /k/
driven by phonetic context (coarticulation)



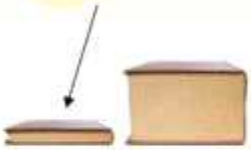
4

speech variation 3

variation driven by phonology (assimilation)

thin sandwich

thin book



thin carpet

'Did you know that eating a thin sandwich is a lot of fun?'

[n]

'Did you know that reading a thin book is a lot of fun?'

[m]

'Did you know that cleaning a thin carpet is a lot of fun?'

[ŋ]

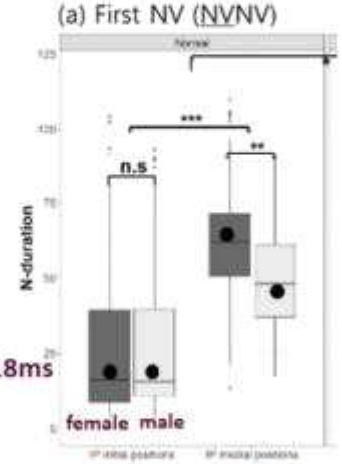
speech variation 4

driven by socio-indexical (e.g., gender-related) factors

Seoul Korean 'denasalizes' IP-initial nasal consonant

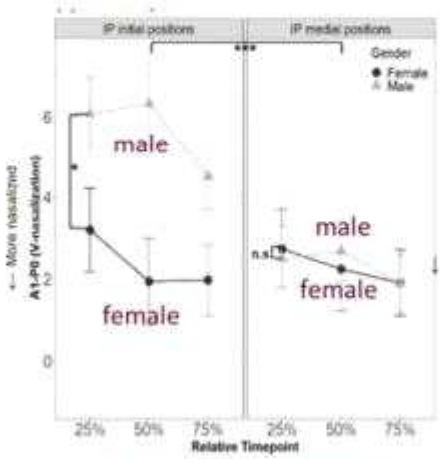
but females still maintain the nasality (nasal feature) through V-nasalization

(a) First NV (NVNV)



Mean = ca. 18ms

IP-initial IP-medial



IP-initial IP-medial

(Jang, Lee, Kim, Cho, in progress)

speech variation 5

Who (how many people) did I meet?

I met Mary and Elena's mother at the mall

(Mary and Elena)'s mother

(Mary) and (Elena's mother)

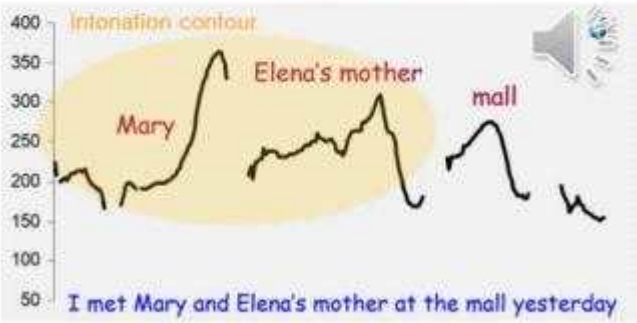
7

speech variation 5

variation of *intonation (prosody)* driven by syntax

(Mary and Elena)'s mother

(Mary) and (Elena's mother)



8

speech variation 5

variation of *intonation* driven by syntax

does not merely refer to a change in intonation (tune)

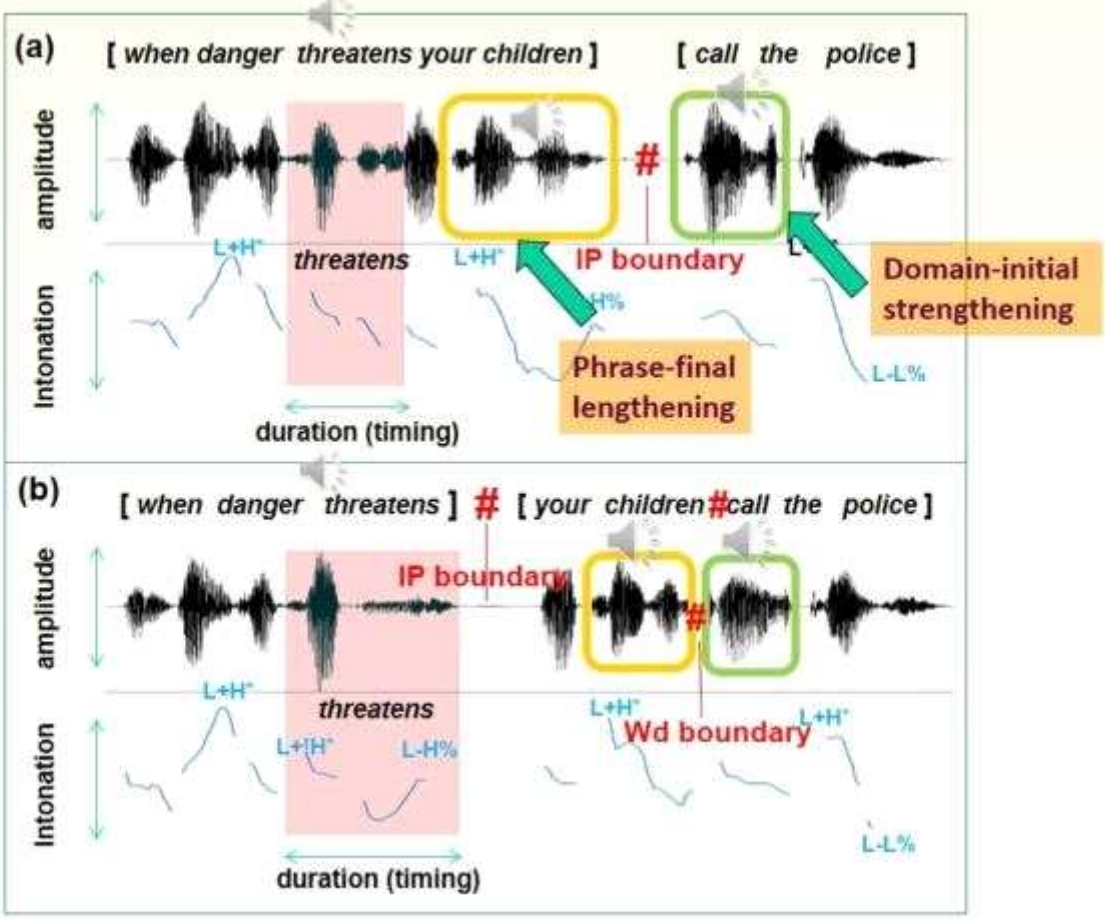
9

Example of variation driven by syntax

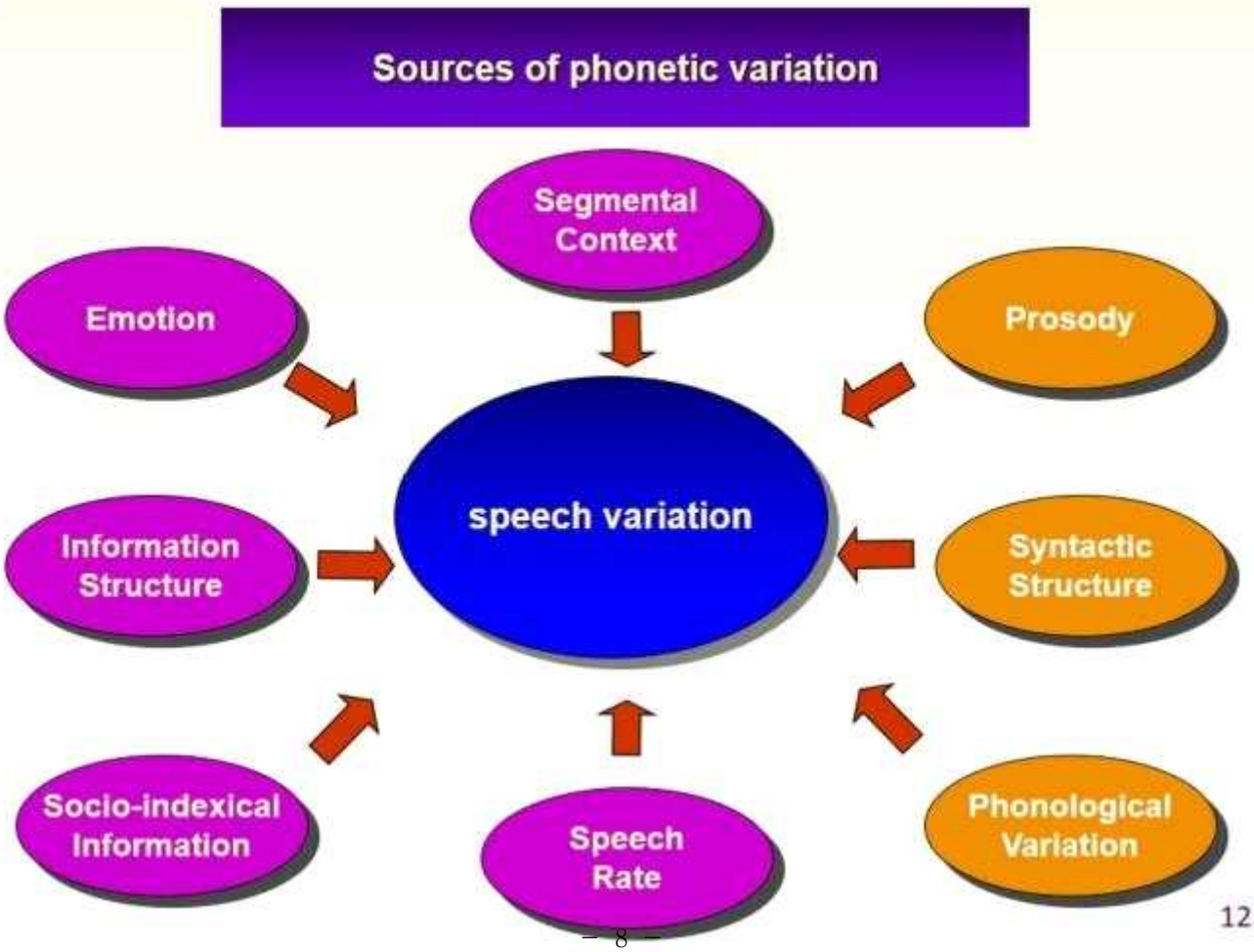
"when danger threatens, your children call the police"

"when danger threatens your children, call the police"

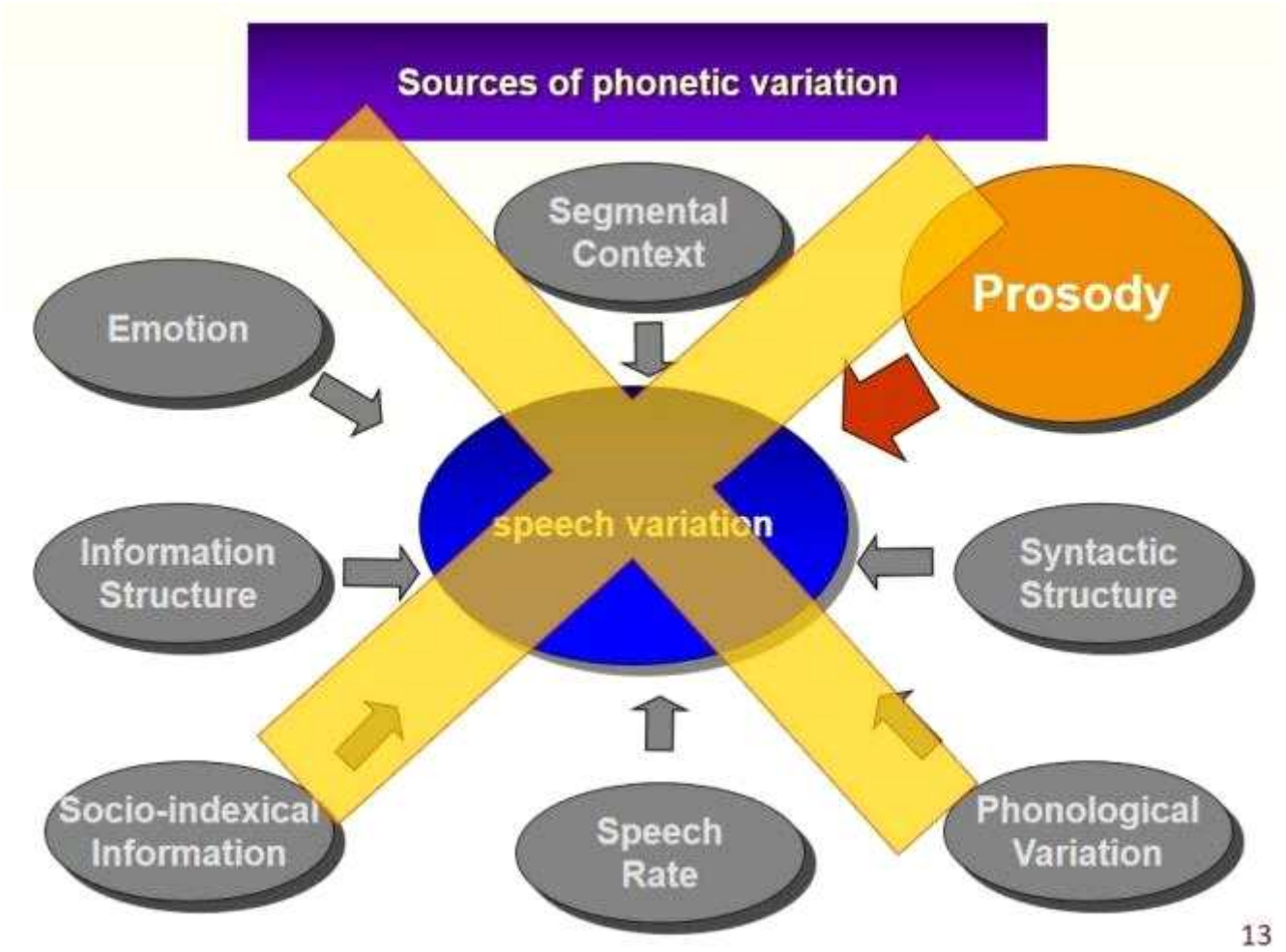
10

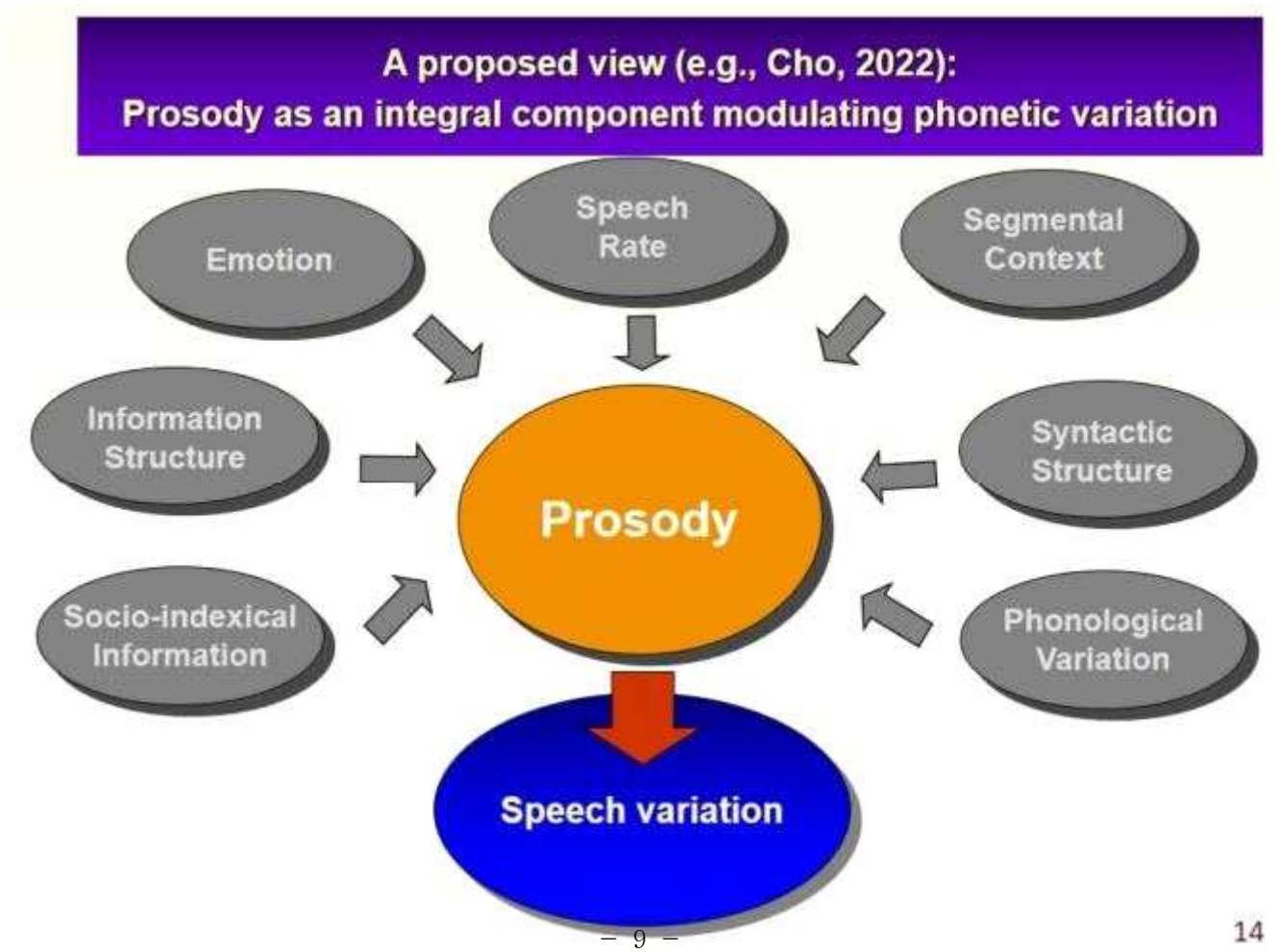

(a)
[when danger threatens your children]
[call the police]
amplitude
Intonation
#
L+H*
threatens
L+H*
IP boundary
Domain-initial strengthening
Phrase-final lengthening
L-L%
duration (timing)
(b)
[when danger threatens] # [your children #call the police]
amplitude
Intonation
IP boundary
L+H*
threatens
L+H*
L-H%
Wd boundary
L+H*
L-L%
duration (timing)

11


Sources of phonetic variation
Segmental Context
Emotion
Prosody
speech variation
Information Structure
Syntactic Structure
Socio-indexical Information
Speech Rate
Phonological Variation

12


Sources of phonetic variation
Segmental Context
Prosody
Emotion
speech variation
Information Structure
Syntactic Structure
Socio-indexical Information
Speech Rate
Phonological Variation
13


A proposed view (e.g., Cho, 2022):
Prosody as an integral component modulating phonetic variation
Speech Rate
Emotion
Segmental Context
Information Structure
Syntactic Structure
Prosody
Socio-indexical Information
Phonological Variation
Speech variation
14

Prosody

Suprasegmental Features

1. pitch
2. duration
3. amplitude

15

Prosody = Prosodic Structure

an abstract organizational structure which determines:

(a) the grouping of phonological units

(b) their relative salience

(Beckman 1996; Shattuck-Hufnagel & Turk, 1996)

~ serves as **a frame for articulation** (*e.g.*, Beckman, 1996 and Keating and Shattuck-Hufnagel, 2002, Keating, 2006, Cho, 2016)

16

"Prosody is a grammatical entity in its own right"
(Beckman 1996)

prosodic structuring = an integral part of spoken language

17

Prosodic Grouping

Autonomy of prosody

'Daddy dislikes hamburgers'

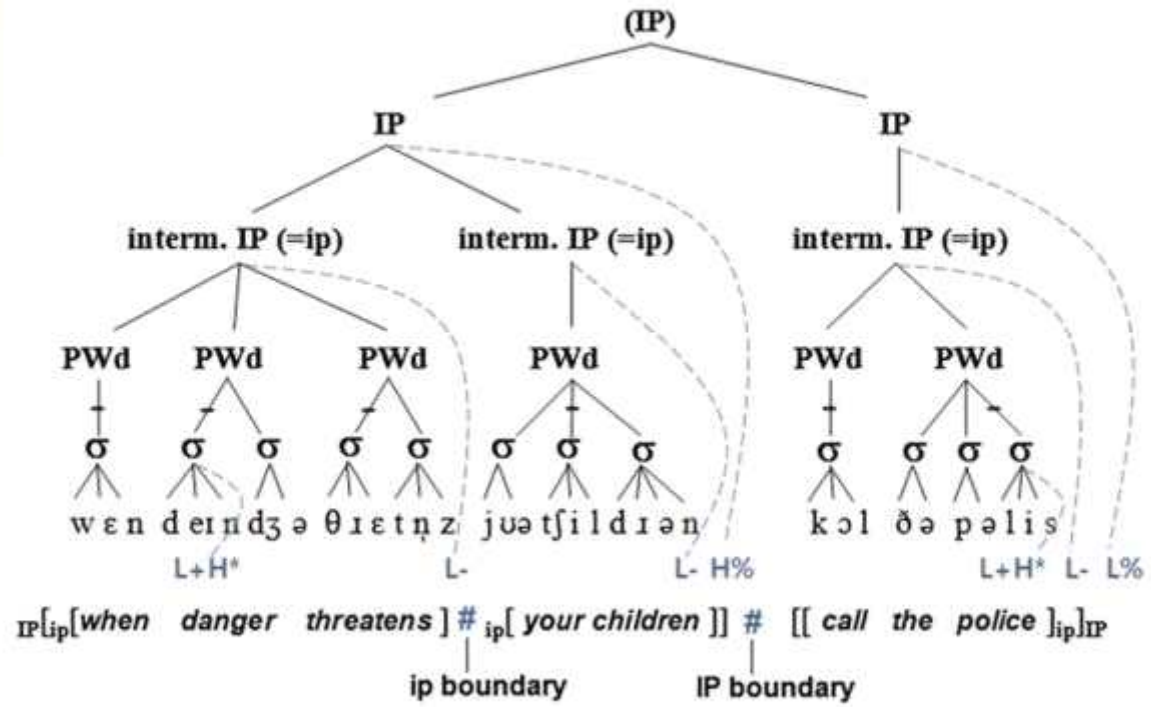
Daddy | dislikes hamburgers

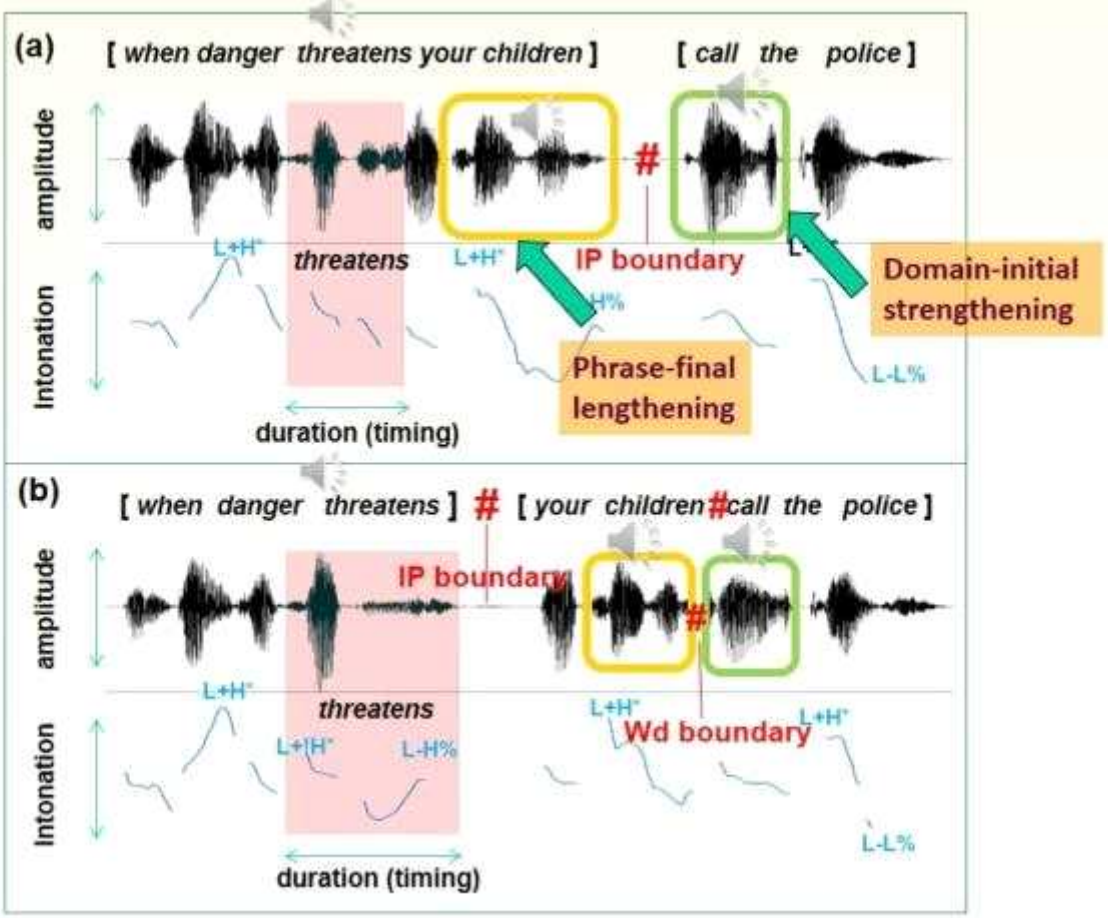
Daddy dislikes | hamburgers

Daddy | dislikes | hamburgers

18

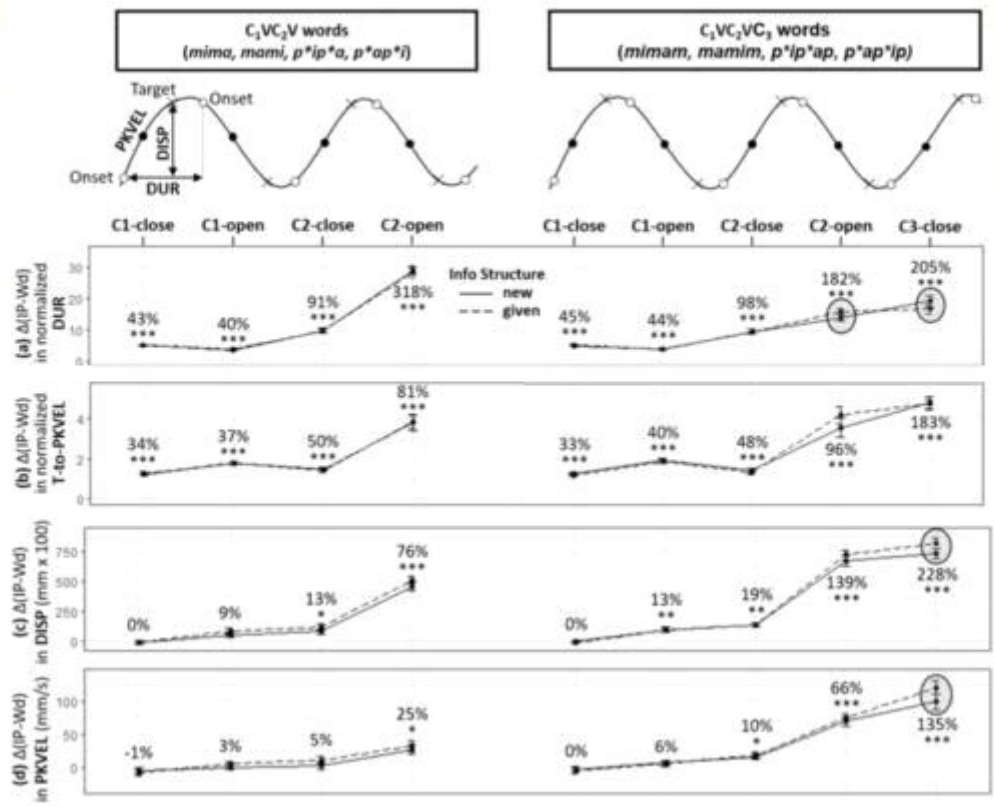
Prosody = Prosodic Structure





Right-edge strengthening in Korean as an edge-prominence language

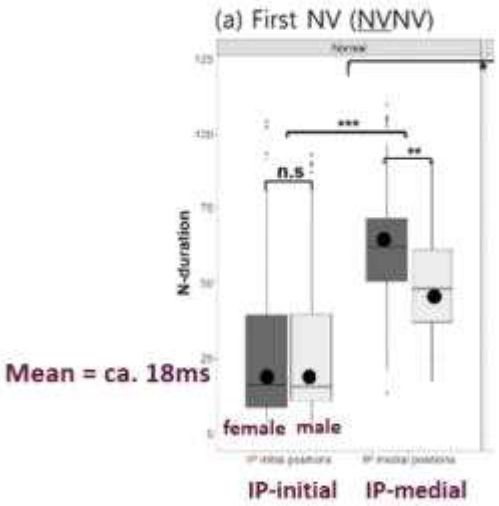
(Kim, Kim & Cho, submitted)



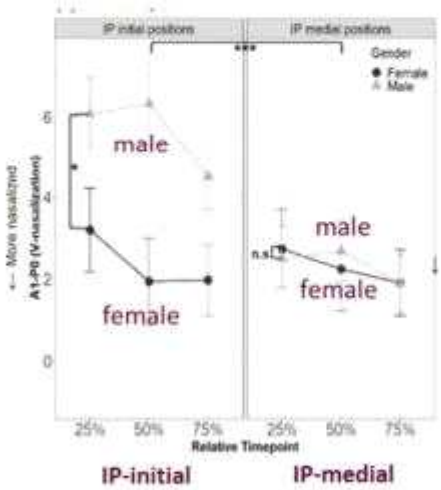
Domain-initial strengthening as a source of on-going sound change: denasalization

Seoul Korean 'denasalizes' IP-initial nasal consonant

but females still maintain the nasality (nasal feature) through V-nasalization



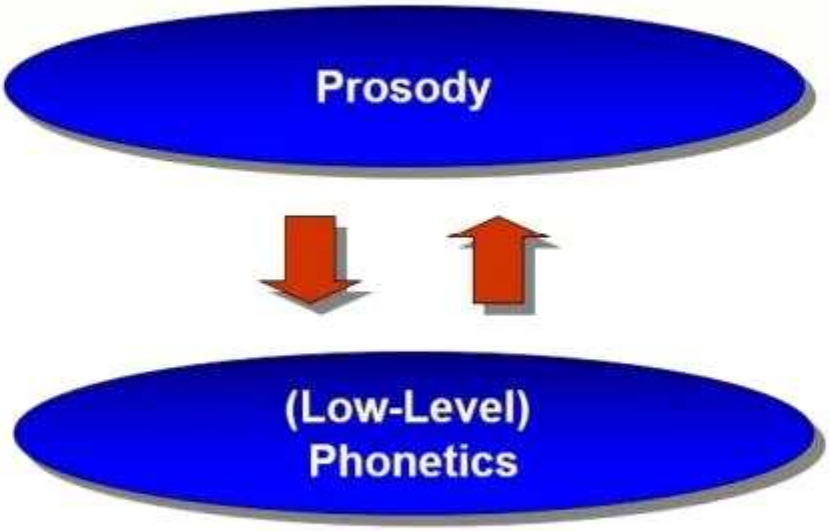
Mean = ca. 18ms



(Lee, Kim, Cho, in progress)

Prosody-Phonetics Interface

Prosody serves as **a frame for articulation**



at both the suprasegmental & the segmental levels

23

Prosodic Encoding & Decoding

Prosodic Encoding in Speech Production
= Phonetic Encoding of Prosodic Structure
(e.g., Keating 2006; Cho, 2022)

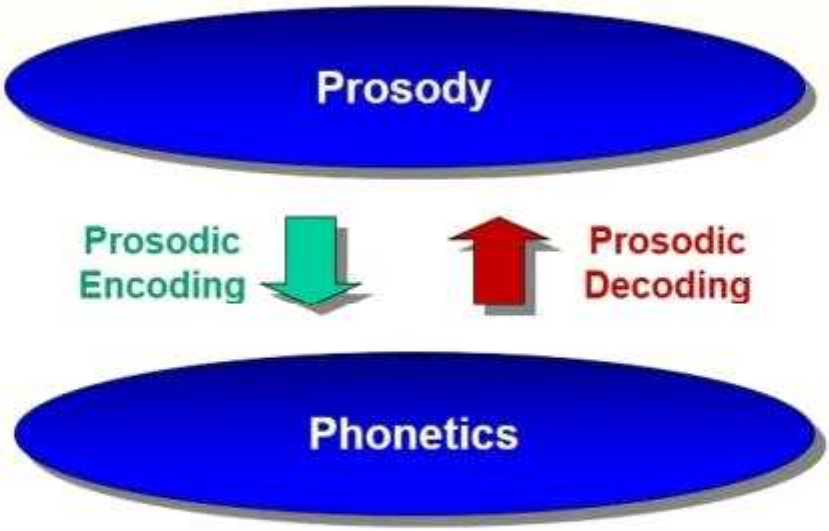
Speakers express (encode) prosodic structure via fine-grained phonetic details

Prosodic Decoding in Speech Comprehension
(e.g., Cho, McQueen & Cox, 2007; McQueen & Dilley, 2020)

Listeners exploit prosodically-driven phonetic details
- **to parse (decode) prosodic structure,**
  - **which helps understand the speakers' massage**

24

Prosody-Phonetics Interface

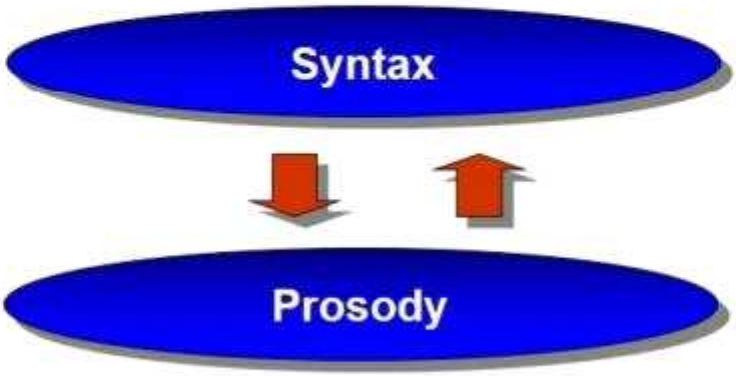


at both the suprasegmental & the segmental levels

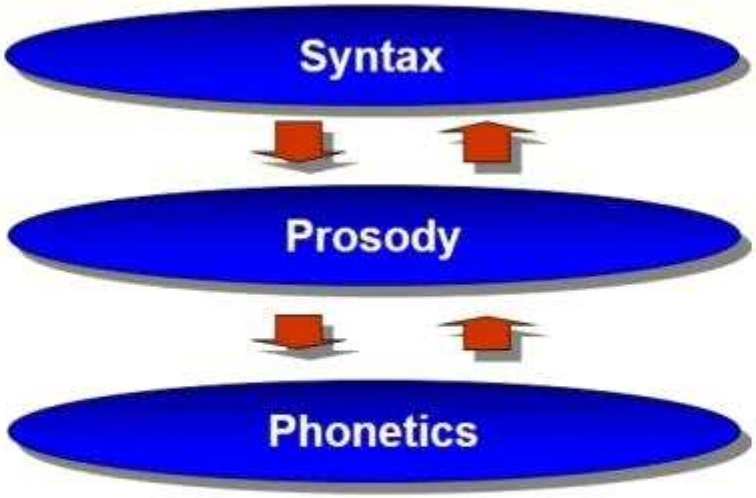
Syntax-Prosody Interface

[e.g., Jun, 2009; Elfner, 2018; Jun, S. A. 2009; Nespor & Vogel, 2012)

The notion of syntax-prosody interface is often related to <u>how syntactic junctures are aligned with prosodic junctures</u>



Syntax-Prosody-Phonetics Interface?

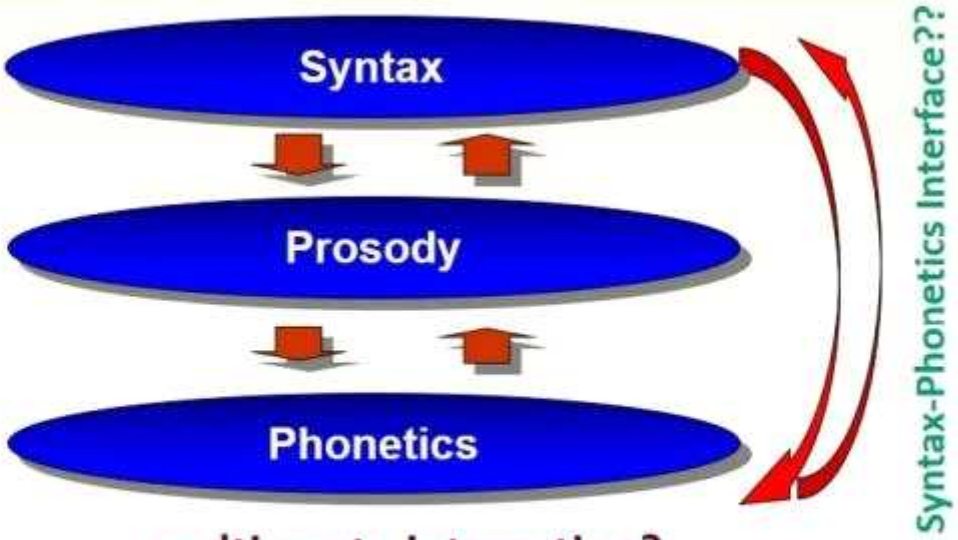


in a single-route cascaded fashion?

- Is the syntactic influence on speech production mediated by prosody ?

27

Syntax-Prosody-Phonetics Interface?



multi-route interaction?

- Syntax-Phonetics interface:
  Does syntax interact directly with phonetics ?

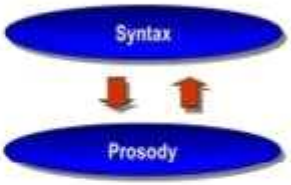
28

Todays' overarching questions

concern "phonetic encoding of prosodic structure" in relation with syntax in English and Korean

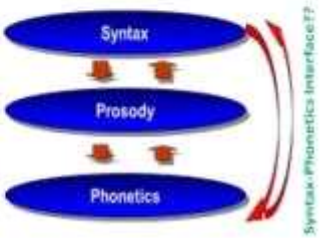
1. How are *syntactically ambiguous coordinate structures* resolved through the syntax-prosody interface (mapping)?

   ➡ by examining an alignment of underlying syntactic junctures with prosodic junctures (phrasings)



2. Is *prosodic-structurally conditioned phonetic realization* (=phonetic encoding of prosodic structure) further modulated in reference to syntactic structure? If so, how?

   ➡ by examining variation in:
   i) the preboundary (domain-final) lengthening
   ii) the postboundary (domain-initial) glottalization



29

Vowel glottalization

(Dilley et al., 1996, Redi & Shattuck-Hufgnagel, 2001; Garellek, 2014)

more V-glottalization

phrase(domain)-initial

....] # [ʔV...

If you take her to a doctor IP *Annie Williams will be crying*

cf: *Doctor Annie Williams*

V-glottalization = "segmental" marker for a prosodic juncture

30

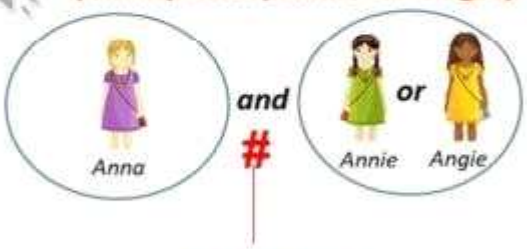
Structural ambiguity with coordinate structures

Structural ambiguity in English

N1 and N2 or N3

Early Closure

(Anna) and (Annie or Angie)



IP boundary

Late Closure

(Anna and Annie) or (Angie)

IP boundary

31

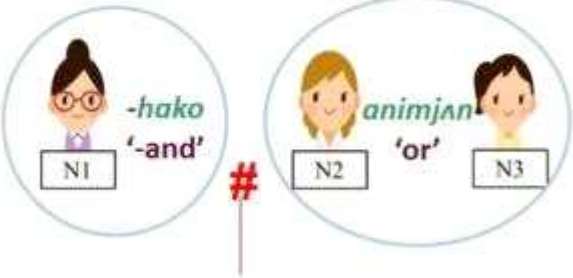
Structural ambiguity with coordinate structures

Structural ambiguity in Korea

N1-and N2 or N3

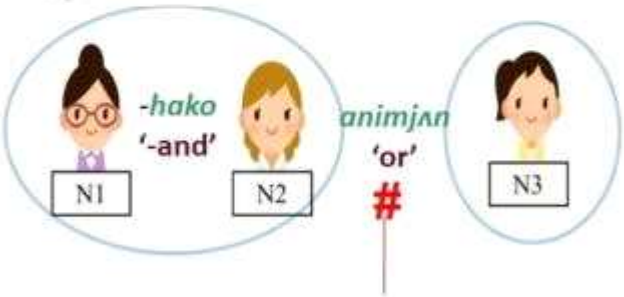
N1-hako N2 animjʌn N3

Early Closure



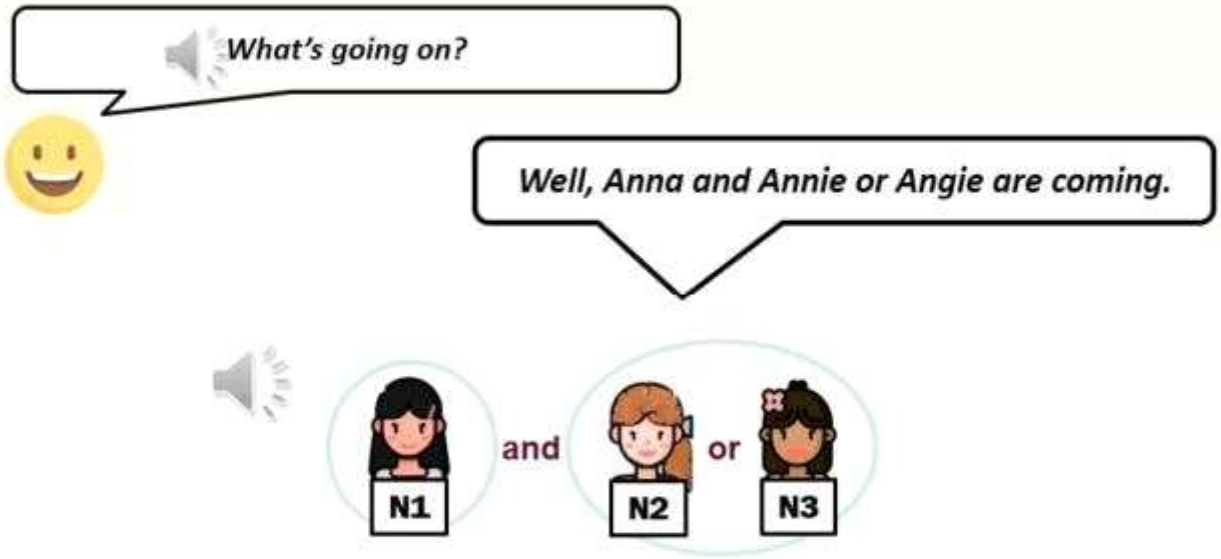
IP boundary

Late Closure



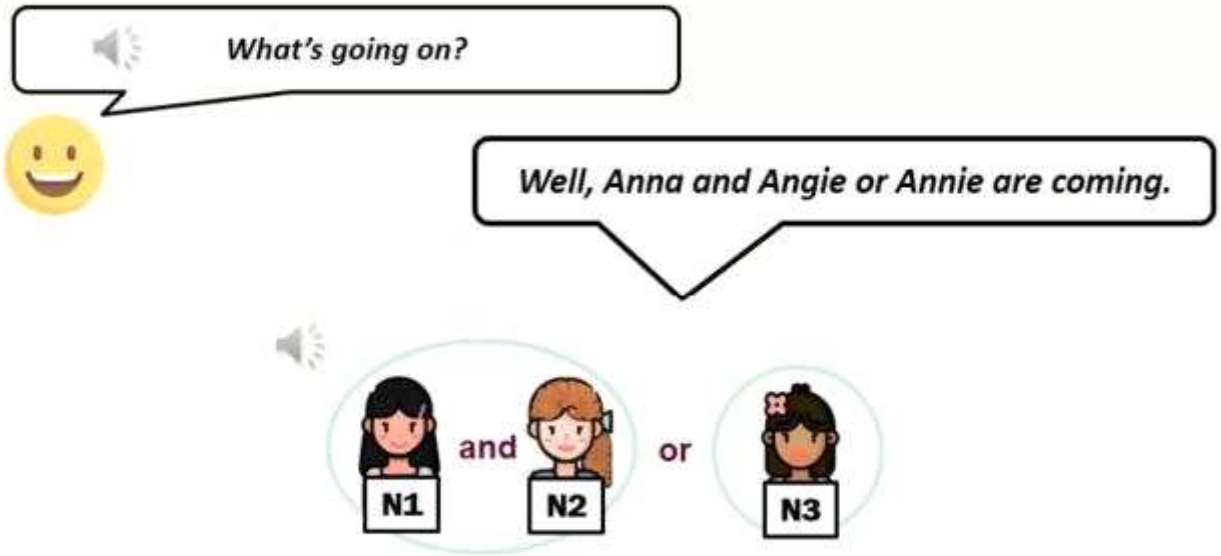
IP boundary

English Example



Early Closure

English Example

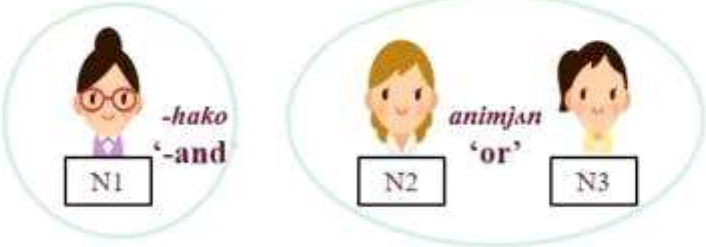


Late Closure

Korean Example

무슨 일이야?
Musɨn ilija ('What's happening?')

아, N1하고, N2 아니면 N3가 온대
a, N1-hako N2 animjʌn N3–ka ontɛ
'Well, N1 and N2 or N3 are coming.'

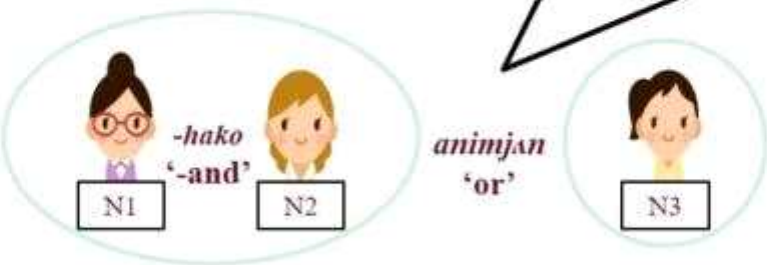


Early Closure

Korean Example

무슨 일이야?
Musɨn ilija ('What's happening?')

아, N1하고 N2, 아니면 N3가 온대
a, N1-hako N2 animjʌn N3–ka ontɛ
'Well, N1 and N2 or N3 are coming.'



Late Closure

English Examples

	Question	Answer
Broad Focus	*What is going on?*	*Well, (Anna) and (Annie or Angie) are coming.*
		Well, (Anna and Annie) or (Angie) are coming.
Narrow Focus	*WHO will come to the party?*	*Well, **(Anna) and (Annie or Angie)** will.*
		*Well, **(Anna and Annie) or (Angie)** will.*
Contrastive Focus	*Did they say (Anna and Annie) or (Angie) will come?*	*No. They said, **(Anna) and (Annie or Angie)** will.*
	Did they say (Anna) and (Annie or Angie) will come?	*No. They said, **(Anna and Annie) or (Angie)** will.*

37

Korean Examples

	Question	Answer
Broad Focus	무슨 일이야? ***musin ilija*** 'What's happening?'	아, N1하고, N2 아니면 N3가 온대 *a, (N1-hako) (N2 animjʌn N3)-ka ontɛ* 'Well, (N1-and) (N2 or N3) are coming.'
		아, N1하고 N2, 아니면 N3가 온대 *a, (N1-hako N2) animjʌn (N3)-ka ontɛ* 'Well, (N1-and N2) or (N3) are coming.'
Narrow Focus	이번 축제에 누가 온대? ***ipʌn ʧʰuktʃɛ-ɛ nuka ontɛ?*** 'Who's coming to the party?'	아, N1하고, N2 아니면 N3가 온대 *a, (N1-hako) (N2 animjʌn N3)-ka ontɛ* 'Well, (N1-and) (N2 or N3) are coming.'
		아, N1하고 N2, 아니면 N3가 온대 *a, (N1-hako N2) animjʌn (N3)-ka ontɛ* 'Well, (N1-and N2) or (N3) are coming.'
Contrastive Focus	뭐? N1하고 N2, 아니면 N2가 축제에 온대? ***mwʌ? (N1-hako N2) animjʌn (N3) -ka ʧʰuktʃɛ-ɛ ontɛ?*** 'Did they say (N1-and N2) or (N3) are coming to the party'	아니, N1하고, N2 아니면 N3가 온대 *ani, (N1-hako) (N2 animjʌn N3)-ka ontɛ* 'No, they said (N1-and) (N2 or N3) will.'
	뭐? N1하고, N2 아니면 N2가 축제에 온대? ***mwʌ? (N1-hako) (N2 animjʌn N3) -ka ʧʰuktʃɛ-ɛ ontɛ*** 'Did they say (N1-and) (N2 or N3) are coming to the party'	아니, N1하고 N2, 아니면 N3가 온대 *ani, (N1-hako N2) animjʌn (N3)-ka ontɛ* 'No, they said (N1-and N2) or (N3) will.'

38

Some experimental aspects

<participants>

- 20 (10 F) American English speakers (ages = 19~35), 14 analyzed
- 20 (10 F) Seoul Korean speakers (ages = 20~29), 14 analyzed

<measurements>

<Partial ToBI transcriptions agreed upon by all collaborators>

- Boundary Type (IP vs. Wd)
- Presence/absence of Pitch Accent

39

Language-specificity?

Yes, languages do differ in various ways!

(for example)

- Both macro- and micro-prosody (e.g., Jun, 2014; cf. Shaw 2022)
- Articulatory underpinnings of preboundary lengthening (e.g., Cho, 2016)
- The way that the acoustic space is used for varying voice quality in a given language (e.g., Keating et al., in press)

40

Language-specificity of English vs. Korean?

English as a head-prominence language

- V-glottalization may be driven by prominence rather than by boundary in English (e.g., Garellek, 2014, 2022)
- Domain-initial strengthening effects may be neutralized in the stressed syllable (e.g., Barnes, 2001)

Korean as a head-prominence language (e.g., Jun, 2014)

- Domain-initial strengthening (left-edge effects) tends to be more robust in Korean than in other languages (e.g., Keating et al., 2003; Cho, 2022)

➡ *More robust boundary-related glottalization in Korean?*

41

But....

what about the constraint:

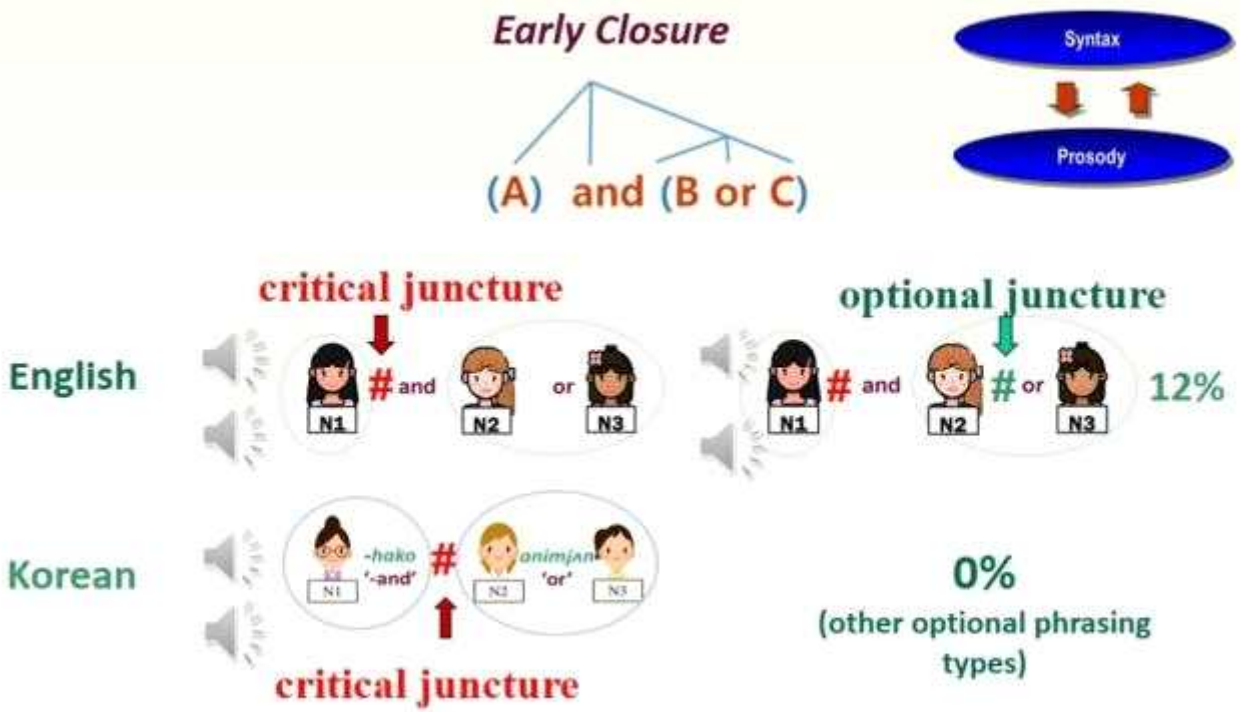
**Vowel Hiatus?*

Languages may also vary, depending on how this markedness constraint is weighed in the system!

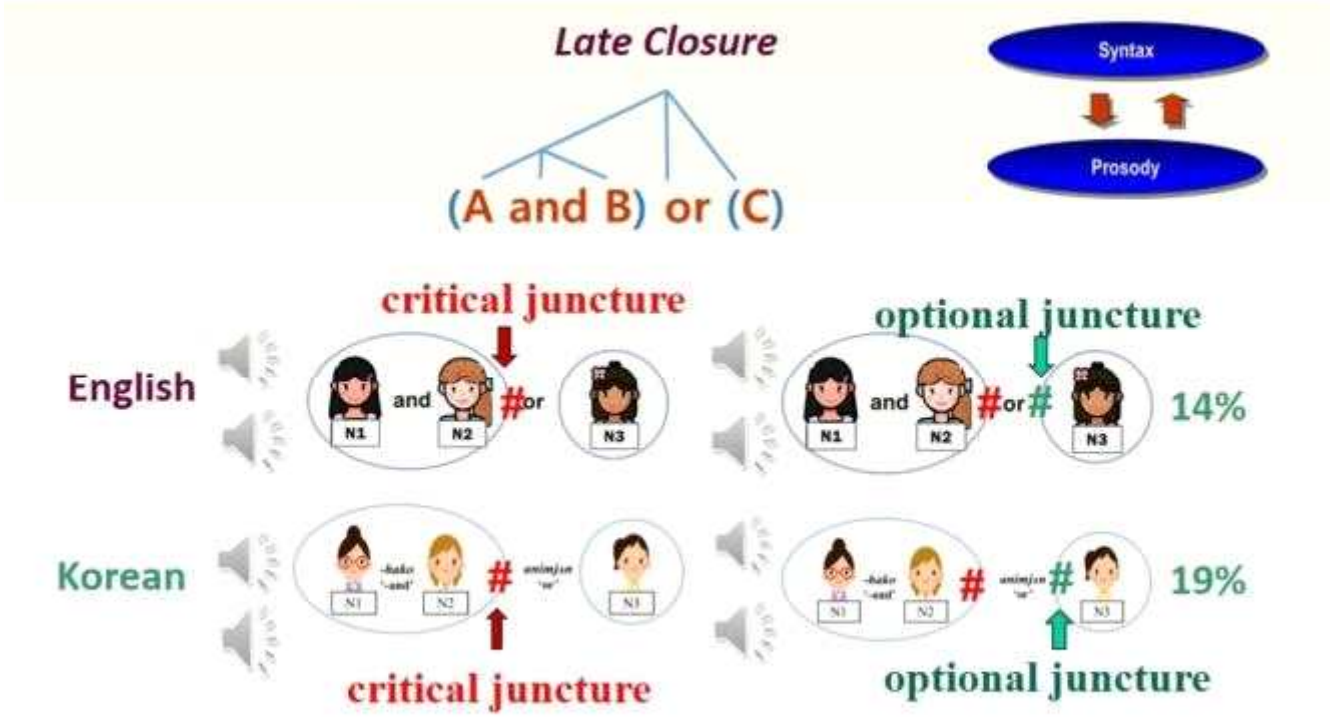
42

Results

Syntax-Prosody mapping

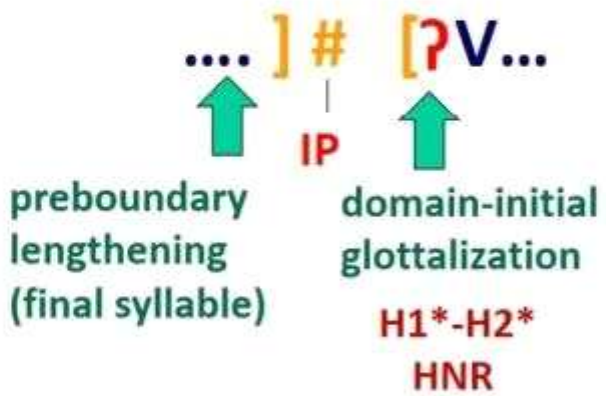


Syntax-Prosody mapping



45

Phonetic encoding of prosodic structure

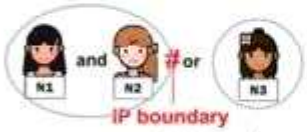


English

Early Closure

more glottalized
(only when pitch accented)

Late Closure



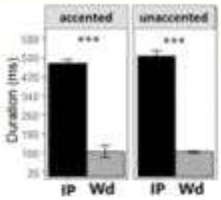
more glottalized
(only when pitch accented)

optional juncture

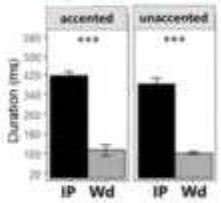
14%

no boundary effect
on glottalization

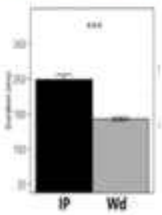
Preboundary
Lengthening of N1 (e.g., Annie)



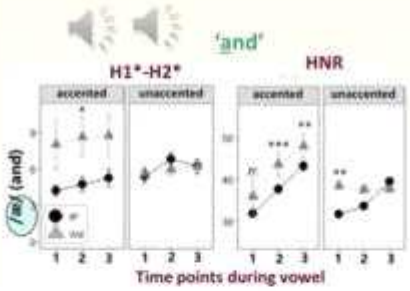
Preboundary
Lengthening of N2 (e.g., Annie)



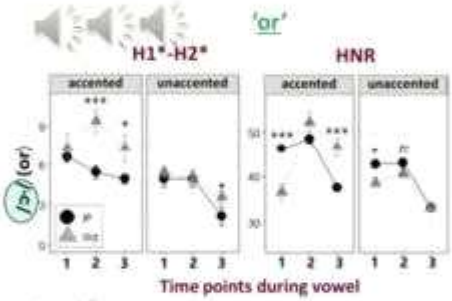
Preboundary
Lengthening of 'or'



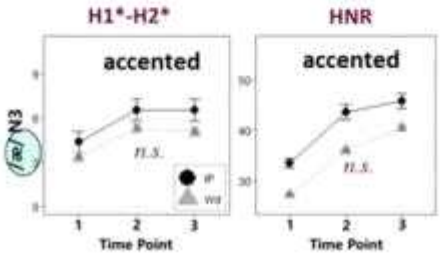
'and'



'or'



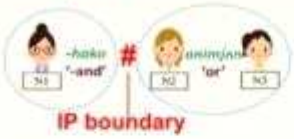
'Annie' (N3)



47

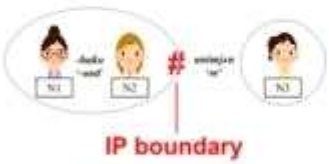
Korean

Early Closure

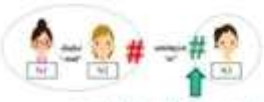


not (less) glottalized

Late Closure



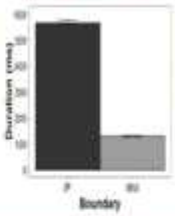
more glottalized



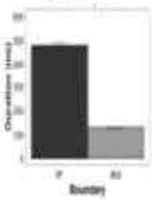
19%

optional juncture

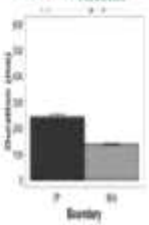
preboundary lengthening
of '-hako'



preboundary lengthening
of N2



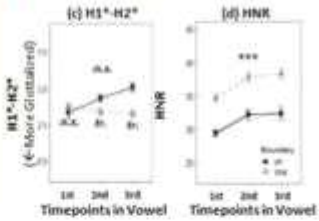
preboundary lengthening
of *animjʌn*



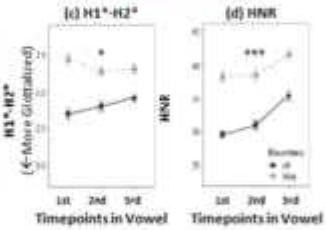
N2 = /ali/, /atʃi/, /ami/

#

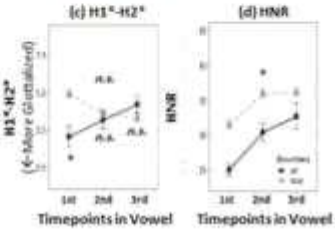
Initial Glottalization for N2



animjʌn 'or'



'ami' (N3)



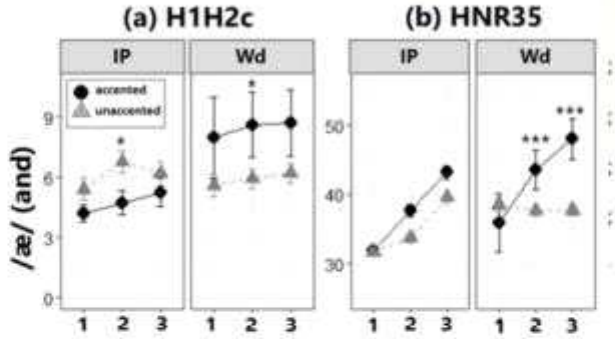
48

Effects of Prominence (pitch accent) in English

'and' (N1 and N2 or N3)

IPi 45% (acc.)
IPi 55% (una.)

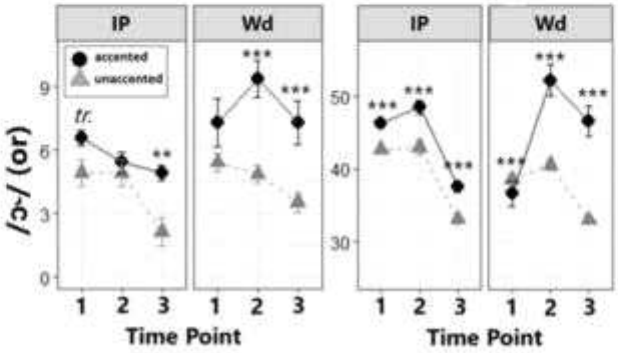
Wd 6% (acc)
Wd 94% (una)



'or' (N1 and N2 or N3)

IPi 65% (acc.)
IPi 35% (una.)

Wd 8% (acc)
Wd 92% (una)

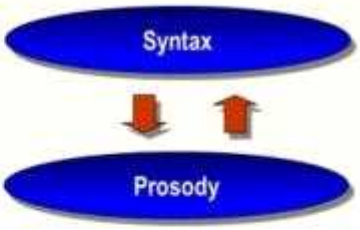


Discussion

Some cross-linguistic similarities

Cross-linguistic similarity 1:

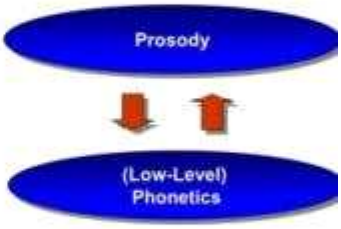
In both English and Korean, *syntactically ambiguous coordinate structures are* resolved through the alignment between major syntactic and prosodic junctures



Cross-linguistic similarity 2:

In both languages, the prosodic structure is encoded consistently by the temporal variation (i.e., preboundary lengthening).

→ similar use in phonetic encoding of prosodic structure



Cross-linguistic similarity 3:

The voice quality-related feature is used in some ways in reference to prosodic structure.

- **But given the robust temporal reflexes of prosodic structure, the glottalization is not consistently used in marking prosodic structure.**

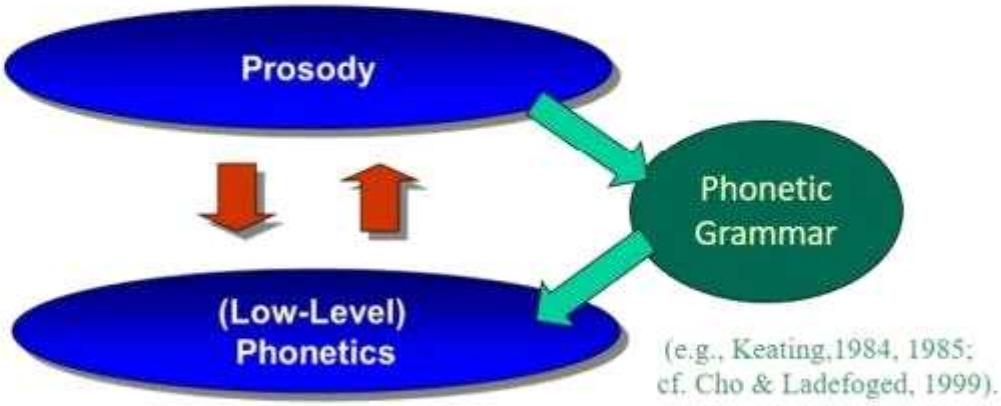
51

Some further points

Still we found some evidence for cross-linguistic use of voice quality difference as a prosodic marker

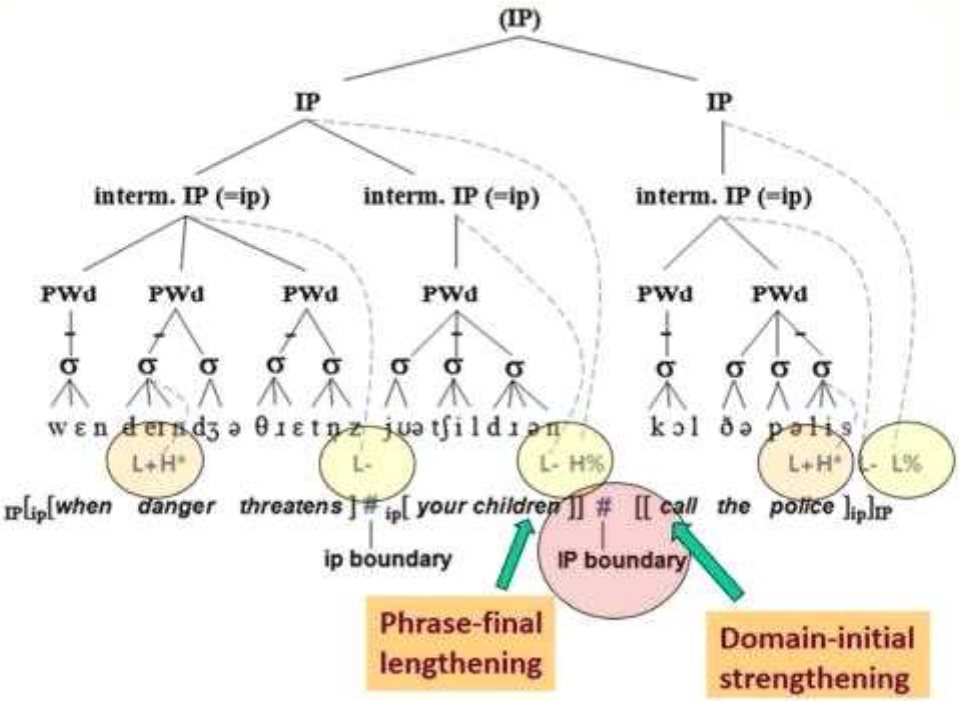
- **But in contrast with our initial prediction, Korean (as an edge-prominence language) does not show more robust glottalization as a domain-initial strengthening effect, as compared to English.**
  - **Vowels were in some cases less glottalized IP-initially than IP-medially in Korean**
- **English appears to use glottalization more robustly across the board regardless of boundary conditions, presumably due to the system's preference to avoid vowel hiatus (at least more than Korean does).**
  - **There was evidence that the glottalization is conditioned by prominence in some way in English, though not consistently so.**

Results also showed language-specific fine-tuning of phonetic realization in relation to prosodic structure.



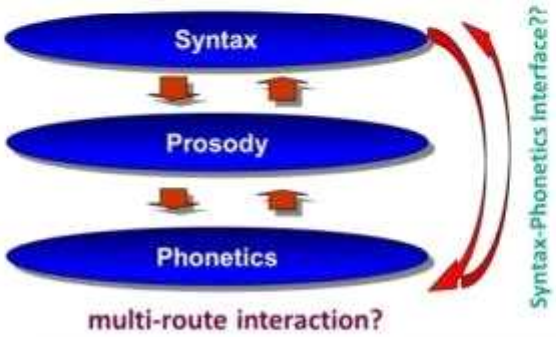
53

'Phonologically-defined' Prosodic Structure



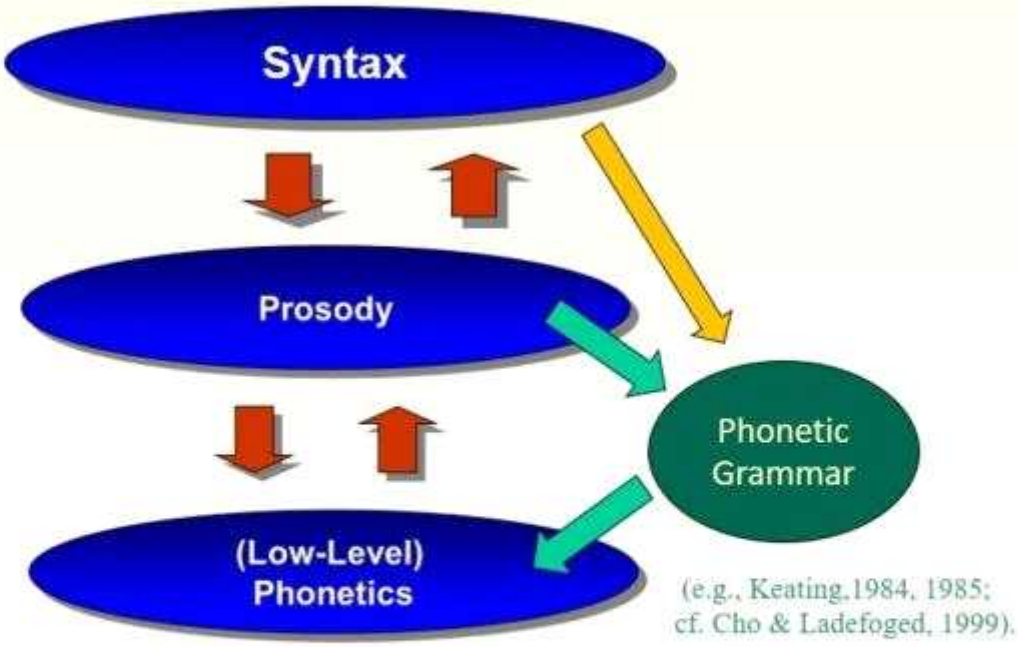
An important point to be underscored:

- The use of glottalization and preboundary lengthening differed for the same phonologically-defined IP boundary in both English and Korean.
  - It varied depending on whether or not an IP was used critically for marking syntactic junctures.

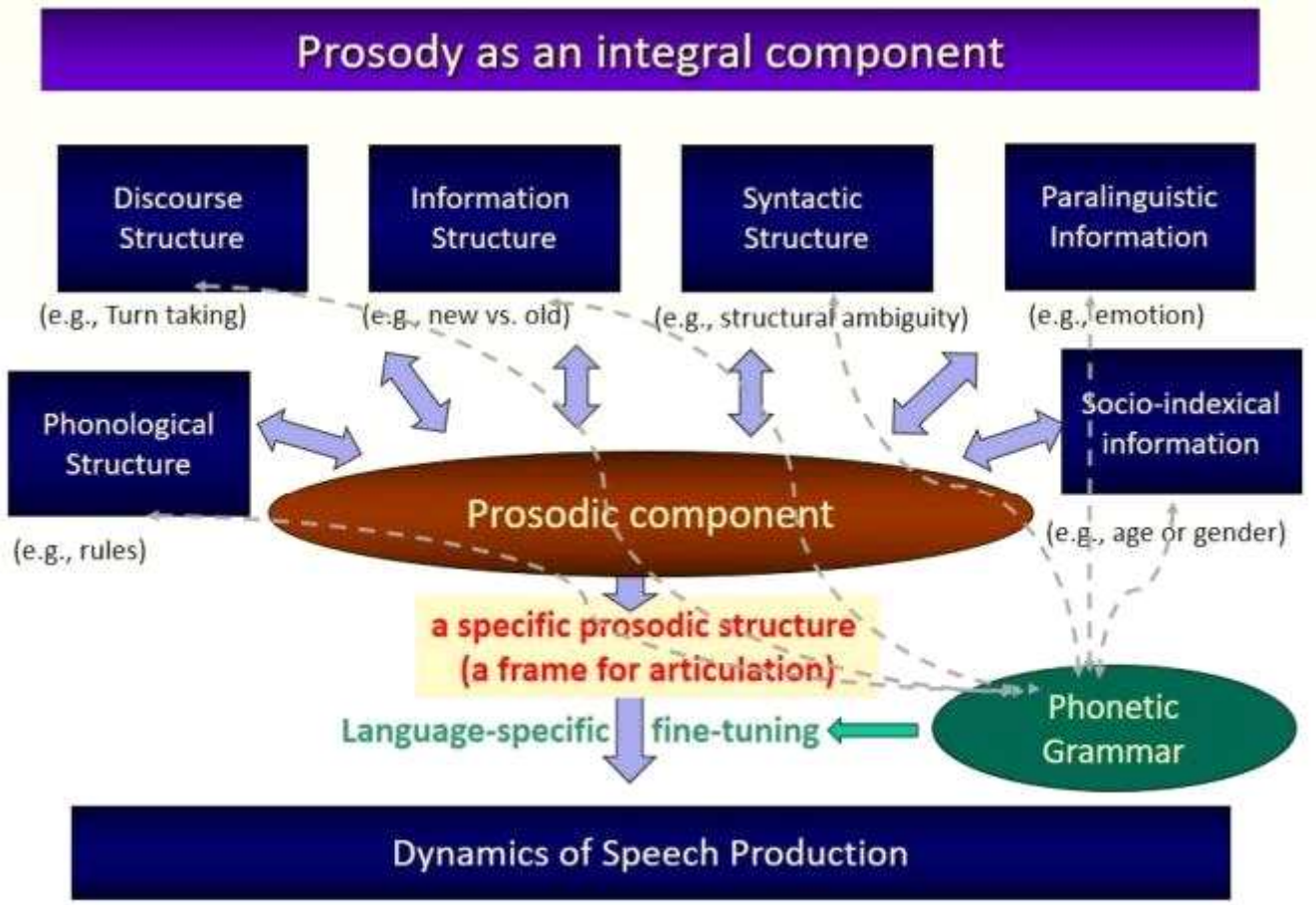


- Syntax-Phonetics interface:
  Does syntax interact directly with phonetics ?

55



E.g. boundary strength of a phonologically-defined IP can be further modulated by *Phonetic Grammar* in reference to Syntax



57

감사합니다!

Thank you!

일정표로 이동

58

특강 2

▌장애음성의 기록과 분석:기기에서 프로그램까지

이승진(한림대)

한국음성학회
Korean Society of Speech Sciences

HALLYM UNIVERSITY 40 1982 - 2022

장애음성의 기록과 분석
기기에서 프로그램까지

HALLYM LAB FOR VOICE & COMMUNICATION DISORDERS
한림 음성 의사소통장애 연구실

Seung Jin Lee, PhD, SLP
Division of Speech Pathology & Audiology
Research Institute of Audiology & Speech Pathology
Hallym University

음성장애

- 성대, 호흡기관, 성도의 구조적, 기능적 이상으로 인하여 음도, 음량, 음질의 변화를 초래하는 말 장애의 한 유형
- 음성의 음도, 음량, 음질, 유동성이 그 사람의 성별, 체격, 문화의 특성 등과 부조화

https://www.asha.org/practice-portal/clinical-topics/voice-disorders/

2

다면적 음성검사

후두 (육안) 검사
Laryngeal inspection

음향학적 분석
Acoustic analysis

전기성문파형 검사
Electroglottography

Psychometric assessment
심리측정적 검사

Aerodynamic assessment
공기역학적 검사

Auditory-perceptual evaluation
청지각적 평가

3

장애음성의 특징

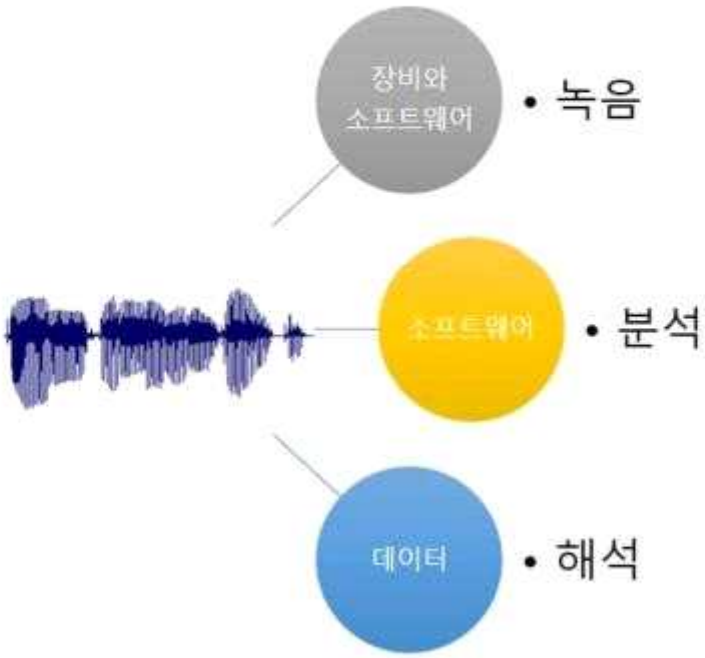
- 음도 및 강도가 정상 범위에서 벗어나는 경우
  - 분석 프로그램의 설정에 따라 분석의 신뢰도가 크게 저하될 수 있음
  - 해석에 주의가 필요
- 중증도에 따라 주기성이 크게 떨어지기도 함
  - 분석 자체가 불가능한 경우 많음(up to 30%)
  - 어떻게 해서든 결과를 낼 것인가?
- 과제에 따라 증상이 드러나거나 드러나지 않을 수 있음
  - 음성 분석의 목적을 정확하게 설정할 필요
  - 문제가 있는 경우 문제가 드러나도록 세심한 지시사항을 준비

4

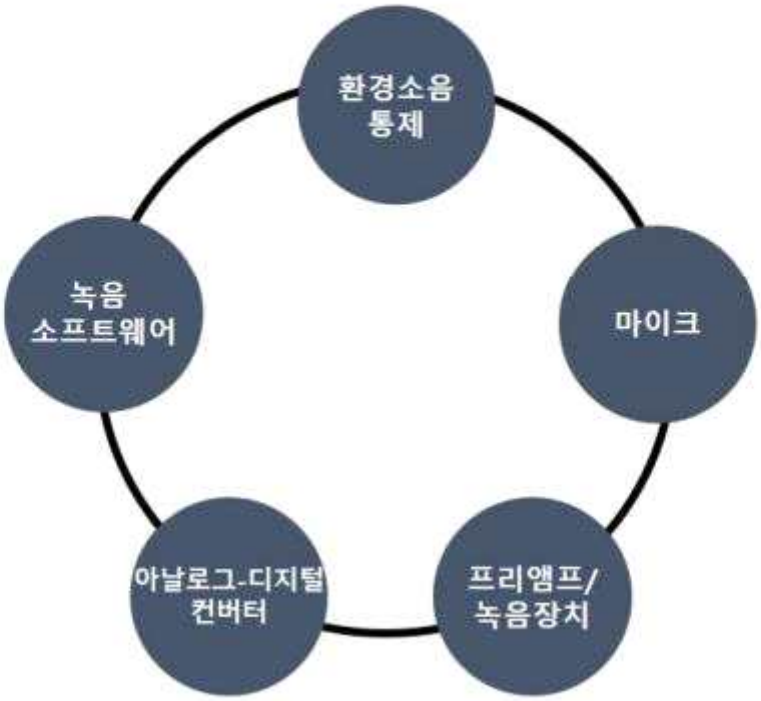
기본요소와 다양한 권고사항들

한림대학교
HALLYM UNIVERSITY

음성 분석의 필수요소?



음성 녹음의 기본요소



NCVS technical memo, 2012
https://archive.ncvs.org/e-learning/NCVSOnlineTechnicalMemo15-vs1_1.pdf

디지털 녹음 세팅

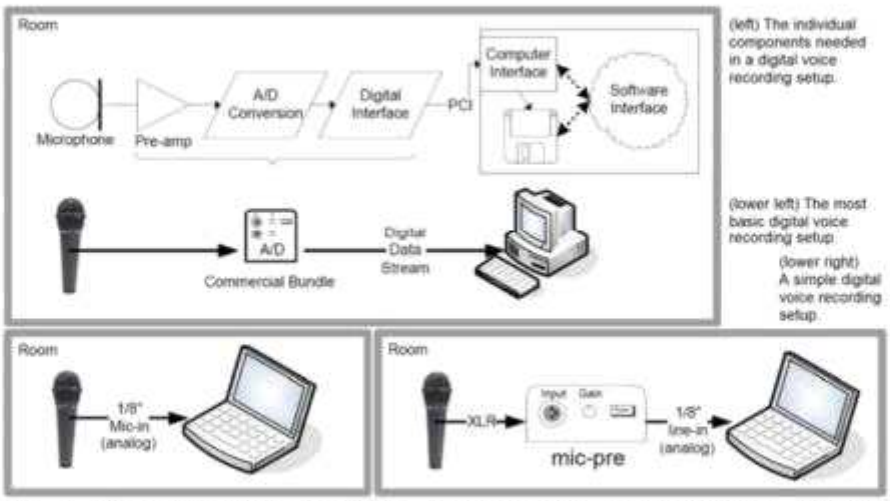


Figure 1. A digital voice recording setup requires a microphone, microphone preamplifier, analog-to-digital converter, digital interface, software interface, and recording environment. Many of these components are often bundled together.

그래서 뭘 사라고요?

NCVS technical memo, 2006
https://archive.ncvs.org/ncvs/library/tech/NCVSOnlineTechnicalMemo07.pdf

Equipment List and Supplies

Below is a general list of equipment needed to set up a basic recording laboratory. After each item, two estimated prices are given, the ideal version followed by a more economical version, if available (prices are as of March 2005).

NECESSARY EQUIPMENT

ITEM	PRICE	ECONOMICAL VERSION
Sound isolation (basic)	$4,500	N/A
Microphone	$600	$150 (inexpensive, separate mount)
Sound level meter	$4,500	$1,700
Microphone power/amp/converter	$550	$235
Computer	$1,000	$500
Recording software	$300	$0-$50
Backup hard disk recorder	$799	$439
Cables	$300	N/A
Surge protectors	$60	$60
Speakers for playback	$150	$0 (use ones that come with computer)
Camcorder (mini DV)	$1,000	$450-$800
Video monitor	$50	N/A
Microphone for video camera	$25	$25
Tripod for video camera	$35	$35
Stand light	$35	$35
Chair for subject	$150	$0 (use office chair with no wheels)
Music stand to display stimuli	$60	$25
Equipment Rack	$1200	$0 (use an old desk)
Storage rack and/or binders	$150	$0 (boxes on shelves)
Water pitcher and glasses	$50	$50
Tape measure (ruler)		
Totals	**$15,514**	**~$3,704-4,104**

REPLENISHABLE SUPPLIES

Item	Cost
Mini DV tapes	~$2.50/tape
DAT tapes, if necessary	~$4.50/tape
CD-Rs and jewel cases	~$0.67-0.99/ea
Batteries	~$0.41-1.13/ea (AA)

9

Sound isolation

- 방음부스
  - 견고한 몸체
  - 두꺼운 더블 벽체
  - 방진구조를 갖춘 뜬바닥구조
  - 외부장비와의 케이블 연결을 위한
  - 조용한 조명

- 흡음재를 적용한 작은 방
  - 부스 설치가 어려운 경우

NCVS technical memo, 2006
https://archive.ncvs.org/ncvs/library/tech/NCVSOnlineTechnicalMemo07.pdf

10

Reproducibility of Voice Parameters: The Effect of Room Acoustics and Microphones

*Pasquale Bottalico, †Juliana Codino, ‡,§Lady Catherine Cantor-Cutiva, †Katherine Marks, *Charles J. Nudelman, †Jean Skeffington, ‖Rahul Shrivastav, ‡,¶Maria Cristina Jackson-Menaldi, §Eric J. Hunter, and †Adam D. Rubin,
*Champaign, Illinois, †, §East Lansing, and ¶Detroit, Michigan, ‡Bogota, Colombia, and ‖ Atlanta, Georgia

FIGURE 1. Experimental set-up in RoomA.

TABLE 1.
Specifications of the six Microphones Analyzed in the Experiment

Mic Id	Microphone	Producer	Polar Pattern	Frequency Response (Hz)	Sensitivity at 1 kHz	Output Impedance (Ohm)	Cost (USD)
M22	M2211	NTI Audio	Omnidirectional	5–20000	20 mV/Pa	100	$1,210
ECM	ECM8000	Behringer	Omnidirectional	15–20000	11 mV/Pa	600	$60
AKG	Perception 120	AKG	Cardioid	20–20000	24 mV/Pa	200	$100
CSL	Computerized Speech Lab Mic	KayPENTAX	Cardioid	55–14000	1.3 mV/Pa	150	$100
SHU	Beta 53	Shure	Omnidirectional	20–20000	3.16 mV/Pa	136	$350
LOG	PC Headset 120	Logitech	Cardioid	100–10000	−39 dBV/Pa	-	$20

- Room A: 방음부스
- Room 1,2,3: 22, 25, 32 m^3
- 환경소음 49 dB(A)

11

Reproducibility of Voice Parameters: The Effect of Room Acoustics and Microphones

*Pasquale Bottalico, †Juliana Codino, ‡,§Lady Catherine Cantor-Cutiva, †Katherine Marks, *Charles J. Nudelman, †Jean Skeffington, ‖Rahul Shrivastav, ‡,¶Maria Cristina Jackson-Menaldi, §Eric J. Hunter, and †Adam D. Rubin,
*Champaign, Illinois, †, §East Lansing, and ¶Detroit, Michigan, ‡Bogota, Colombia, and ‖ Atlanta, Georgia

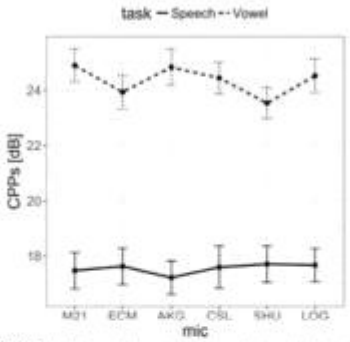


FIGURE 4. Mean and the standard error of Smoothed Cepstral Peak Prominence (CPPS) measured with the six microphones in the sound booth for the speech and the vowel tasks.

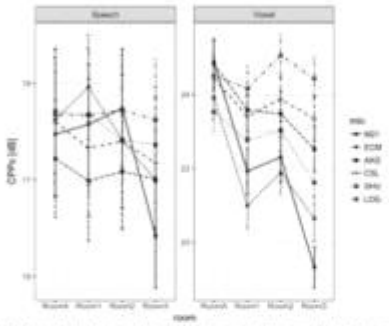


FIGURE 5. Mean and the standard error of Smoothed Cepstral Peak Prominence (CPPS) measured in the four rooms for the speech and the vowel tasks.

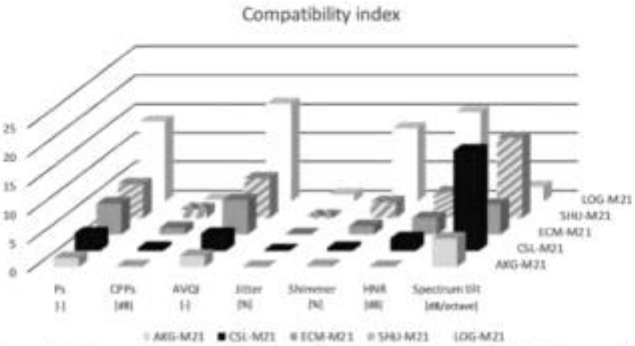


FIGURE 16. Compatibility indices of the seven parameters (Ps, CPPS, AVQI, Jitter, Shimmer, HNR and Spectrum tilt) performed with the different microphones (AKG, CSL, ECM, SHU, LOG) compared to the M211 in RoomA.

12

Microphone

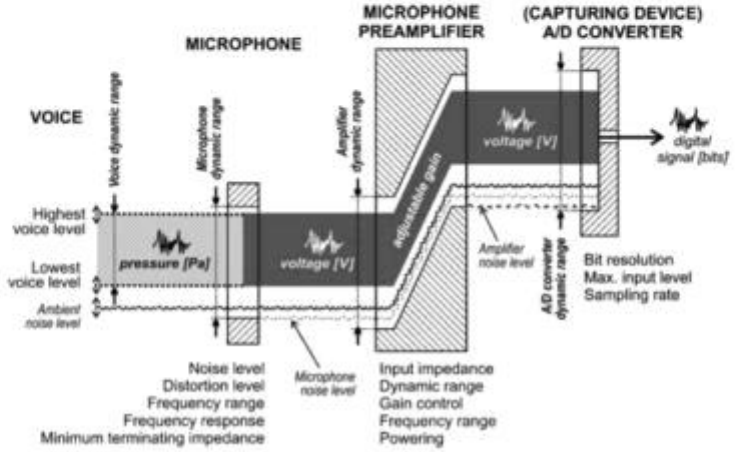
Tutorial

Guidelines for Selecting Microphones for Human Voice Production Research

Jan G. Švec
Palacky University Olomouc, Olomouc, the Czech Republic
Svante Granqvist
Royal Institute of Technology, Stockholm, Sweden

- 변환기(transducer)
  - Electret, condenser: 전원필요
  - Dynamic: 전원 불필요

FIGURE 9. The path of the sound signal through the microphone and the preamplifier to the digital capturing device. The parameters that should be well selected in voice measurement are listed below each of the equipment parts. The goal is to keep the F_0, spectrum, and level of the voice signal unperturbed as well as to keep the noise level well below the signal level. A/D = analog-to-digital.



Microphone

- 고급 헤드마운트 마이크 권고
  - 명확한 specification sheet가 있는 마이크 선택
    - 주파수응답성, 지향성, 근접효과, 탄탄한 마운트 고려
    - 예
      - Sennheiser HS2, DPA 4066, Countryman Isomax B3
  - 지향성
    - 무지향성 마이크(omni-directional / non-directional)
      - 조용한 방의 가까운 거리에서 권고됨
    - 단일지향성 마이크(directional / uni-directional / cardioid)
      - 정확하게 같은 거리, 방향 유지 중요

FIG. 1. Hybner International harmonica holder with Beyer dynamic unidirectional microphone positioned for routine voice recording.

Price & Sataloff, J Voice 1988

NCVS technical memo, 2006
https://archive.ncvs.org/ncvs/library/tech/NCVSOnlineTechnicalMemo07.pdf

NCVS technical memo, 2012
https://archive.ncvs.org/e-learning/NCVSOnlineTechnicalMemo15-vs1_1.pdf

주파수 응답성

- 지향성 마이크는 근접효과 (proximity effect)나 저주파음 증폭 주의

Shure SM58
단일지향성
다이나믹 마이크

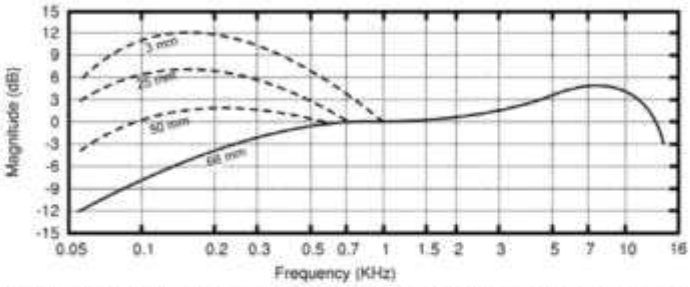


Figure 3. (solid line) *Frequency Response:* The sensitivity of a component to various frequencies (i.e., the amount in dB each frequency is amplified or attenuated where 0 dB means no change), usually graphed as an amplitude or gain (vertical axis) plotted against frequency (horizontal axis). (dashed lines) *Proximity Effect:* For directional microphones, the frequency response of the microphone changes at the low frequency end, as the microphone-to-mouth distance changes.

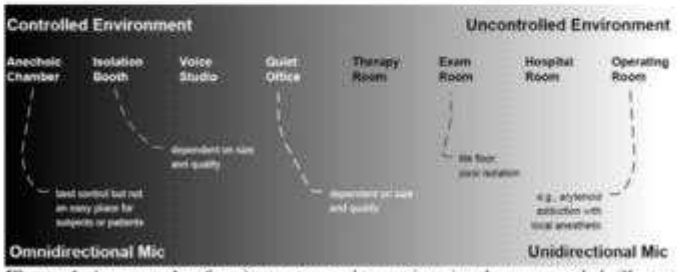


Figure 4. An example of various rooms where voice signals are recorded, illustrating the range from the most controlled to the most uncontrolled acoustic environment with preferred microphone type.

주파수 응답성

- 넓고 평활한 주파수 응답성 중요
  - 어떤 마이크도 완전히 flat할 수는 없음
  - 스튜디오/무대 마이크의 경우 음성이 더욱 호소력 짙게 하기 위해 3~10kHz에 피크를 주는 경우가 있는데 이러한 마이크는 지양해야 함

dB 내 변

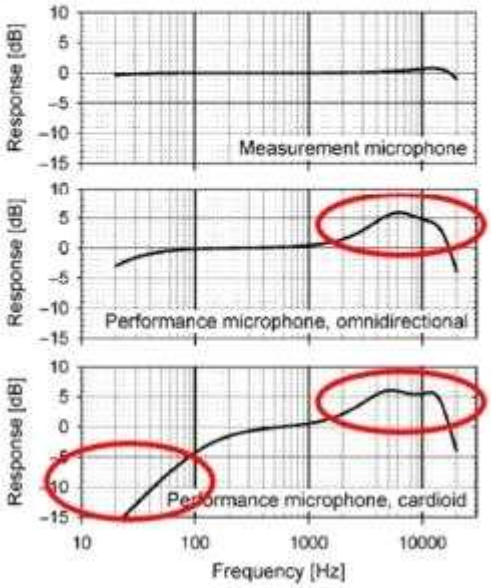
Tutorial

Guidelines for Selecting Microphones for Human Voice Production Research

Jan G. Švec
Palacky University Olomouc, Olomouc, the Czech Republic

Svante Granqvist
Royal Institute of Technology, Stockholm, Sweden

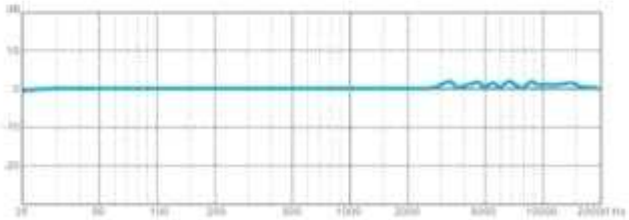
FIGURE 2. Frequency responses for three typical microphones: an omnidirectional measurement microphone, an omnidirectional studio microphone, and a cardioid microphone for stage use (here the response is valid for the far field). Note the broad presence peak in the 3–10-kHz region for the two latter microphones.



Spec sheet

- PRESONUS PRM1

Microphone	
Polar Pattern	Omnidirectional
Type	Back-Electret Condenser
Sensitivity	37 dB/Pa (14 mV/Pa) 1 Pa=94 dB SPL (at 1 kHz Open-Circuit Voltage)
Element	Premium JFET Preamplifier
Equivalent Input Noise	26 dB (A-Weighted)
Maximum SPL	132 dB SPL (THD ≤ 1%, 1 kHz) (1 kΩ Load)
Phantom Power Requirements	9 - 52V
Signal To Noise Ratio	70 dB
Dynamic Range	106 dB (1 kΩ Load)
Frequency Response	20 Hz to 20 kHz



https://www.presonus.com/en/misc/more/microphones/2777300105.html

Spec sheet

- AKG P420

P420

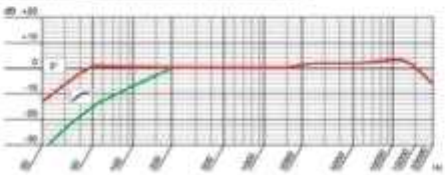
LARGE DIAPHRAGM DUAL-CAPSULE TRUE CONDENSER MICROPHONE

KEY SPECIFICATIONS

by HARMAN

Capsule	1" dual capsule condenser
Polar Pattern	cardioid, omnidirectional, figure 8
Frequency Response	20 Hz to 20 kHz (see frequency response trace)
Sensitivity (Cardioid)	28 mV/Pa (-31 dBV)
Max. SPL for 0.5 % THD	135/155 dB (0/-20 dB pad)
Equivalent Noise Level To IEC 60286-4	15 dB (A)
Signal To Noise	79 dB (A) re 1Pa
Preattenuation Pad	-20 dB
Bass Cut Filter	300 Hz, 12 dB/octave
Electrical Impedance	≤ 200 ohms
Recommended Load Impedance	≥ 1000 ohms
Connector	3-pin XLR (pin 2 hot)
Phantom Powering	48 V +/- 4 V to IEC 61938
Current Consumption	≤ 3 mA
Temperature Range	-10 °C to +60 °C (14 °F - 140 °F)
Diameter	54 mm (2.13 in.)
Length	165 mm (6.5 in.)
Net Weight (Mic Only)	530 g (18.7 oz.)
Shipping Weight	1730 g (61oz.)
Box Dimensions (L x W x H)	23 x 25 x 14 cm (9 x 9.8 x 5.5 in)

Frequency Response in cardioids mode:



Polar Pattern:

http://recordinghacks.com/pdf/akg/P420-cutsheet.pdf

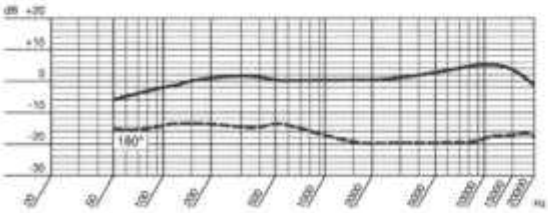
Spec sheet

- AKG C420

Specifications

Transducer type:	Condenser pressure gradient
Frequency response:	20 Hz - 20 kHz
Polar pattern:	Cardioid
Impedance:	200 ohms
Output connector:	XLR (C420), mini XLR (C420/L)
Sensitivity:	7 mV/Pa; -43 dB (re 1 V)
Powering:	Phantom powering from 9 to 52 Vdc with power adapter
Sound pressure level for 1% /3% THD: 126/130 dB	
Size:	4.9 in. (125 mm) x 5.0 in. (130 mm)
Net/Shipping weight:	1.1 oz./19.6 oz. (30 g/540 g)
Standard accessory:	W 44 Windscreen
Optional Accessories:	B 29 L battery operated phantom power supply with a 2 input mixer (used with the C 420 L version) B 18 E battery operated phantom power supply. (used with C 420) B 15 battery operated power supply

Cardioid



https://www.akg.com/on/demandware.static/-/Sites-masterCatalog_Harman/default/dw3393f9ac/pdfs/C420_cutsheet.pdf

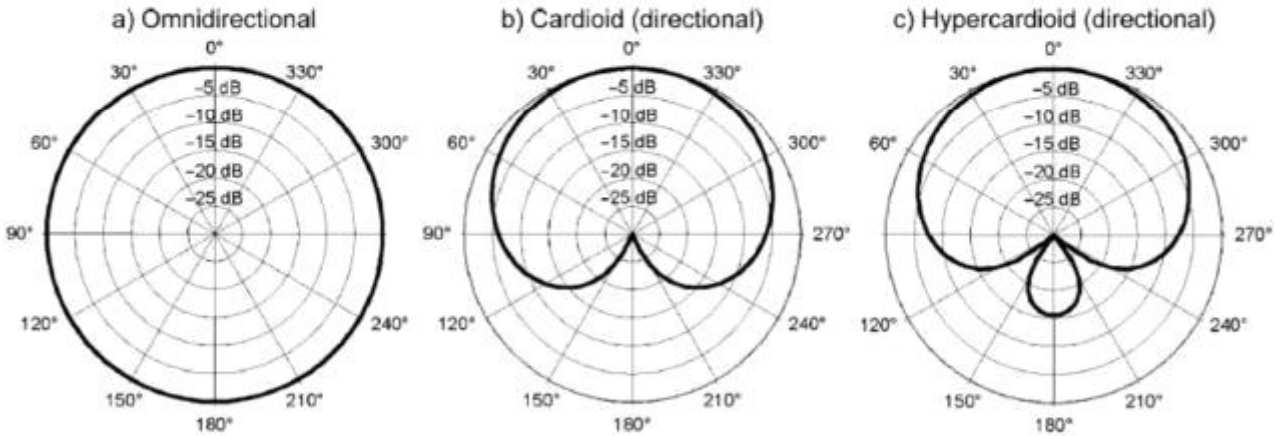
마이크의 지향성

Tutorial

Guidelines for Selecting Microphones for Human Voice Production Research

Jan G. Švec
Palacky University Olomouc, Olomouc, the Czech Republic
Svante Granqvist
Royal Institute of Technology, Stockholm, Sweden

FIGURE 3. Examples of polar plots of three types of microphones: (a) omnidirectional, (b) cardioid, and (c) hypercardioid. The hypercardioid directionality can vary between designs, in particular with respect to the level of the lobe directed toward 180°.



Signal-to-noise ratio

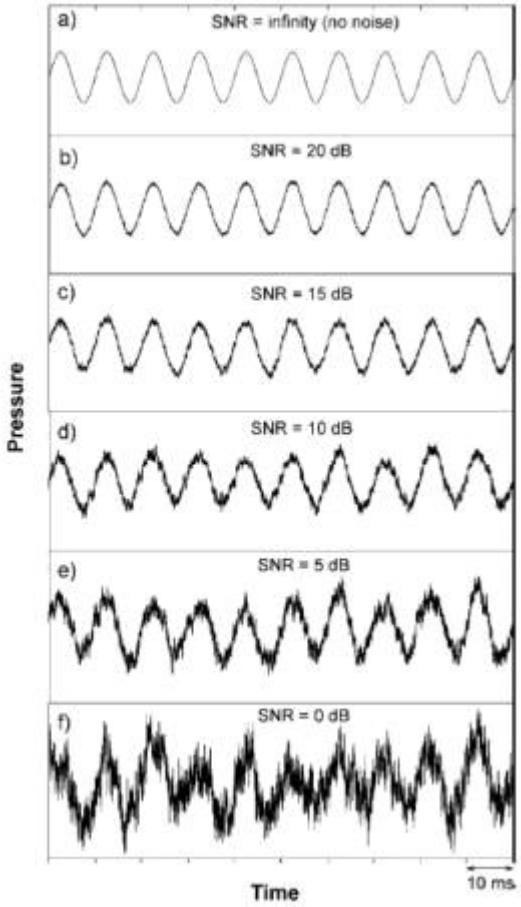
- At least 30 dB
  - For perturbation measurements

FIGURE 7. The influence of noise on the pressure waveform: (a) a clean periodic signal with no noise—that is, with infinite signal-to-noise ratio (SNR); (b–e) the same waveform under the conditions when pink noise is added such that the noise level is (b) 20 dB, (c) 15 dB, (d) 10 dB, or (e) 5 dB weaker than the signal; and (f) when the noise has the same level as the signal. Notice the increased perturbation of the waveform when the noise level is increased. For accurate measurement of voice, the signal is recommended to be at least 15 dB stronger than the internal noise of the microphone and the background noise (cases a–c).

Tutorial

Guidelines for Selecting Microphones for Human Voice Production Research

Jan G. Švec
Palacky University Olomouc, Olomouc, the Czech Republic
Svante Granqvist
Royal Institute of Technology, Stockholm, Sweden



Variability of microphones used for acoustic analysis of the voice in the last twenty years

Article in Revista de Investigación e Innovación en Ciencias de la Salud · December 2020
DOI: 10.46634/riics.36

Immediate acoustic effects of straw phonation exercises in subjects with dysphonic voices [13]	Samson MM-01 (descontinuado)	No
Alteración de parámetros acústicos de la voz y el habla en la enfermedad de Parkinson [14]	No especificado	No
Eficacia de la rehabilitación de la voz en etapas tempranas de la Enfermedad de Parkinson [15]	SHURE SM-58	No
Primeros formantes y área de espacio vocal en una pequeña muestra de pacientes con Parkinson [16]	No especificado	No
Efficacy of water resistance therapy in subjects diagnosed with behavioral dysphonia: a randomized controlled trial [17]	Rode NT2-A	Si
Comparison of effects produced by physiological versus traditional vocal warm-up in contemporary commercial music singers [18]	AKG CK-77	No
The impact of double source of vibration semioccluded voice exercises on objective and subjective outcomes in subjects with voice complaints [19]	AKG CK-77	No
Water Resistance Therapy as Vocal Warm-Up Method in Contemporary Commercial Music Singers [20]	AKG CK-77	No
Documentación lingüística del quechua de Chachapoyas: Registro acústico de la monoptongación [21]	HTDZ HT-81	No
	SHURE WH-20	No
Immediate Effects of the Semi-Occluded Ventilation Mask on Subjects Diagnosed With Functional Dysphonia and Subjects With Normal Voices [22]	AKG Perception-120	Si
Duration of biodynamic changes associated with water resistance therapy [23]	Behringer ECM-8000	Si
	Sennheiser ME-62	No
Does a Systematic Vocal Exercise Program Enhance the Physiologic Range of Voice Production in Classical Singing Graduate-Level Students? [24]	SHURE SM-48	No

Variability of microphones used for acoustic analysis of the voice in the last twenty years

Article in Revista de Investigación e Innovación en Ciencias de la Salud - December 2020
DOI: 10.46634/riics.38

Tabla 1. Micrófonos y criterios de selección

Nombre del Artículo	Modelo del micrófono seleccionado	¿Criterios de selección adecuados?
Long-term average spectrum (LTAS) analysis of sex- and gender-related differences in children's voices [4]	No especificado	No
Análisis acústico de la voz: fiabilidad de un conjunto de parámetros multidimensionales [5]	SHURE SM-58	No
Computer-assisted voice analysis: Establishing a pediatric database [6]	No especificado	No
Early effects of smoking on the voice: A multidimensional study [7]	SHURE SM-58	No
Evaluación perceptual de la disfonía: correlación con los parámetros acústicos y fiabilidad [8]	No especificado	No
Examen de la función vocal y análisis acústico de 905 docentes de La Rioja [9]	AKG-190	No
Performance of Gaussian mixture models as a classifier for pathological voice [10]	No especificado	No
Spectral- and cepstral-based measures during continuous speech: capacity to distinguish dysphonia and consistency within a speaker [11]	No especificado	No
Cambios acústicos de la voz como signos de fatiga vocal en locutores de radio: resultados preliminares [12]	AKG Perception-120	Si

Patel et al. AJSLP 2018
NCVS technical memo, 2006
https://archive.ncvs.org/ncvs/library/tech/NCVSOnlineTechnicalMemo07.pdf

한림대학교 HALLYM UNIVERSITY

Sound level meter

- Class I 사운드 레벨 미터 권고
  - 클래스: 표준에 따른 허용 오차의 수준
  - C 주파수 가중 설정 권고: 대략 10kHz까지 동일하게 측정
- 절대적인 SPL의 측정을 원하는 경우 calibration 절차 수행
  - 마이크와 SLM을 모두 대상자의 입으로부터 30cm에 위치시킴
  - /아/모음 발성하여 두 기기에서 동시 측정
  - SLM에서 측정된 SPL과 입력된 SPL이 일치되도록 입력레벨 조절
  - 또는, SPL@30 cm = SPL@d − 20 log(30/d)
    - d=마이크에서 입의 중심까지의 거리(cm)
    - 예) 10cm의 경우: SPL@30 cm = SPL@10 cm − 9.5 dB
  - 이것이 어려운 경우 최소한 마이크와의 거리 일정하게 유지
    - 상대적 비교만 가능

Microphone power / amplifier / converter

- 팬텀파워
  - 컨덴서 마이크
  - 12V / 48V

- 고품질 프리앰프
  - 이퀄라이저, 자동 게인조절, 노이즈 캔슬 등 금지

NCVS technical memo, 2006
https://archive.ncvs.org/ncvs/library/tech/NCVSOnlineTechnicalMemo07.pdf

25

Microphone power / amplifier / converter

- 오디오 인터페이스 / 믹서
  - 프리앰프+팬텀파워+AD 컨버터
  - Tascam, Edirol, Steinberg, Behringer, MAudio ...
  - 가격 천차만별

- 가능하면 데이터 백업 시스템 갖출 것

NCVS technical memo, 2006
https://archive.ncvs.org/ncvs/library/tech/NCVSOnlineTechnicalMemo07.pdf

26

기타

- 케이블
  - 수준에 따라 가격 차이가 큼
  - 다양한 커넥터 정확히 확인
    - 무선 마이크팩에 연결하는 커넥터는 제조사마다 다를 수 있음

3.5mm(3극, 4극)

RCA

XLR(캐논)

TRS(스테레오)
TS(모노)

https://blog.naver.com/samasound/223044100485

NCVS technical memo, 2006
https://archive.ncvs.org/ncvs/library/tech/NCVSOnlineTechnicalMemo07.pdf

기타

- 컴퓨터
  - 성능을 다소 희생하더라도 팬 소음 조절
- 녹음 소프트웨어
  - 최소 2채널 녹음 기능
- 캠코더
- 의자
  - 조절가능한 바퀴 없는 의자
- 청지각 실험 시
  - 모니터링 스피커/헤드폰

NCVS technical memo, 2006
https://archive.ncvs.org/ncvs/library/tech/NCVSOnlineTechnicalMemo07.pdf

기타

- 마이크스탠드
  - 가급적 T자형 스탠드 권고
  - 안정적 거치가 매우 중요

29

다양한 기기들

삽질의 역사

30

디지털 녹음기(digital recorder)

- Sony Hi-MD
  - MD에 녹음된 파일을 컴퓨터로 별도로 옮겨야 함
- 마이크
  - ECM-MS907
- 휴대성은 좋으나 품질 보장 못함

디지털 녹음기(digital recorder)

- Sony PCM M10
  - 표본추출율(sampling rate)
    - 44.1 kHz
  - 양자화(quantization)
    - 16-bit
  - 자체 마이크도 성능이 꽤 좋음
    - 무지향성
    - Low-cut filter 설정 반드시 유의!
- 단일지향성 다이나믹 마이크 (cardioid dynamic microphone)
  - SHURE, SM48
  - 단점: input이 너무 작아짐

오디오 인터페이스(USB-based audio interface)

- A fan-less tablet PC
  - Surface Pro 4, Core m3 model; Microsoft, Redmond, WA, U.S.A
- A USB-based audio interface
  - UR-22; Steinberg Media Technologies GmbH, Hamburg, Germany

33

프리앰프(preamplifier)

- Tube Ultragain MIC200(Behringer GmbH, Kirchardt, Germany)

34

USB 콘덴서 마이크

SENNHEISER

Smartphone? Microphone?

Smartphone Recordings are Comparable to "Gold Standard" Recordings for Acoustic Measurements of Voice

*Shaheen N. Awan, †Mohsin Ahmed Shaikh, ‡Jordan A. Awan, §Ibrahim Abdalla, ‖Kelvin O. Lim, and ¶Stephanie Misono, *Tampa, FL, †Bloomsburg, PA, ‡West Lafayette, IN, and §‖¶Minneapolis, MN

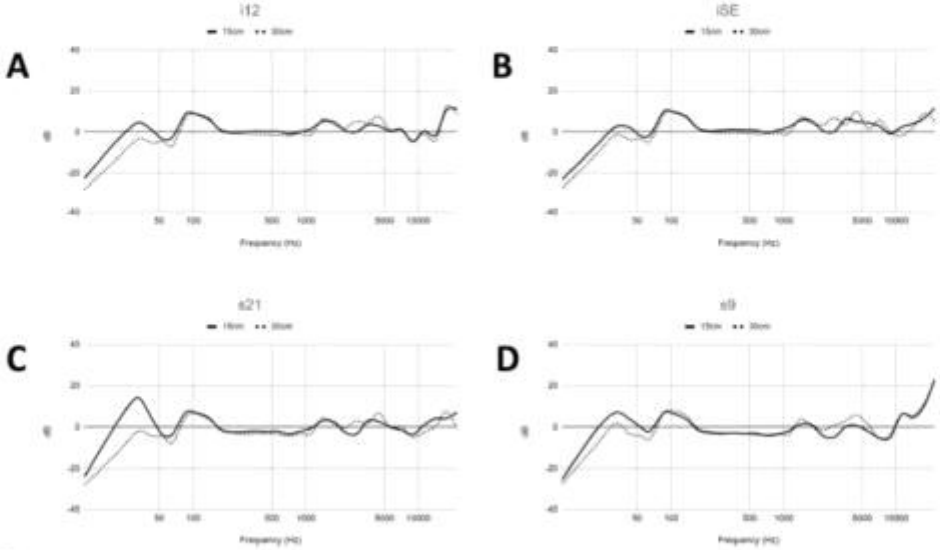


FIGURE 1. Frequency response characteristics for the tested smartphone devices (i12 = iPhone 12, iSE= iPhone SE, s21 = Samsung Galaxy s21, s9 = Samsung Galaxy s9,).

Smartphone Recordings are Comparable to "Gold Standard" Recordings for Acoustic Measurements of Voice

*Shaheen N. Awan, †Mohsin Ahmed Shaikh, ‡Jordan A. Awan, §Ibrahim Abdalla, ‖Kelvin O. Lim, and ¶Stephanie Misono, *Tampa, FL, †Bloomsburg, PA, ‡West Lafayette, IN, and §‖¶Minneapolis, MN

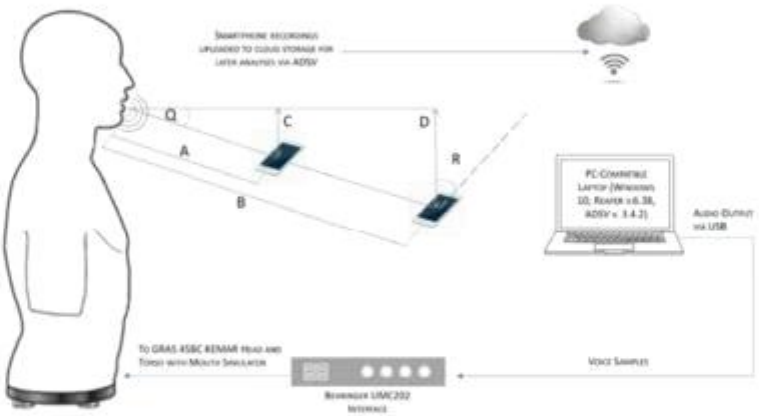


FIGURE 2. Schematic of the voice sample playback and smartphone recording setup. Smartphone recordings were uploaded to cloud age and later analyzed using the Analysis of Dysphonia in Speech and Voice program (ADSV v. 3. 4.2). Smartphone positioning in rel to the KEMAR head and torso (HAT) model: A = 15cm distance; B = 30 cm distance; C = 7.5 cm; D = 15 cm; Q = 30 degrees; R degrees. For the purposes of illustration clarity, distances are not to scale.

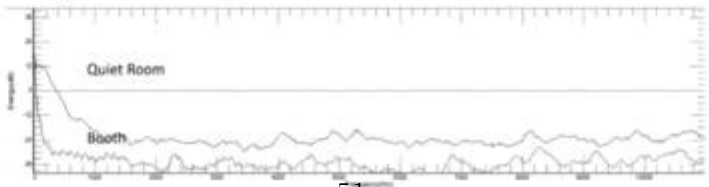


FIGURE 3. Background noise spectra obtained using the SLM in the sound-treated booth vs. quiet room.

Smartphone Recordings are Comparable to "Gold Standard" Recordings for Acoustic Measurements of Voice

[*]Shaheen N. Awan, [†]Mohsin Ahmed Shaikh, [‡]Jordan A. Awan, [§]Ibrahim Abdalla, [‖]Kelvin O. Lim, and [¶]Stephanie Misono, *Tampa, FL, †Bloomsburg, PA, ‡West Lafayette, IN, and §‖¶Minneapolis, MN

TABLE A1.
Mean Pearson's *r* correlations across smartphones vs. the sound level meter standard (SLM, Booth, at 15cm). Standard deviations (in parentheses) and correlation ranges are also provided. All correlations are significant at $p < .001$.

Dependent Variable	Across Setting and Distance	Booth 15cm	Booth 30cm	Quiet Room 15cm	Quiet Room 30 cm
CPP Vowel	0.969 (0.018) Range: 0.943 - 0.993	0.981 (0.003) Range: 0.977 - 0.985	0.989 (0.003) Range: 0.986 - 0.9993	0.950 (0.005) Range: 0.944 - 0.956	0.957 (0.011) Range: 0.943 - 0.969
CPP Sentence	0.955 (0.021) Range: 0.920 - 0.980	0.972 (0.002) Range: 0.970 - 0.974	0.976 (0.004) Range: 0.970 - 0.980	0.931 (0.015) Range: 0.920 - 0.954	0.942 (0.011) Range: 0.935 - 0.959
L/H Ratio Vowel	0.950 (0.020) Range: 0.908 - 0.979	0.957 (0.018) Range: 0.941 - 0.978	0.938 (0.033) Range: 0.908 - 0.979	0.951 (0.008) Range: 0.941 - 0.959	0.956 (0.017) Range: 0.940 - 0.977
L/H Ratio Sentence	0.942 (0.023) Range: 0.892 - 0.970	0.952 (0.015) Range: 0.938 - 0.966	0.938 (0.020) Range: 0.918 - 0.965	0.930 (0.035) Range: 0.892 - 0.962	0.950 (0.018) Range: 0.931 - 0.970
CSID Vowel	0.971 (0.012) Range: 0.935 - 0.989	0.981 (0.008) Range: 0.971 - 0.989	0.976 (0.007) Range: 0.969 - 0.983	0.968 (0.004) Range: 0.962 - 0.973	0.960 (0.017) Range: 0.935 - 0.974
CSID Sentence	0.969 (0.006) Range: 0.959 - 0.981	0.971 (0.004) Range: 0.968 - 0.977	0.968 (0.005) Range: 0.962 - 0.973	0.968 (0.006) Range: 0.963 - 0.976	0.971 (0.011) Range: 0.959 - 0.981

CPP: Cepstral Peak Prominence; L/H Ratio: Low vs. High Spectral Ratio (4 kHz cutoff); CSID: Cepstral Spectral Index of Dysphonia

JASA ARTICLE

Impact of naturalistic smartphone positioning on acoustic measures of voice[a)]

Victoria S. McKenna,[1,b)] Rachel M. Roberts,[1] Aaron D. Friedman,[2] Savannah N. Shanley,[1] and Andres F. Llico[3]
[1]Department of Communication Sciences and Disorders, University of Cincinnati, Cincinnati, Ohio 45267, USA
[2]Department of Otolaryngology–Head and Neck Surgery, University of Cincinnati, Cincinnati, Ohio 45267, USA
[3]Department of Biomedical Engineering, University of Cincinnati, Cincinnati, Ohio 45221, USA

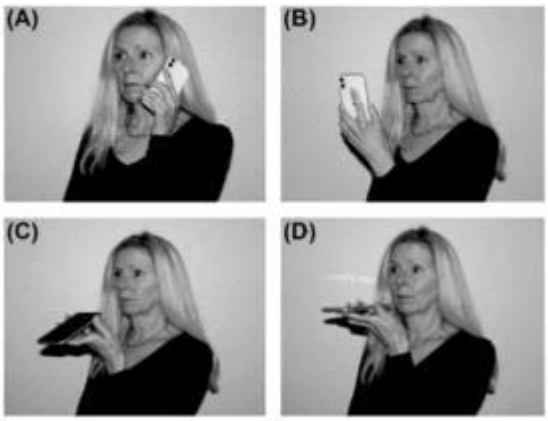
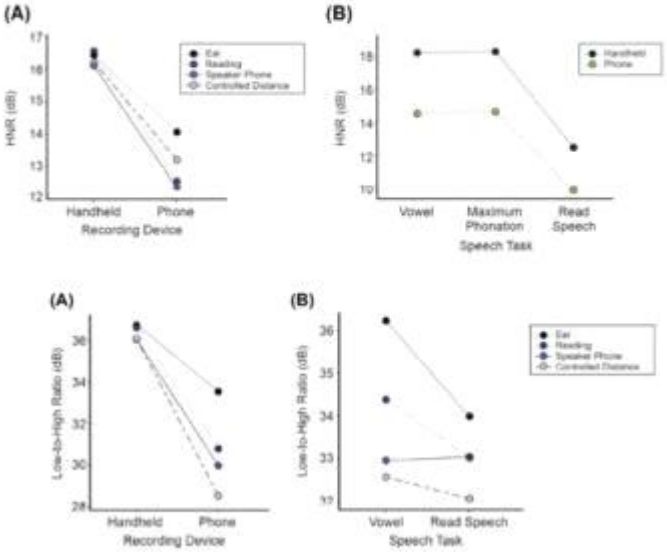


FIG. 1. Phone positions. (A) Ear position. (B) Reading position. (C) Speaker phone position. (D) Controlled distance position of 8.5 cm from the lips, parallel to the ground.



대한후두음성언어의학회지
제 26 권 제 2 호 2015

□원 저□

음향분석에 사용할 녹음장비로 갤럭시 스마트폰 녹음기능의 유용성

성균관대학교 의과대학 강북삼성병원 이비인후과학교실
윤태화 · 이재혁 · 이상혁 · 진성민

- CSL, 갤럭시 S5, Sony PCM-M10
  - 녹음환경 언급 없음

Table 4. Correlation between 8 parameters of portable recorder, CSL and galaxy according to the population (n=30)

	C_Fo	C_jitter	C_shim	C_NHR	C_F1	C_F2	C_F3	C_F4
G_Fo	1.000**	.419*	.412*	.086	.024	.046	−.232	−.017
G_jitter	.534**	.438*	.417*	.016	.338	.511**	.220	.332
G_shim	.305	.529**	.491**	.247	−.012	−.056	−.220	−.108
G_NHR	−.005	.220	.135	.829**	−.365*	−.182	−.174	−.250
G_F1	.566**	.412*	.344	.055	.185	−.052	−.405*	−.089
G_F2	.070	.044	−.063	−.158	.656**	.867**	.222	.594**
G_F3	−.101	−.225	−.238	−.004	.358	.549**	.606*	.484**
G_F4	.206	−.187	−.114	−.066	.493**	.597**	.318	.622**
	P_Fo	P_jitter	P_shim	P_NHR	P_F1	P_F2	P_F3	P_F4
G_Fo	1.000**	.614**	.160	−.228	.250	.096	−.167	−.032
G_jitter	.534**	.596**	.434*	.087	.635**	.550**	.243	.334
G_shim	.304	.391*	.659**	.450*	.046	.081	−.148	−.042
G_NHR	−.006	.103	.174	.731**	.011	−.112	−.116	−.107
G_F1	.566**	.339	.131	−.208	.165	−.048	−.214	.000
G_F2	.070	.095	.118	−.059	.507**	.896**	.227	.571**
G_F3	−.101	.061	−.067	−.108	.274	.510**	.528**	.349
G_F4	.206	.109	−.112	−.213	.490**	.601**	.354	.496**

Galaxy smartphone recording. * : $p<0.05$, ** : $p<0.01$, P : portable recorder, C : CSL, G : galaxy

臨床耳鼻 : 第 27 卷 第 2 號 2016
J Clinical Otolaryngol 2016;27:286-294

□원 저□

Recording Quality of Smartphone for Acoustic Analysis

Geun-Hyo Kim, MS[1,2], Duck-Hoon Kang, MS[3], Yeon-Yoo Lee, MS[1,2],
Hyung-Soon Kim, PhD[1,4], Soon-Bok Kwon, PhD[1,5], Young-Seuk Cho, PhD[6],
Hee-June Park, PhD[7], Yang-Gyu Choi, PhD[8] and Bum-Joo Shin, PhD[3]

J Clinical Otolaryngol 2016;27:286-294

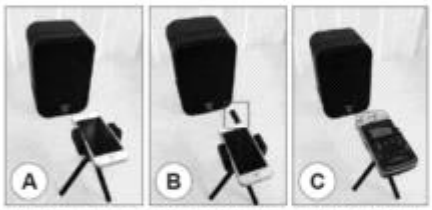


Fig. 2. Setup for a device recording in a conference room. The clean recording is played through the loudspeaker and recorded on a smartphone, capturing the noise and reverberation of the room as well as the limitations of the recording hardware. A : The smartphone signals were recorded with the smartphone placed in front of personal computer's mono speaker at a distance of approximately 15 cm. B : Smartphone recording setting with unidirectional microphone (*red dotted square*). C : Sony voice recorder setting.

Table 3. Results of correlations among each recording method for percent shimmer

	nhrl1	nhrl2	nhrl3	nhrl4	Sony	nhrA1	nhrA2	nhrA3	nhrA4
Spearman's r	.814**	.873**	.854**	.837**	.734**	.780**	.774**	.886**	.864**
p	.000	.000	.000	.000	.000	.000	.000	.000	.000

** : Significant at 0.01 level (two-tailed)

Table 4. Results of correlations among each recording method for noise-to harmonic ratio (NHR)

	nhrl1	nhrl2	nhrl3	nhrl4	Sony	nhrA1	nhrA2	nhrA3	nhrA4
Spearman's r	.814**	.873**	.854**	.837**	.734**	.780**	.774**	.886**	.864**
p	.000	.000	.000	.000	.000	.000	.000	.000	.000

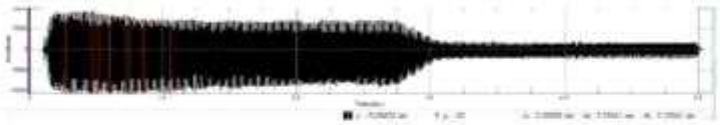
** : Significant at 0.01 level (two-tailed)

Table 5. Results of correlations among each recording method for cepstral prominence peak (CPP)

	cppl1	cppl2	cppl3	cppl4	Sony	cppA1	cppA2	cppA3	cppA4
Pearson's r	.986**	.994**	.994**	.996**	.996**	.995**	.989**	.996**	.992**
p	.000	.000	.000	.000	.000	.000	.000	.000	.000

** : Significant at 0.01 level (two-tailed)

Smartphone + wireless earphone



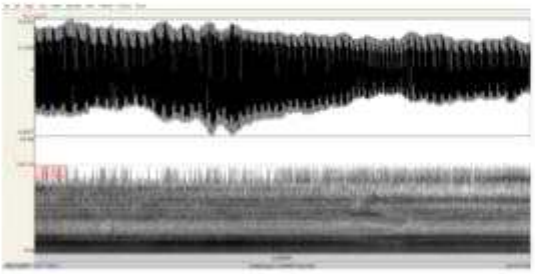


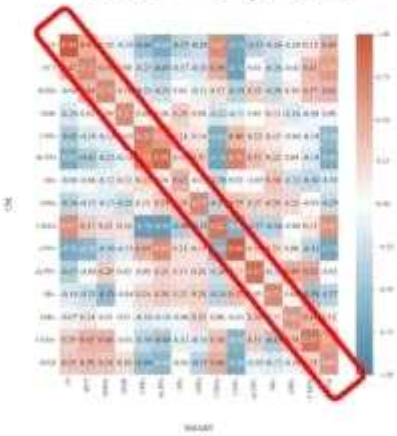
그림 2. 무선이어폰으로 녹음한 모음 발성 샘플에서 발생한 5 - 7 kHz 주파수 대역의 불규칙한 감쇄효과 예시

Figure 2. Example of irregular attenuation effects in the 5–7 kHz frequency range in a vowel sample recorded with wireless earphones

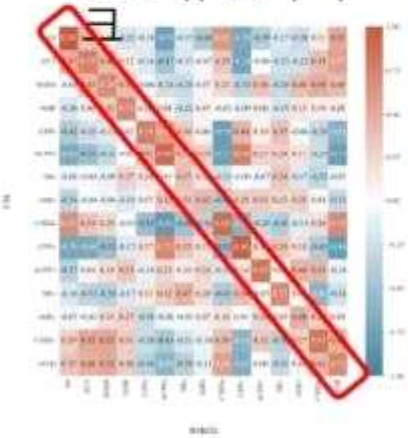
Park & Lee(2024). Phon Speech Sci

Correlation among acoustic measures

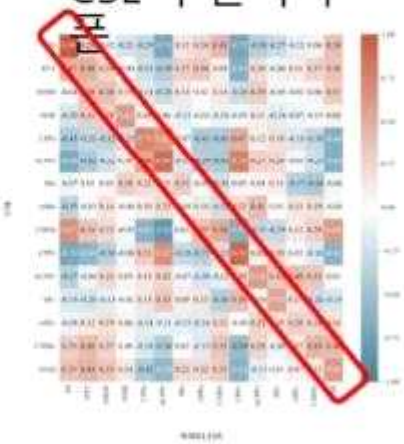
- CSL-스마트폰



- CSL-유선마이크



- CSL-무선이어폰



Park & Lee(2024). Phon Speech Sci

Smartphone + microphone: Wired

High-end Wired (HW)

XS LAV(SENNHEISER, Wedemark, Germany)
가격: 100,000원
주파수 응답성: 50 ~ 18000Hz
무지향성

Low-end Wired (LW)

HC-220UC(DoowonTrade, Daegu, Korea)
가격: 13,520원
주파수 응답성: 50 ~ 16000Hz
단일지향성

Park & Lee(2024). 한국언어청각임상학회 학술대회

45

Smartphone + microphone: Wireless

High-end Wireless (HWL)

MoveMic(SHURE, Niles, IL, USA)
400,000원
주파수 응답성: 50 ~ 20000Hz
무지향성

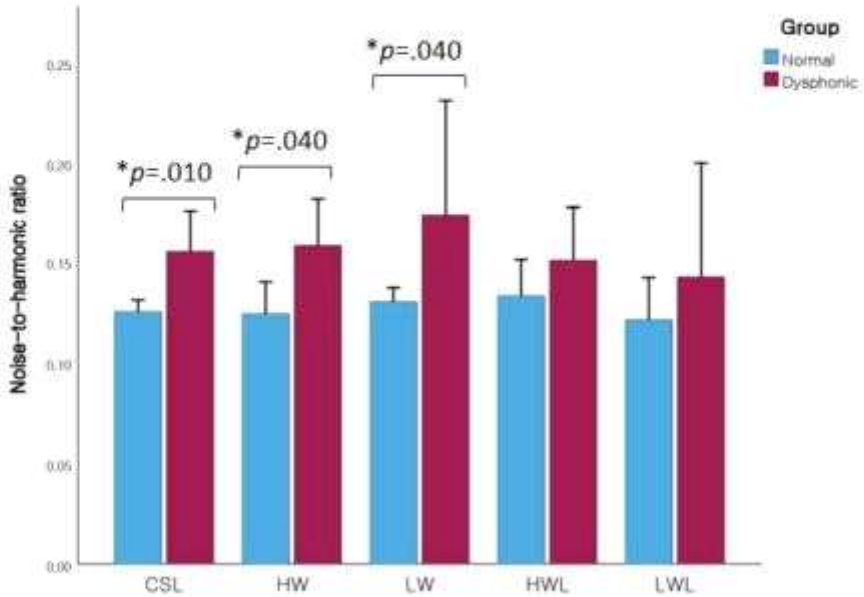
Low-end Wireless (LWL)

Wireless 1(MATA STUDIO, Ansan, Korea)
가격: 39,900원
주파수 응답성: 20 ~ 20000Hz
무지향성

Park & Lee(2024). 한국언어청각임상학회 학술대회

46

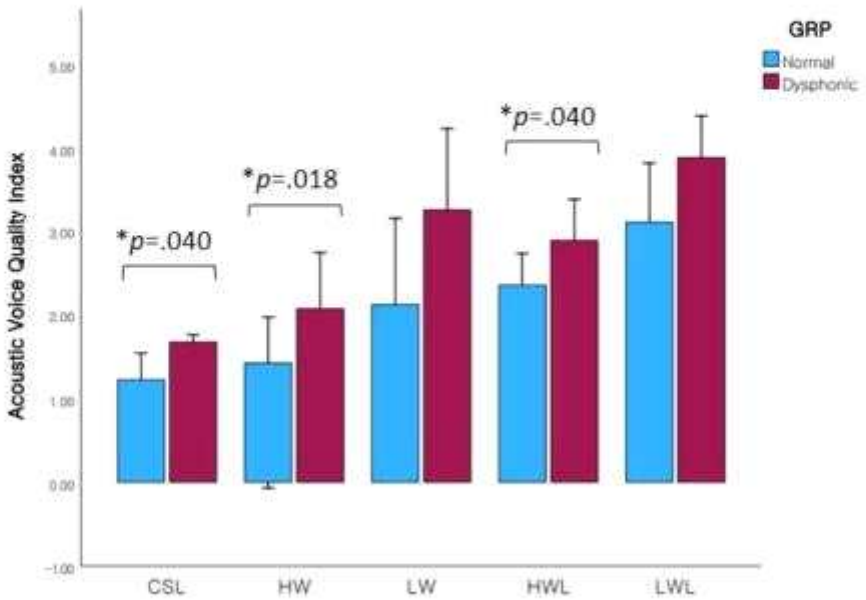
NHR



Park & Lee(2024). 한국언어청각임상학회 학술대회

47

AVQI

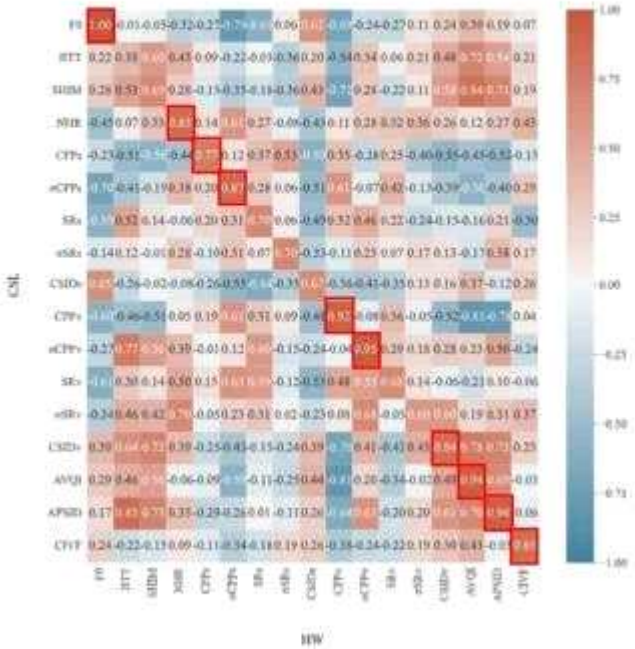


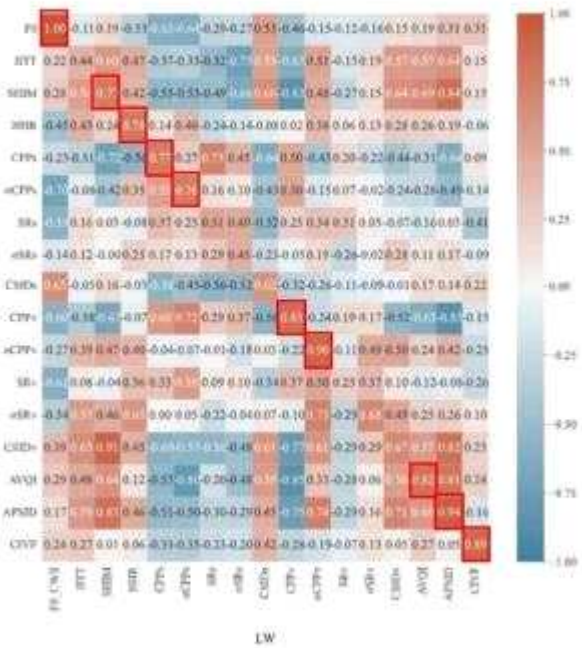
Park & Lee(2024). 한국언어청각임상학회 학술대회

48

CSL과 유선마이크 간 상관관계

Park & Lee(2024). 한국언어청각임상학회 학술대회





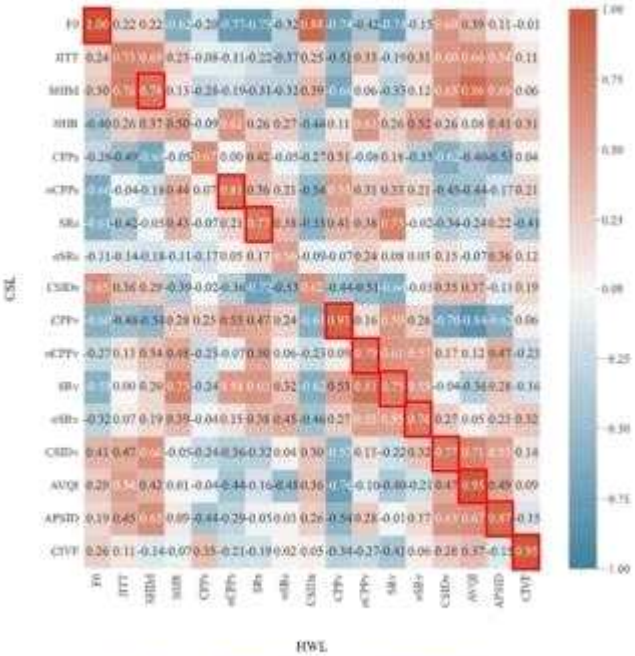
≥.75: F0, NHR, CPPs, σCPPs, CPPv, σCPPv, CSIDv, AVQI, APSID, CIVF

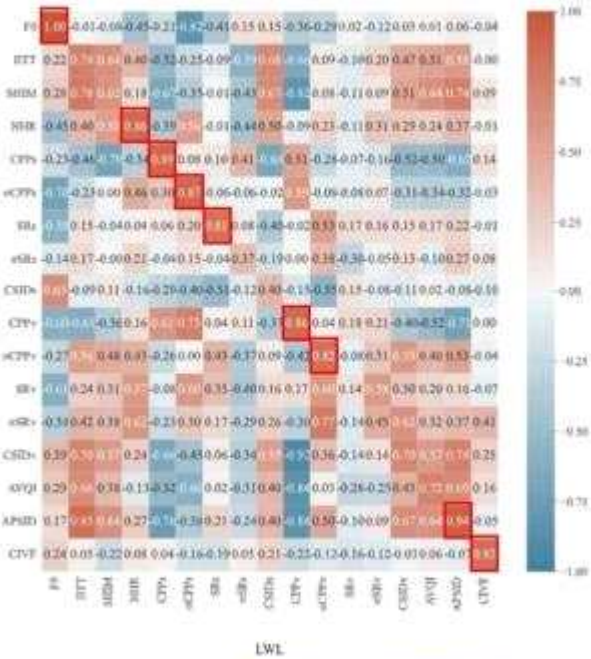
≥.75: F0, SHIM, NHR, CPPs, σCPPs, CPPv, σCPPv, AVQI, APSID, CIVF

49

CSL과 무선마이크 간 상관관계

Park & Lee(2024). 한국언어청각임상학회 학술대회





≥.75: F0, SHIM, σCPPs, SRs, CPPv, σCPPv, SRv, σ SRv, CSIDv, AVQI, APSID, CIVF

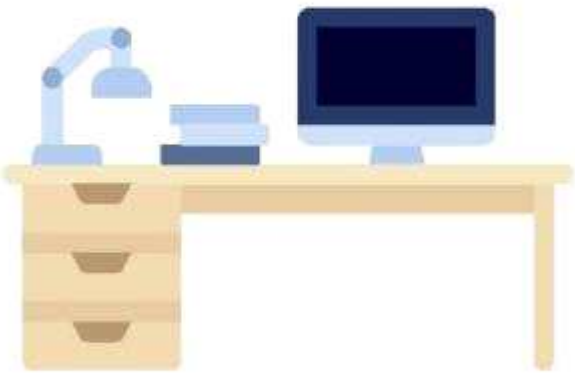
≥.75: F0, NHR, CPPs, σCPPs, SRs, CPPv, σCPPv, APSID, CIVF

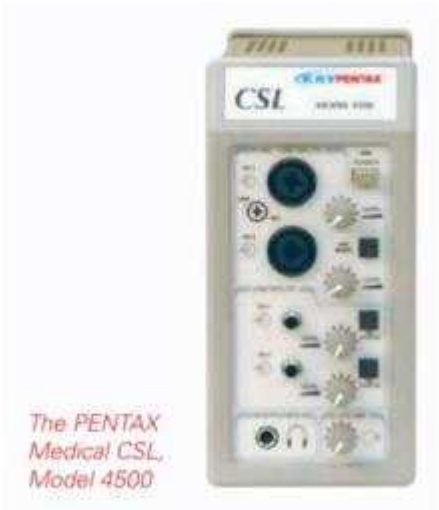
50

녹음 및 분석 프로그램

한림대학교
HALLYM UNIVERSITY

Computerized Speech Lab (CSL)





The PENTAX
Medical CSL,
Model 4500

CSL + audio interface + Preamp

53

CSL + sound-treated booth

54

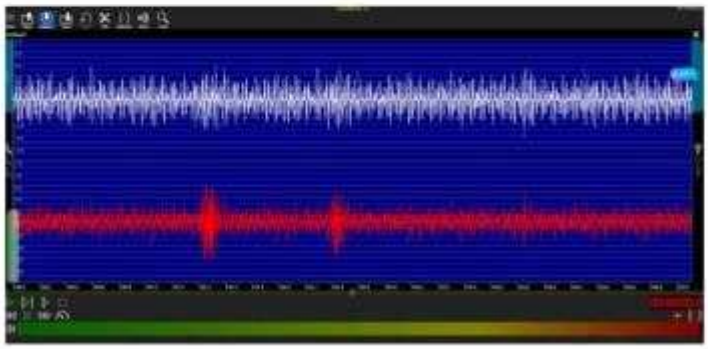
Windows 녹음기

녹음 프로그램: Audacity

https://www.audacityteam.org/

녹음 프로그램: Goldwave

https://www.goldwave.com/goldwave.php



https://www.goldwave.com/goldwaveapp.php

녹음 프로그램: Wavepad

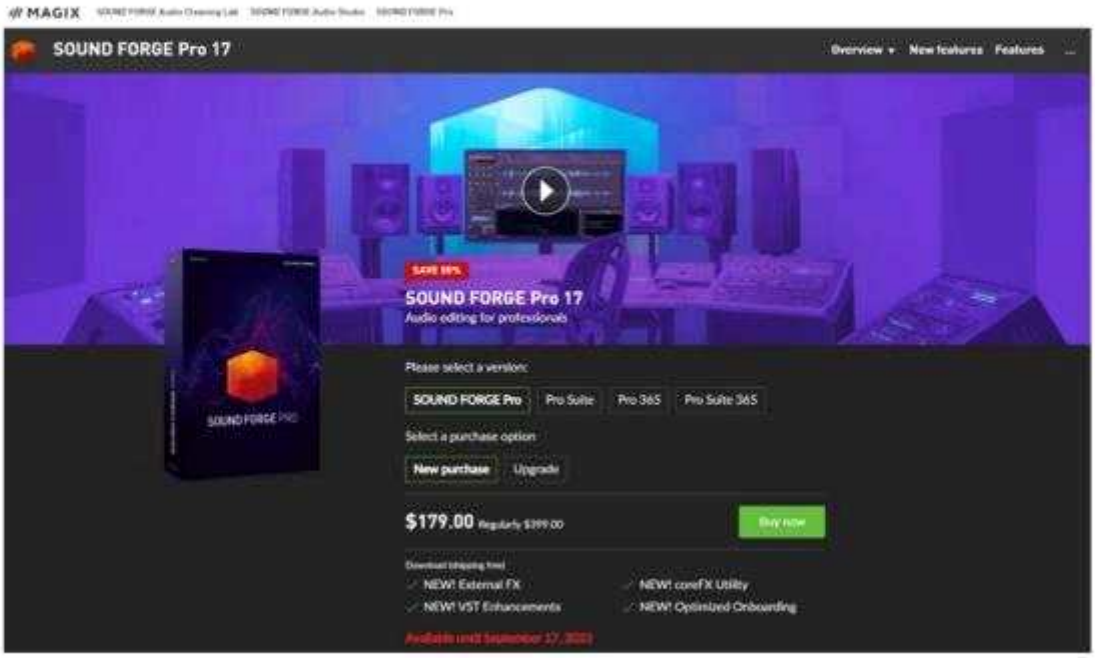
녹음 프로그램: Ardour

https://ardour.org/

녹음 프로그램: Adobe Audition

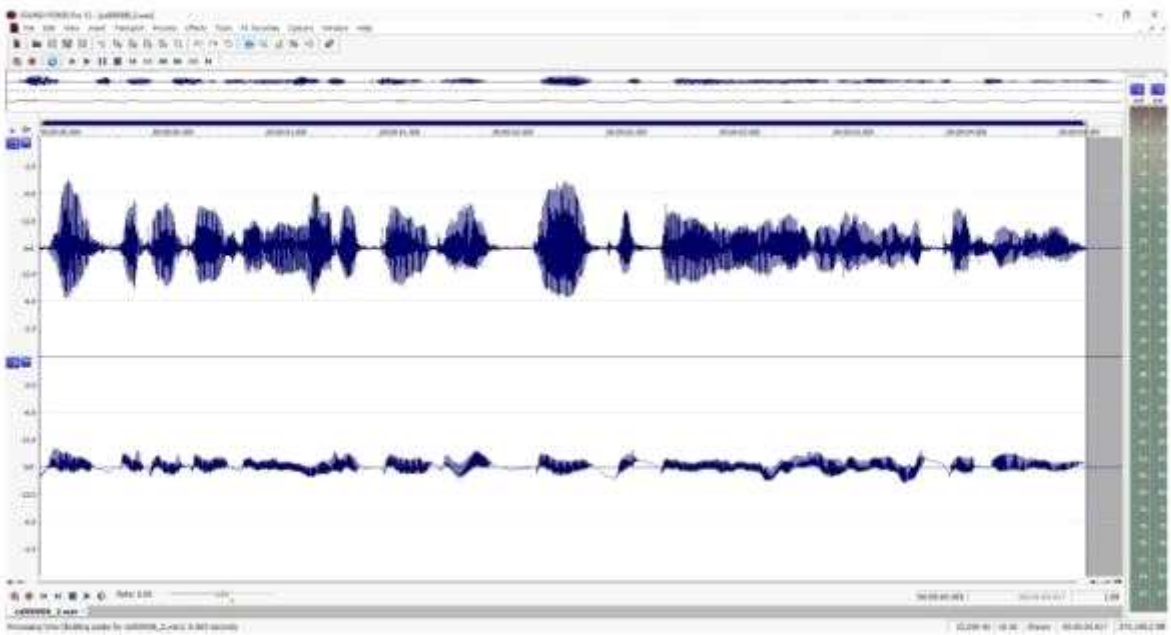
https://www.adobe.com/kr/products/audition.html

녹음 프로그램: Sound Forge

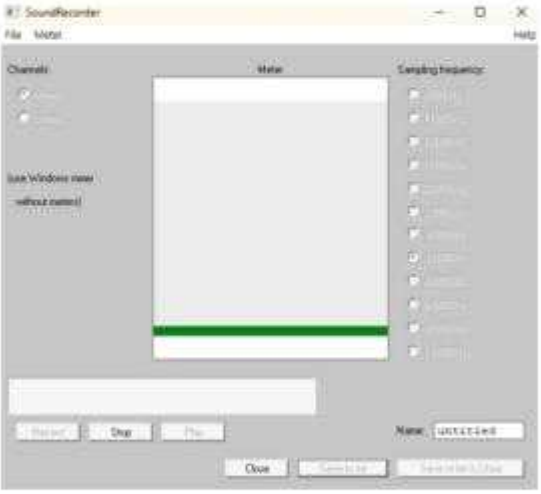


https://www.magix.com/us/sem/sound-forge-pro/

녹음 프로그램: Sound Forge



녹음 및 분석 프로그램: Praat



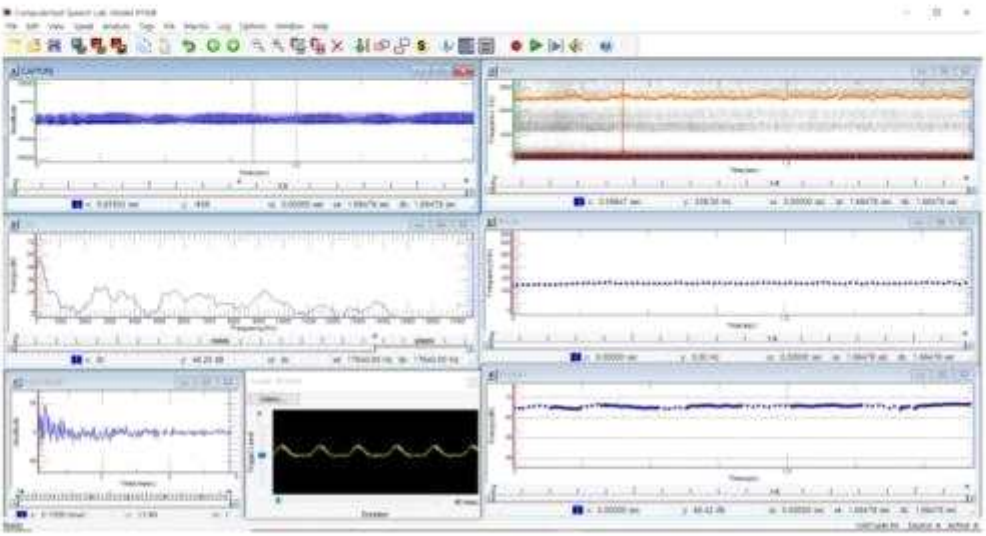
https://www.fon.hum.uva.nl/praat/

https://jkslp.org/upload/pdf/jkslp-2022-33-2-64.pdf

63

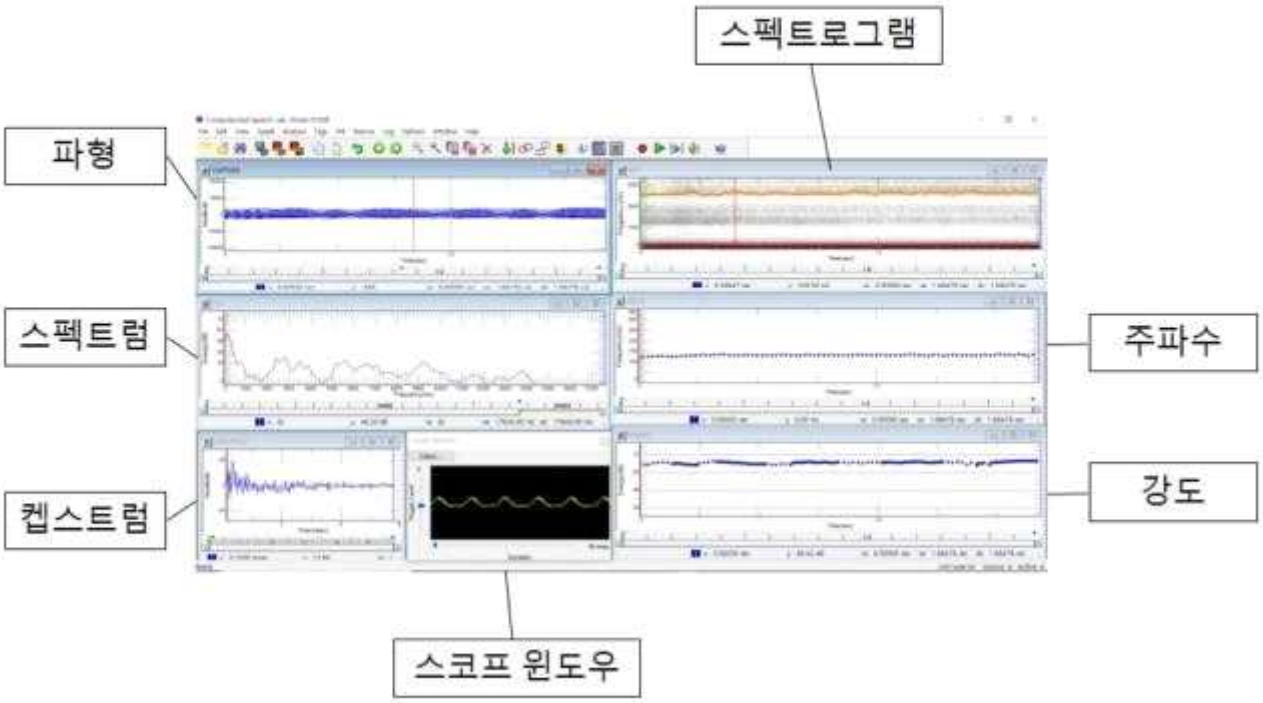
녹음 및 분석 프로그램: CSL main program

- 다양한 형태로 커스터마이즈 가능
- 녹음 제한 시간 없어 긴 발화 녹음 가능



64

녹음 및 분석 프로그램: CSL main program



녹음 및 분석 프로그램: CSL main program

Long-Term Average FFT Statistics

Domain (Hz)	0.00	to	17571.09
Range (dB)	-10.31	to	55.55
Samples	205		
Mean (dB)	11.13		
Std.Deviation (dB)	12.34		
Median (dB)	11.33		
RMS (dB)	16.59		
Spectral Mean (Hz)	237.92		
Spectral Std. Dev. (Hz)	723.72		
Skewness	10.59		
Kurtosis	128.31		

Save As... Print... Info... Done

Copy to Clipboard

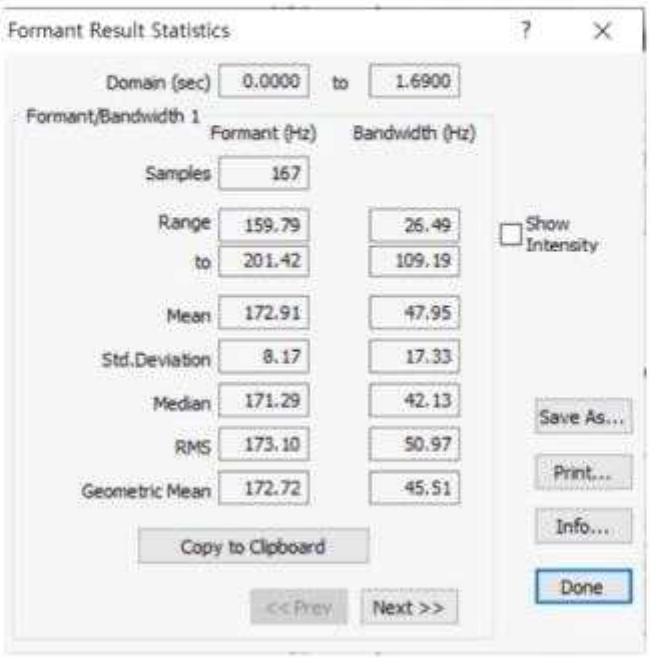
Cepstrum Result Statistics

Domain (msec)	0.0000	to	5.8050
Range (level)	-54.62	to	906.77
Samples	257		
Mean (level)	4.94		
Std.Deviation (level)	58.71		
Median (level)	0.23		
RMS (level)	58.81		

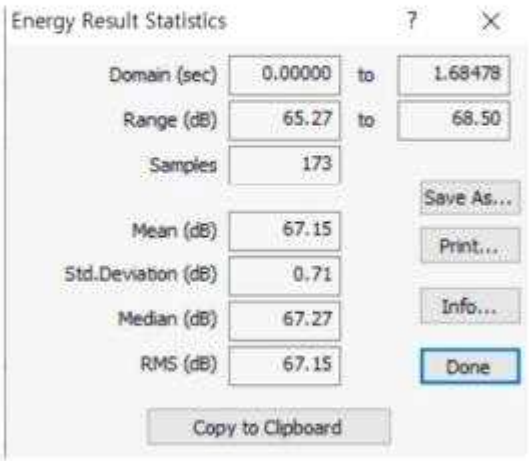
Save As... Print... Info... Done

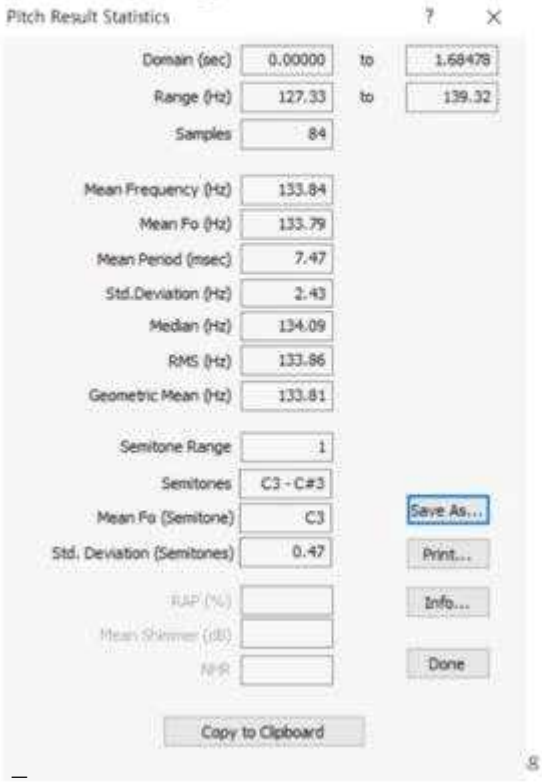
Copy to Clipboard

녹음 및 분석 프로그램: CSL main program



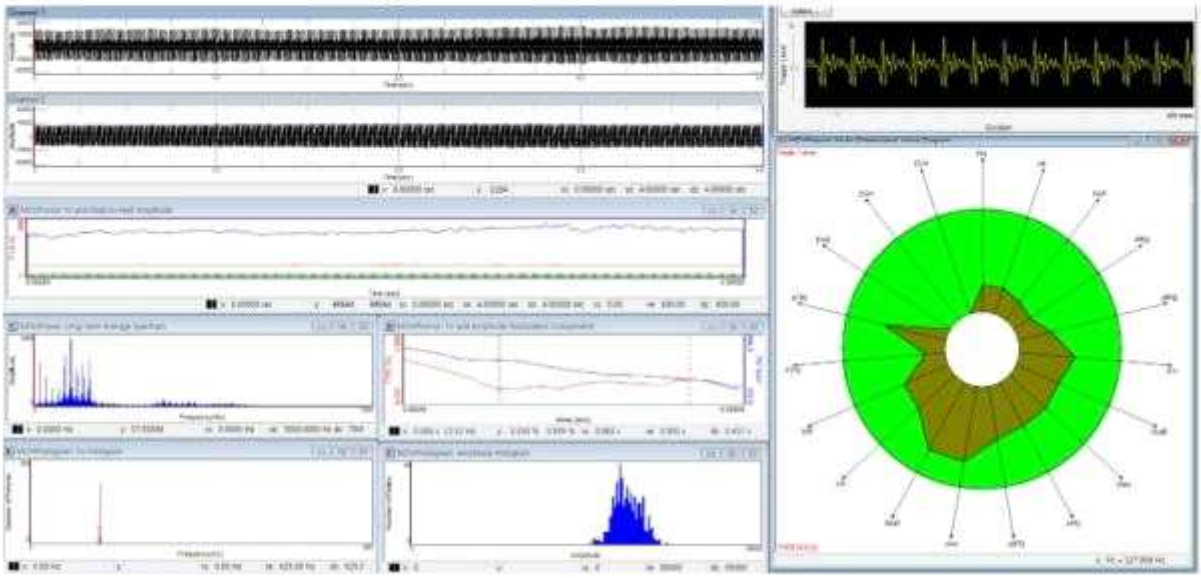
녹음 및 분석 프로그램: CSL main program



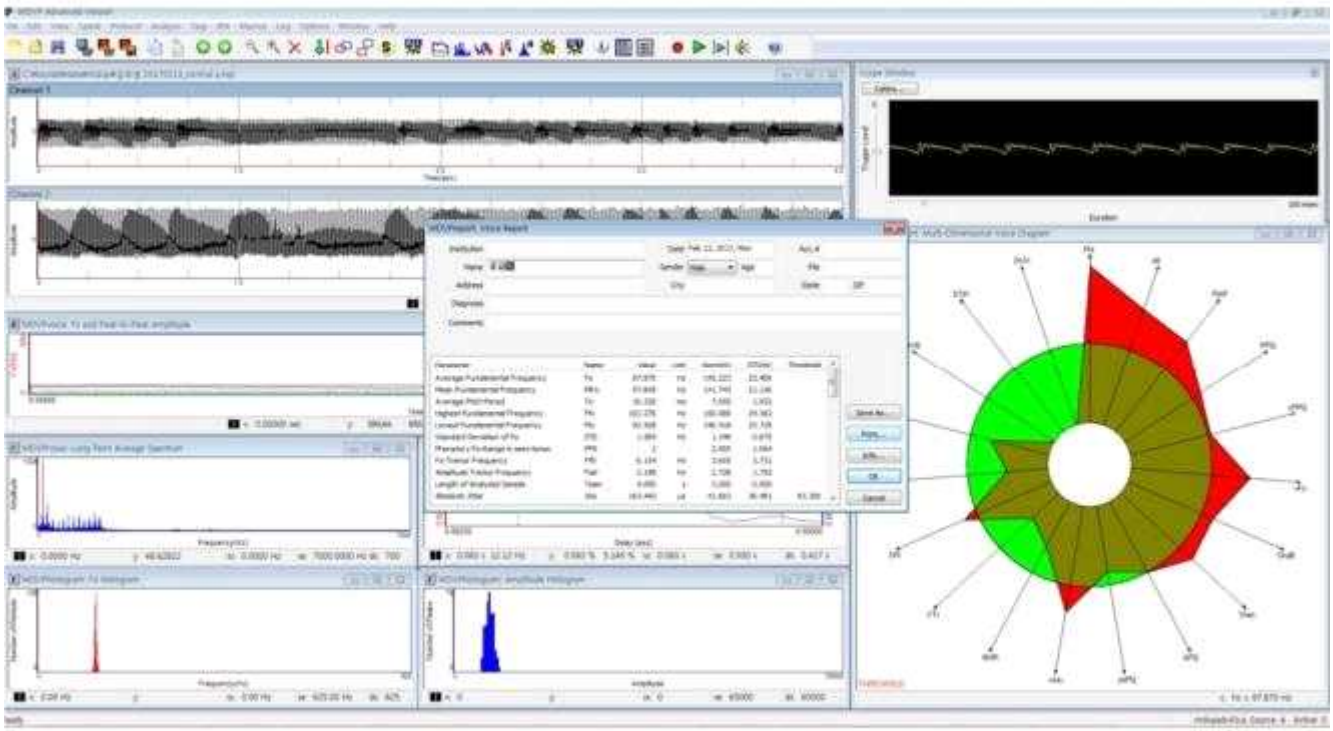


녹음 및 분석 프로그램: MDVP

- Multi-Dimensional Voice Program
- Model 5105, KayPentax, NJ, USA



69



70

All rights reserved @ SunRiver

ESC 녹음 중지	F1 매뉴얼	F2 창 비우기	F3 전체 재생	F4 선택 재생	F5 파일 열기	F6 다른 이름 저장	F7	F8	F9	F10	F11	F12 녹음 시작
`	1	2	3	4	5	6	7	8	9	0	-	=	←
Tab	Q	W	E	R ⓐ결과	T ⓐ자동 구간 설정	Y	U	I	O	P ⓒ화면 출력	[	]	₩
Caps Lock	A	S ⓒ저장	D	F	G	H	J	K	L	;	'	Enter 녹음중지	
Shift	Z	X	C	V	B	N	M	,	.	/	Shift		
Ctrl		Alt	SPACE 녹음중지							한/영	한자		

ⓒ Ctrl ⓢ Shift ⓐ Alt

한림대학교 HALLYM UNIVERSITY

녹음 및 분석 프로그램: MDVP

Figure 7-19: The different types of parameters are grouped together on the radial graph.

출처: KayPENTAX(2008). MDVP instruction manual. A Division of PENTAX Medical Company 2 Bridgewater Lane Lincoln Park, NJ.

녹음 및 분석 프로그램: MDVP

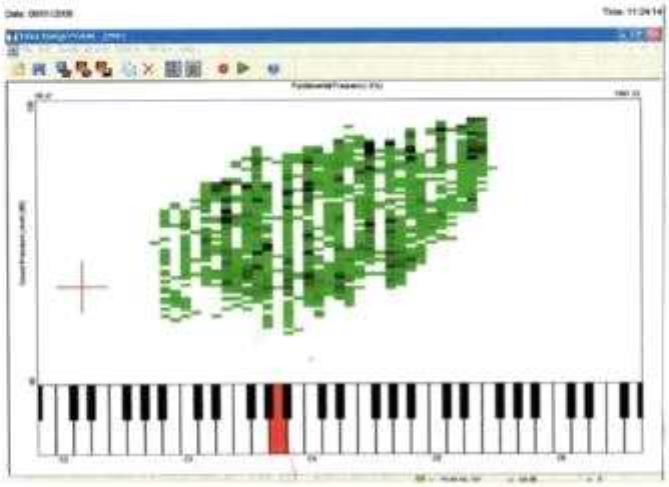
- Jitter percent(지터퍼센트, 주파수변동률)
  - 1.040%
- Shimmer percent(쉬머퍼센트, 진폭변동률)
  - 3.810%
- NHR(noise to harmonic ratio, 소음대배음 비율)
  - 0.190

- 분석이 안되거나 신뢰성이 문제가 될 수 있는 경우
  - Signal amplitude가 너무 작은 경우
  - 주기성이 너무 손상된 경우
  - 녹음한 프로그램과의 호환성이 문제시되는 경우

73

녹음 및 분석 프로그램: VRP

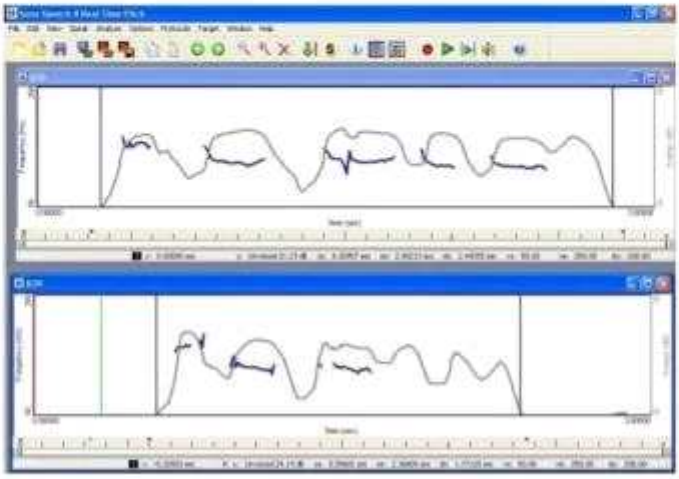
- Voice Range Profile(VRP), KayPentax, USA
- 원리
  - 산출 가능한 음역대와 각 음역에서의 강도 분포를 측정하여 phonetogram을 작성



74

녹음 및 분석 프로그램: RTP

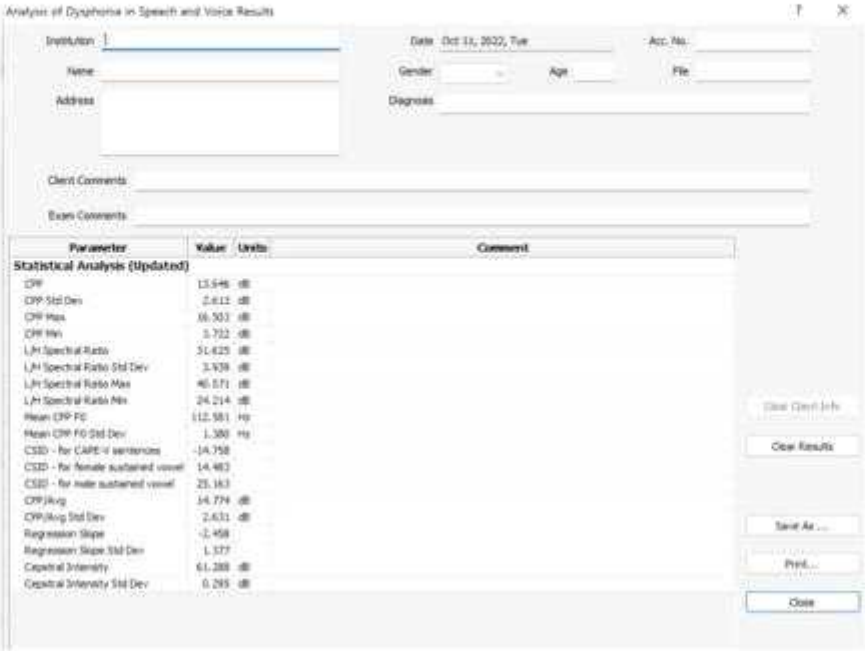
- Real-Time Pitch, KayPENTAX
  - 시각적 및 청각적 피드백
  - 문단읽기 샘플의 음도 및 음량 분석



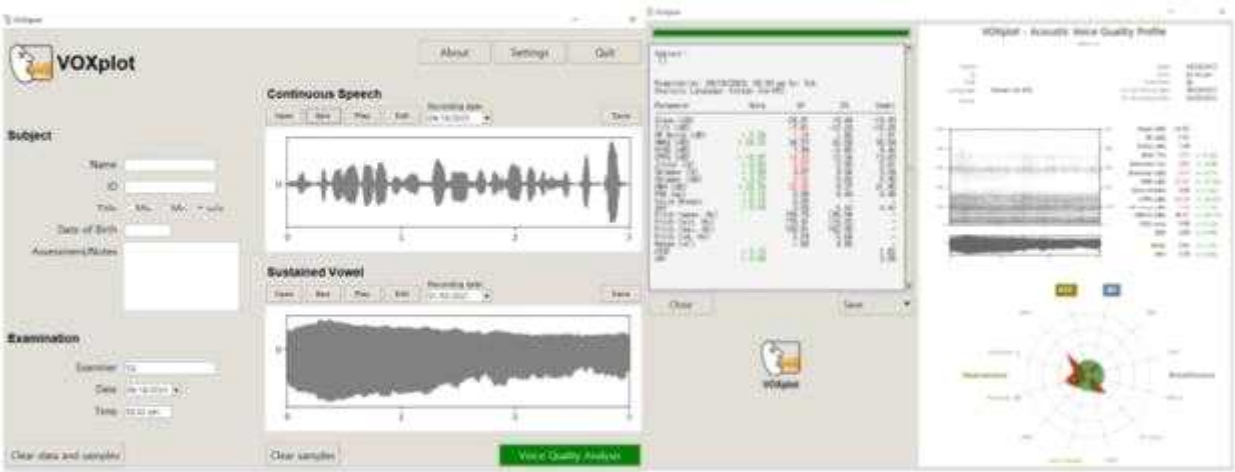
녹음 및 분석 프로그램: ADSV

- Analysis of Dysphonia in Speech and Voice
- Model 5109, KayPentax, NJ, USA
- 켑스트럼 분석 프로그램
- 모음, 말자료 (영어) 를 위한 프로토콜 제공

녹음 및 분석 프로그램: ADSV



분석 프로그램: VOXplot

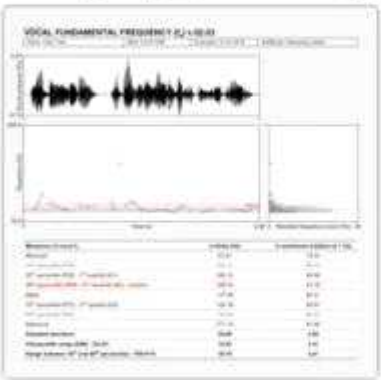


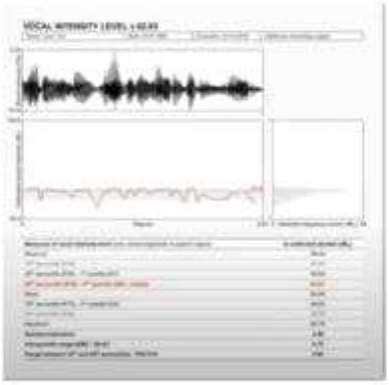
분석 프로그램: VOXplot

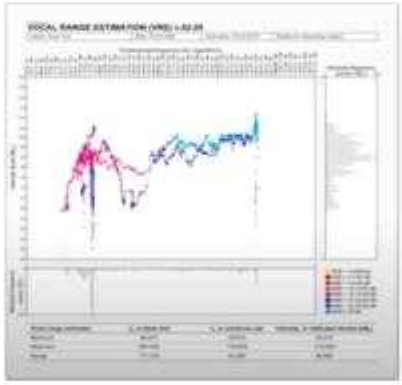
Parameter	Norm	SV	CS	Combi
Slope (dB)		-24.81	-8.44	-19.43
Tilt (dB)		-7.61	-12.24	-10.26
HF Noise (dB)	> 2.28	1.94	1.86	1.91
HNRd (dB)	> 24.23	34.57	20.88	30.94
H1H2 (dB)		7.49	-2.85	4.42
CPPS (dB)	> 14.47	14.20	12.34	13.83
Jitter (%)	< 0.50	0.67	2.90	0.91
Shimmer (%)	< 3.08	3.07	9.94	3.39
Shimmer (dB)	< 0.27	0.32	0.82	0.35
HNR (dB)	> 23.34	22.82	9.37	21.45
PSD (ms)	< 0.12	0.08	0.40	0.48
Voice Breaks	< 0.16	0.00	-	-
GNE	> 0.89	0.89	0.93	0.92
Pitch (mean, Hz)		180.75	136.54	-
Pitch (min, Hz)		175.91	115.20	-
Pitch (max, Hz)		193.81	152.48	-
Pitch (sd, Hz)		2.29	5.64	-
Range (st)		1.68	4.85	-
AVQI	< 3.15			1.82
ABI	< 3.69			3.28

Phonanium

- Praat 스트립트 패키지
- 온라인 코스 제공

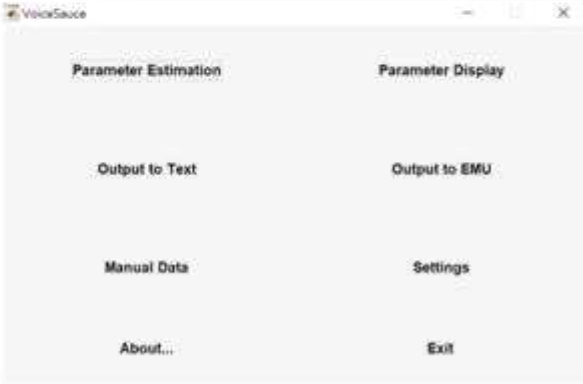






https://www.phonanium.com/about-us/

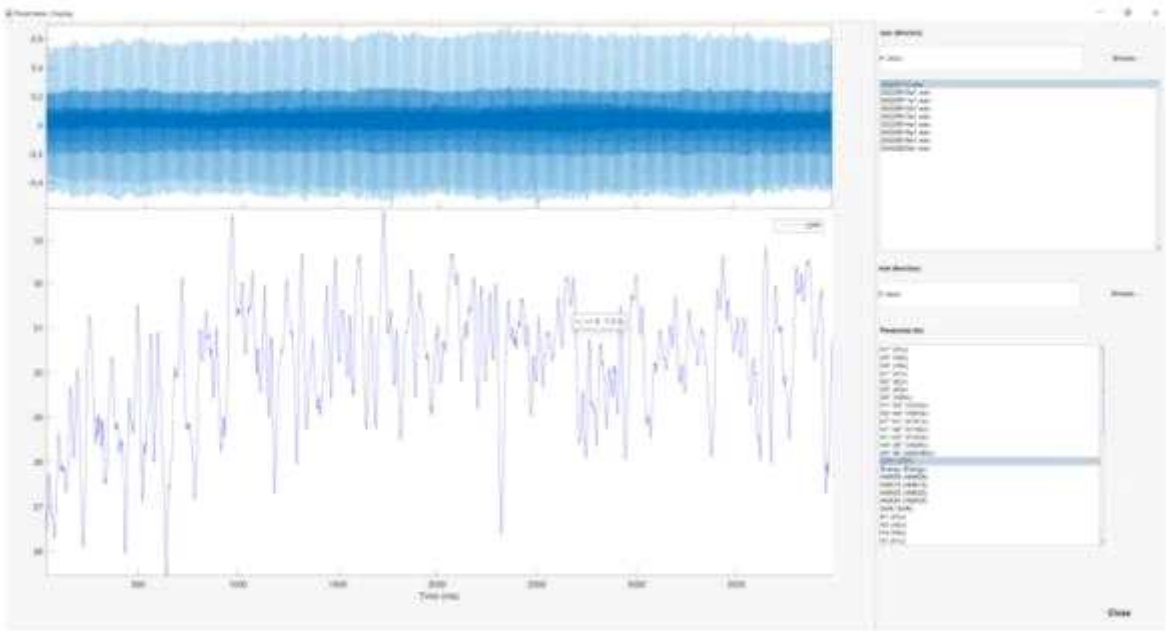
분석 프로그램: Voice Sauce



http://www.phonetics.ucla.edu/voicesauce/

81

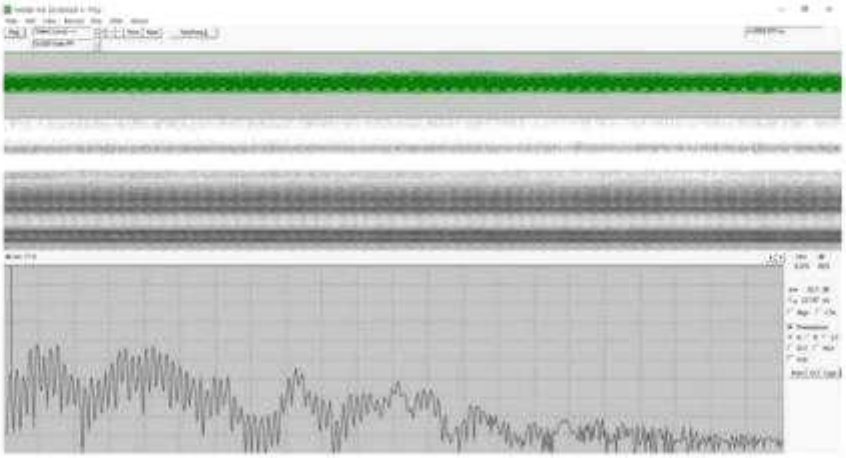
분석 프로그램: Voice Sauce



82

분석 프로그램: TF32

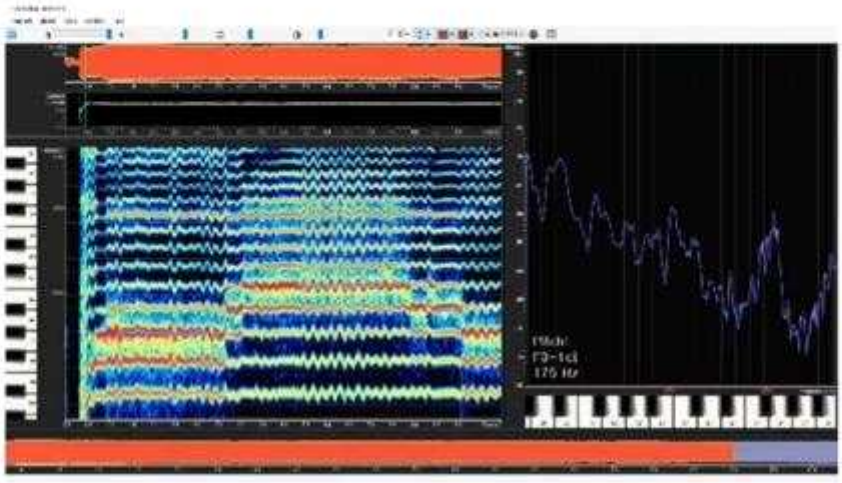
- Milenkovic P: TF32. Madison, University of Wisconsin-Madison, 2005.
- 스펙트럼 에너지 분석에 용이



83

분석 프로그램: VoceVistaVideo Pro

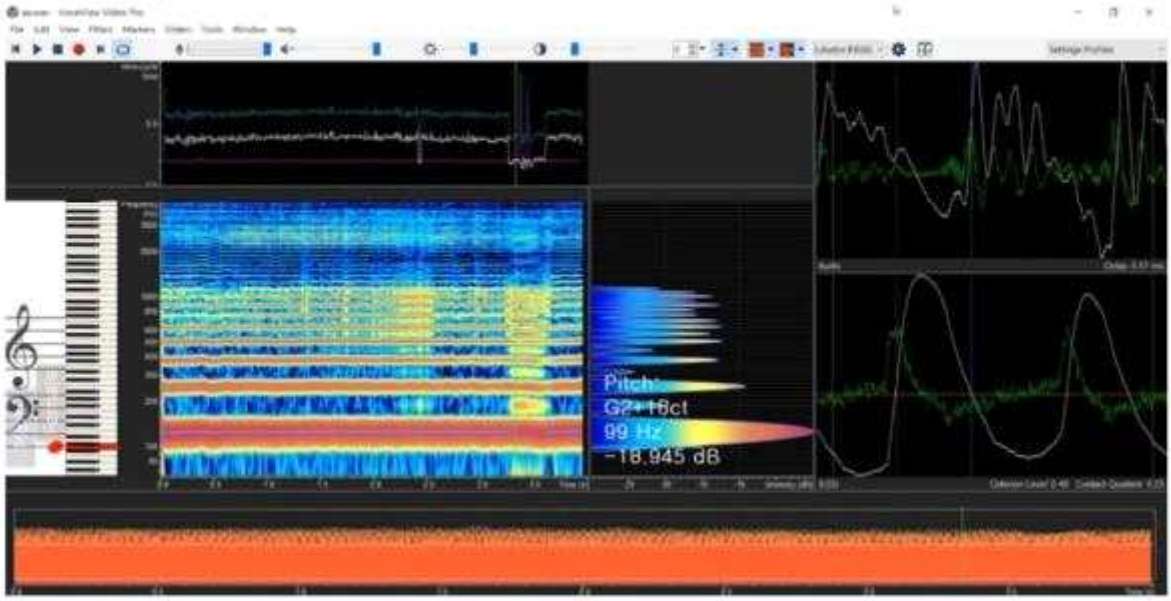
- 시각적 분석에 용이
  - Khuumii (overtone singing)



84

분석 프로그램: VoceVistaVideo Pro

- Diplophonia



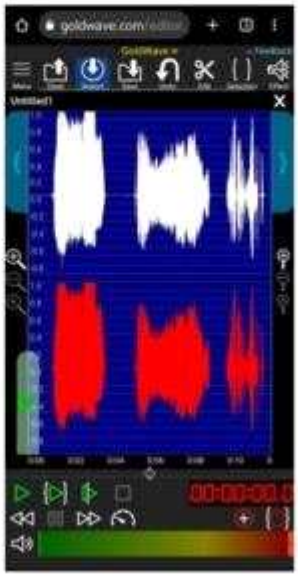
85

녹음 앱: Android & web apps

Smart recorder

Earbuds voice recorder

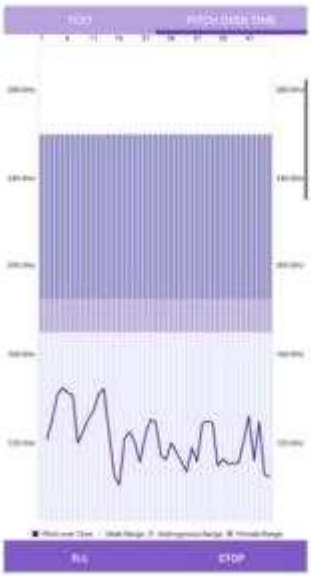
Goldwave(web app)



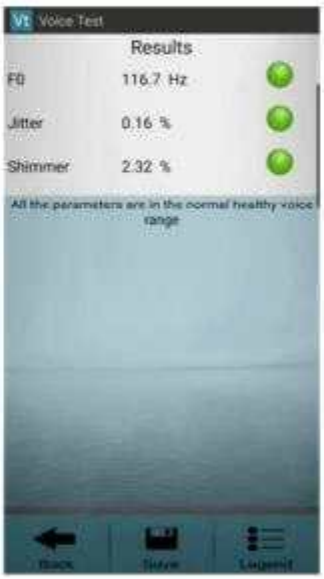
86

분석 앱

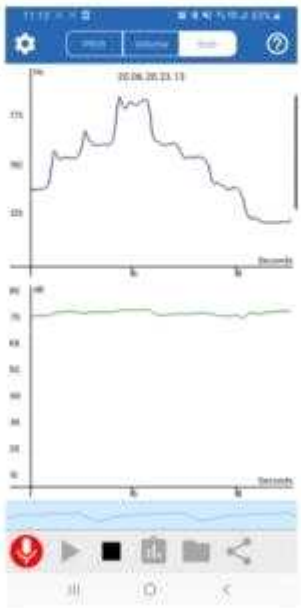
Vocal pitch monitor



Voice test (not available)



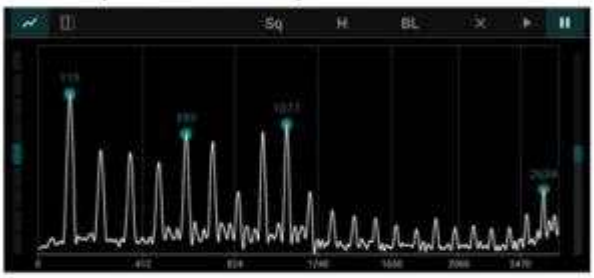
Voice Analyst



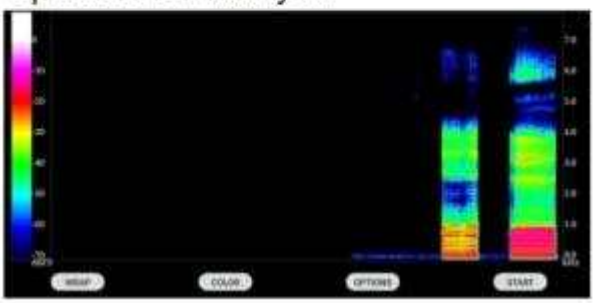
87

Android apps

FFT Spectrum Analyzer



SpectralView Analyzer



Voice Tools

88

한림대학교
HALLYM UNIVERSITY

대상자에 맞는 녹음 및 분석 프로그램 선정 및 시행

- 어떻게 이상 증상을 잘 파악할 수 있을까?
  - 다양한 음성장애의 특성에 대한 배경지식 필수
- 그냥 전부 다 시행하면 되는 것 아닐까?
  - 임상적 인상을 토대로 한 집중적 분석이 필요
- 근거기반중재에는 어떤 도움이 될까?
  - 중재 성과를 객관적으로 검증 가능

"적절한 과제와 도구를 선정하여
시행할 수 있어야"

89

"THANK YOU
FOR
ATTENTION!"

HALLYM LAB
FOR VOICE &
COMMUNICATION DISORDERS
한림 음성 의사소통장애 연구실

https://www.hvcd.kr

일정표로 이동

특강 3

▌음성 딥페이크 탐지 현황

유하진(서울시립대)

음성 딥페이크 탐지 현황

서울시립대학교
컴퓨터과학부
유하진

2024.10.19

Contents

- **음성 딥페이크 탐지 시스템 소개**
- **최신연구 방법 소개**
  - AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks
  - Wav2vec2.0 front-end + AASIST
- **음성 딥페이크 탐지 시스템의 최신 연구 방향**

음성 딥페이크 탐지 시스템의 개요 및 필요성 – 현황

- **음성 생성 기술의 발전 속도 증가**
  - 음성 생성 기술은 음성 합성(text-to-speech, TTS) 또는 음성 변조 (voice conversion)를 사용
    - 음성 생성 기술의 각 분야 성능은 인간의 MOS(mean opinion score)값과 유사한 단계까지 발전
    - 음성 생성 기술의 각 분야 논문 수는 10년 전에 비해 **483%**, **404%** 증가

그림 1. 음성 합성 및 변조 기술에 대한 연도별 논문 수 및 MOS 값 비교[1)2)]

1) 'Papers With Code', (accessed 2024. 05. 21), https://paperswithcode.com/sota
2) 'Dimensions AI', (accessed 2024. 05. 21), https://app.dimensions.ai/discover/publication

음성 딥페이크 탐지 시스템의 개요 및 필요성 – 현황

- **특정 대상 음성을 쉽게 생성 가능**
  - Voice cloning TTS 또는 zero-shot voice conversion 기술의 등장으로 특정 대상의 음성 쉽게 생성 가능
  - 쉽게 접근할 수 있는 고도화된 음성 생성 서비스들이 상용화 됨
    - (**타입 캐스트**) 감정·톤 조절이 가능한 수준의 고도화된 서비스 개시
    - (**네이버 클로바**) 가족의 목소리를 AI 보이스로 제작해 주는 서비스 개시

그림 2. 타입캐스트 음성 합성 및 변조 기술 서비스[3)]

그림 2. 네이버 클로바 음성 합성 기술 서비스[4)]

||

ElevenLabs

3) 'Typecast', (accessed 2024. 05. 22), https://typecast.ai/kr/text-to-speech
4) 'Naver CLOVA', (accessed 2024. 05. 22), https://clovadubbing.naver.com/campaign/for_family

음성 딥페이크 탐지 시스템의 개요 및 필요성

- **페이크보이스로 인한 사회적·경제적 문제 발생**
  - 사회적 문제: 가짜 뉴스 및 언론 보도
    - 유명인의 음성을 왜곡해 잘못된 정보를 보도하여 사회적 혼란을 야기
  - 경제적 문제: 보이스 피싱
    - 가족이나 지인 또는 검사 등의 목소리를 위조하여 재산을 갈취하는 범죄

그림 4. 페이크보이스를 악용한 가짜 뉴스[5]

그림 5. 페이크보이스를 악용한 보이스피싱 범죄[6]

5) 최종혁, "우크라 대통령, 돈바스 반환 결정"? 가짜 영상 논가..., JTBC 뉴스룸, 2022. 3. 17. 수정, (2024. 05. 21. 접속), https://www.youtube.com/watch?v=FdojnAQE92U
6) 김진일, "분명 내 목소리인데"...AI 딥보이스 범죄에 주의보, 연합뉴스TV, 2024. 5. 15. 수정, (2024. 05. 21. 접속), https://www.youtube.com/watch?v=vbD-E_Lwnyg

음성 딥페이크 탐지 시스템의 연구 현황

- **학계·산업계에서 다양한 페이크보이스 탐지 대회 및 서비스로 해당 연구의 관심이 높아지고 있음**

- ASVspoof (Automatic Speaker Verification and Spoofing Countermeasures Challenge): 화자 인식 시스템의 보안을 향상 시키기 위한 페이크보이스 탐지 대회
- ADD (Audio Deepfake Detection Challenge): 페이크보이스 공격을 탐지하고 대응책을 연구하기 위한 대회
- SASV challenge 2022: 페이크보이스 시스템과 화자 인증 시스템을 통합하여 연구하기 위한 탐지 대회

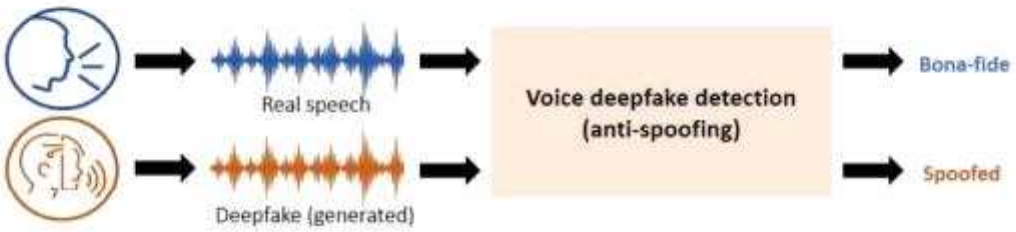
Deepfake Challenges

- **ASVspoof (2015, 2017, 2019, 2021, 2024)**
  Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan

- **ASVspoof consortium** http://www.asvspoof.org
- **Tasks**
  - A logical access (LA) task
    - text-to-speech (TTS) and voice conversion (VC) attacks
    - coding and transmission of TTS and VC
  - A physical access (PA) task
    - real bona fide and replayed samples
  - A new speech deepfake task (DF)
    - compressed audio similar to the LA task (but without ASV).

음성 딥페이크 탐지

- **정의**
  - 입력된 음성이 인위적으로 합성 또는 변조되었는지 혹은 실제 발화되었는지 진위를 판단하는 이진 분류 작업
  - 안티스푸핑 (anti-spoofing)으로도 알려짐



- **음성 딥페이크 탐지 연구의 필요성**
  - 음성 딥페이크로 인한 범죄 예방 / 화자 인식 시스템의 보안성 향상
  - 음성 생성 기술의 발전으로 인해 탐지 성능 고도화 요구
    - 생성된 음성을 사람이 구분하기 어려운 정도로 발전
      * 실제 발성의 MOS와 생성된 발성의 MOS 유사
      * MOS (mean opinion score): 청취자들이 음성 샘플을 1~5점으로 평가한 결과의 평균값

음성 딥페이크 탐지 시스템 구조

- **음향 특징 추출기 (audio feature extractor)**
  - 입력된 원시 파형을 프레임 별로 가공하여 음향 특징 지도 (audio feature map) 출력
  - 예시: Mel-spectrogram, LFCC (linear-frequency cepstrum coefficient)
- **딥페이크 특징 추출기 (deepfake feature extractor)**
  - 음향 특징 지도를 가공하여 스푸핑 증거 정보가 내포된 임베딩 출력
    * 스푸핑 증거 (spoofing evidence): 음성의 위・변조 여부를 탐지할 수 있는 정보 (예시: 비정상적인 강조, 억양, 과도한 스무딩)
  - 대표적인 딥페이크 특징 추출기: AASIST, Conformer, LCNN-LSTM, Rawformer
- **분류기 (classifier)**
  - 임베딩을 입력 받아 이진 분류를 통해 발성의 진위 여부 출력
  - Bona-fide: 실제 발성 / Spoofed: 위・변조된 발성 (가짜)

성능 평가 지표

- **동일 오류율 (equal error rate, EER)**
  - 오인식률과 오거부율이 동일해지는 임계값에서의 오류율
  - 낮을수록 정확도가 높은 시스템

성능 평가 척도 (2)

minDCF(minimum detection cost function)

- ❖ NIST Speaker Recognition Evaluation에서 권장하는 성능 평가 척도 계산 방법
- ❖ Miss에 대한 비용을 10, false alarm에 대한 비용을 1로 두고,
  검출 대상의 비율이 1/100이라고 가정하여 비용 계산(NIST08 SRE 기준)

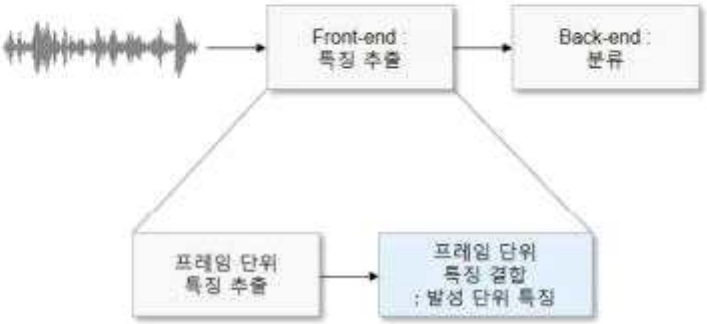
C_{Miss}	$C_{FalseAlarm}$	P_{Target}
10	1	0.01

- ❖ 다음 탐지 비용을 최소화하는 threshold하에서 비용 측정

$$C_{Det} = C_{Miss} \times P_{Miss|Target} \times P_{Target} + C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \times (1-P_{Target})$$

딥러닝을 이용한 딥보이스 검출의 주요 요소

- 입력
- 특징 결합
- 네트워크 구조
- 목적 함수



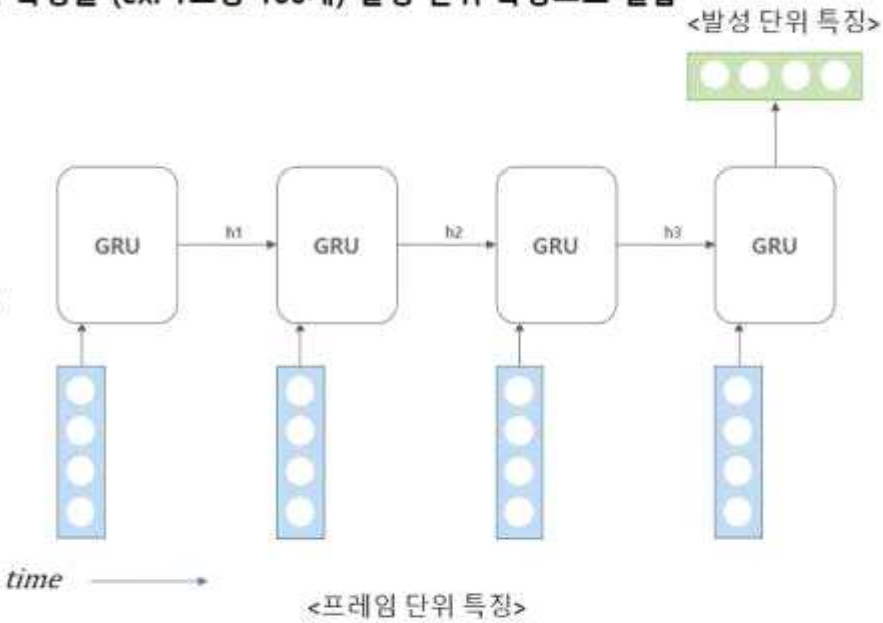
입력 및 음향 특징

- MFCC
- Spectrogram
- Raw waveform
- Spectral and temporal features

특징 결합

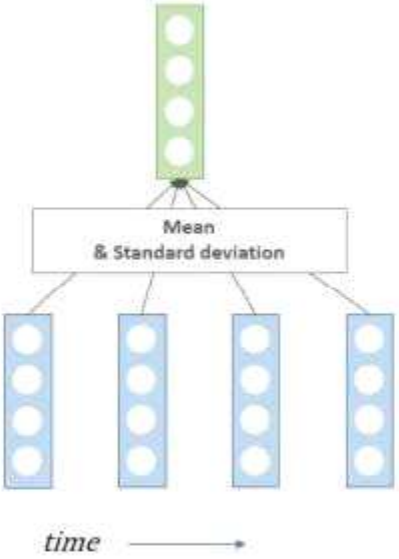
- 입력 발성에 대해 추출된 복수의 프레임 특징을 (ex: 1초당 100개) 발성 단위 특징으로 결합
- 시간 순서에 따른 정보 결합
  - 자연어 처리 과정과 유사함 (frame → 단어)
  - 음성 인식에서 성공한 모델을 적용
- Gated recurrent unit (GRU)
  Long short-term memory (LSTM)
  Learnable dictionary encoding (LDE)

frames



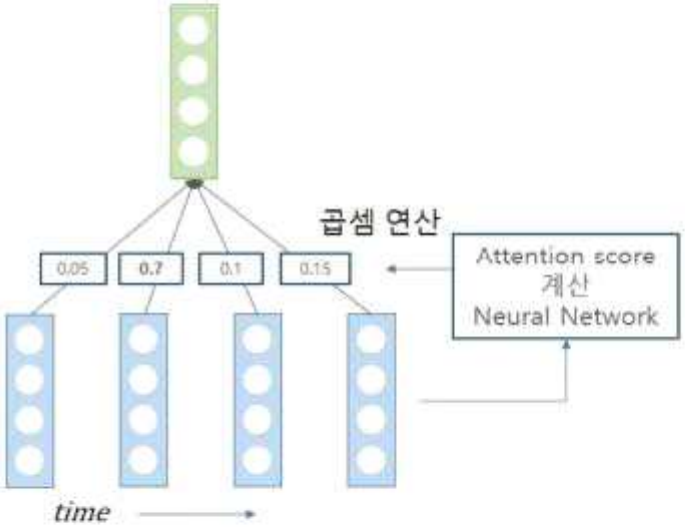
특징 결합 - Statistics pooling (SP)

- **통계치를 활용해 전반적인 특성 결합**



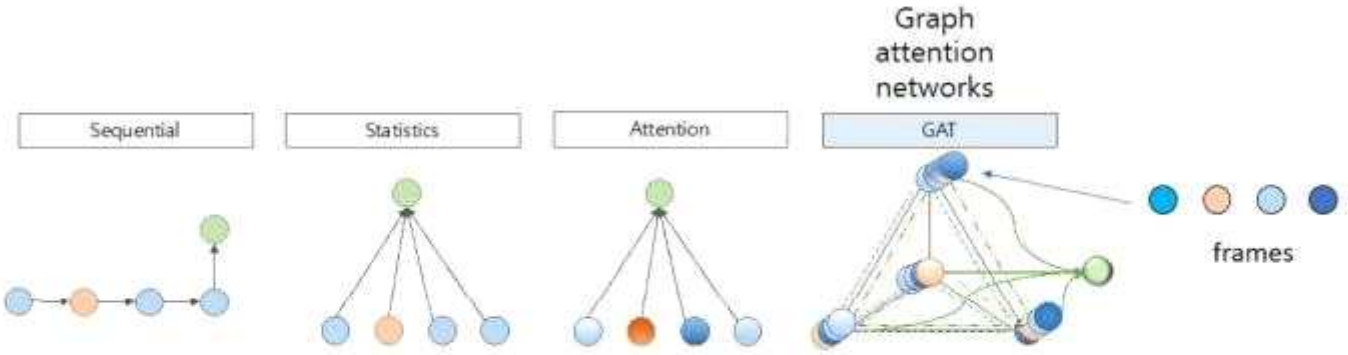
특징 결합 - Self-attentive pooling (SAP)

- **Attention 기법을 활용해 중요한 정보를 집중적으로 결합**



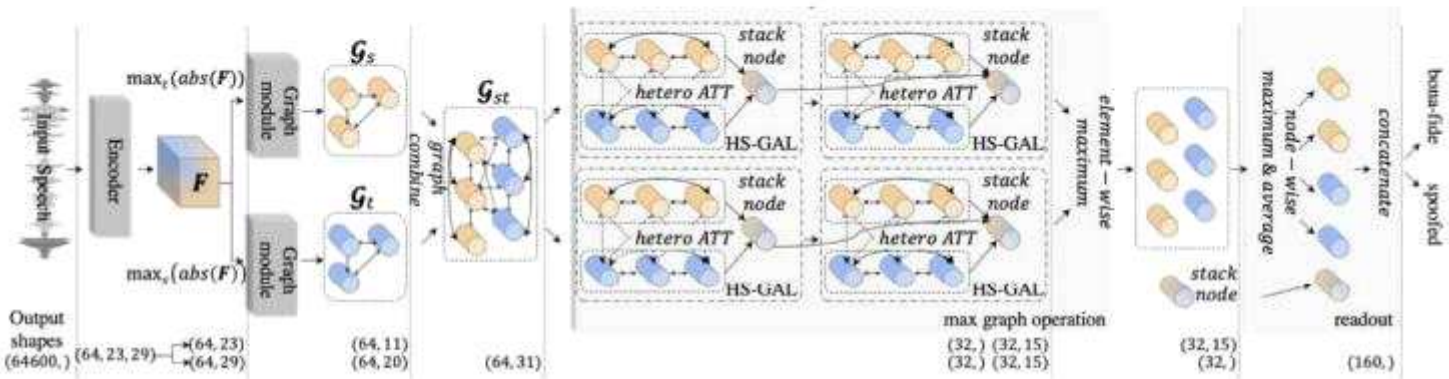
특징 결합 - Graph 활용

- **특징 간의 연관성을 명시적으로 고려 → Graph 활용**



	Sequential (GRU, LDE)	Statistics	Attention	GAT
특징의 순서가 성능에 미치는 영향	△	X	X	X
특징 간의 비교	X	X	△	O

AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks



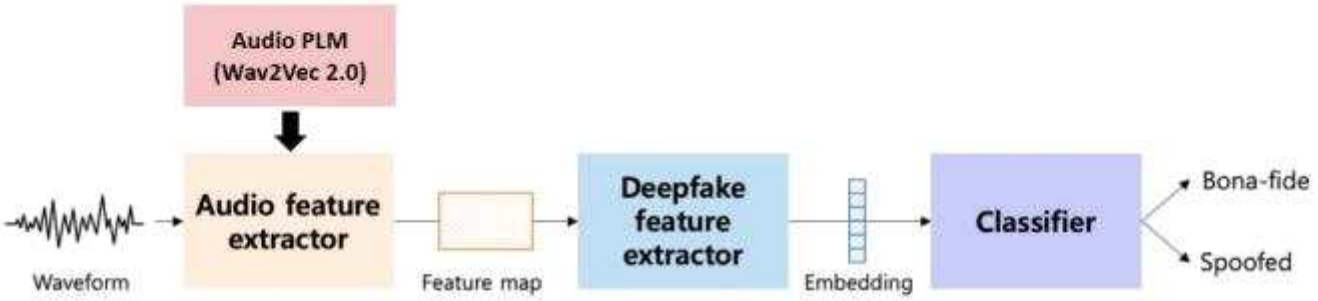
Spectral and temporal features

Heterogeneous stacking graph attention layer

J Jung, HS Heo, H Tak, H Shim, JS Chung, BJ Lee, HJ Yu, N Evans,
"AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," ICASSP 2022

사전 학습된 거대 모델 (pre-trained large model, PLM)의 활용

- **Audio PLM**
  - 자기 지도 학습 (self-supervised learning)을 통해 대규모 데이터로부터 음향 특징의 다양한 패턴과 분포 학습
  - 다양한 음향 신호 처리 분야에서 우수한 성능 달성에 기여
- **음성 딥페이크 탐지 분야에서의 audio PLM의 사용**
  - Audio PLM을 음향 특징 추출기로 활용
  - H. Tak et al[1]: Wav2Vec 2.0 + AASIST
  - E. Rosello et al[2]: Wav2Vec 2.0 + Conformer



1. H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," Proc. Odyssey 2022, 2022.
2. E. Rosello, "Casado, A. Go "mez Alan" ıs, A. ", M. Go "mez Garc "ıa, A. M. Peinado Herreros et al., "A conformer-based classifier for variable-length utterance processing in anti-spoofing," Proc. Interspeech, 2023.

대형 자기 지도 사전 학습 (Large-Scale Self-Supervised Pre-training)을 통한 대형 언어 모델 학습

학습 데이터 문장

"한국음성학회는 90년의 역사를 가지고 있습니다 "

한국음성학회	는	90	년	의	( )	를	가지고	있습니다

Neural Network (Transformer) → 언어 지식 학습

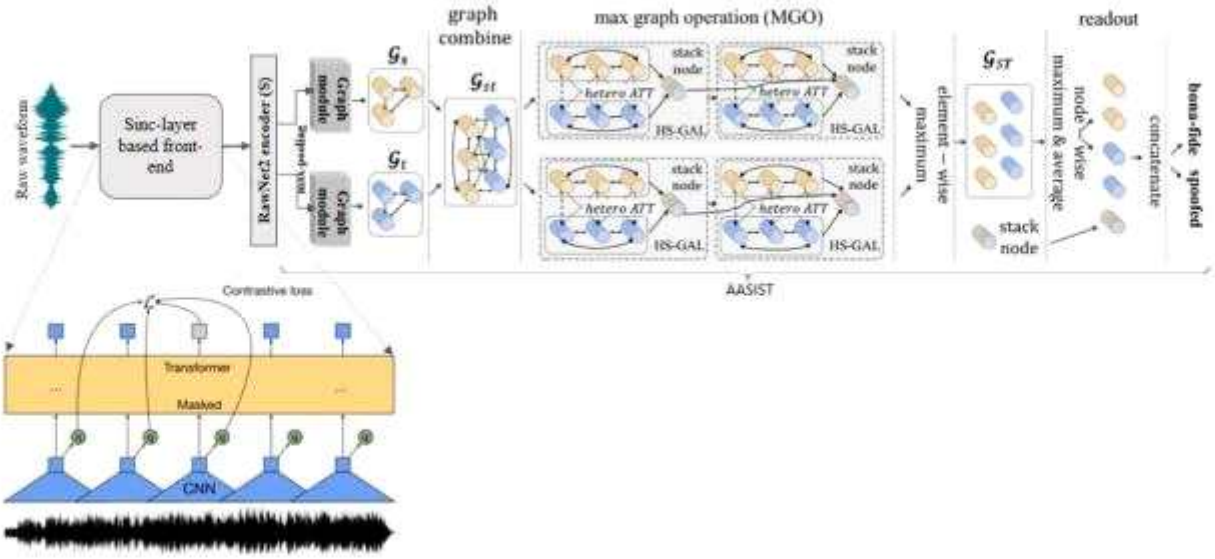
"역사"

- 텍스트 대신 음성 사용
  → 음향 지식 학습

예:
Wav2Vec 2.0
HuBERT
WavLM

Wav2vec2.0 front-end + AASIST

- Use pre-trained self-supervised models trained on large diverse datasets



Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi and Nicholas Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation", Odyssey, 2022.

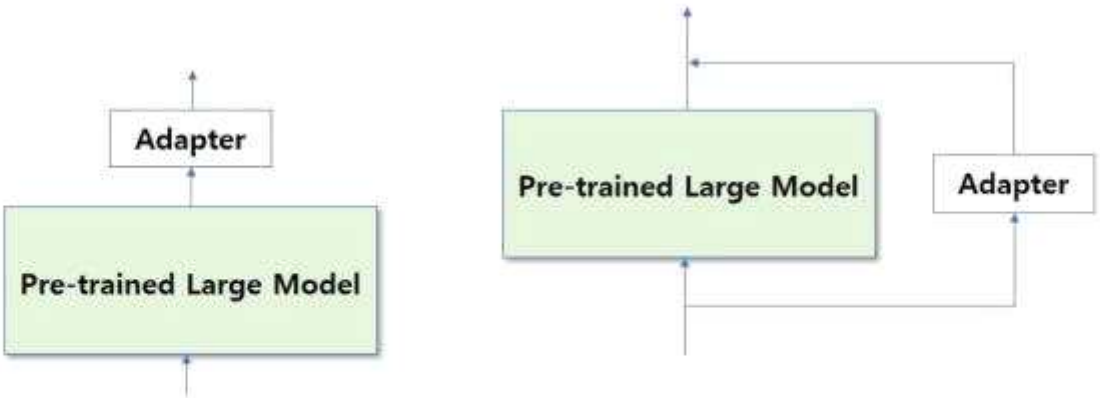
최근 연구 트렌드 분석 (INTERSPEECH 2024)

Data augmentation

- **동일 화자의 발성, 동일 공격 유형의 발성을 결합하여 데이터를 증강**
- **속도 증강**
  - 0.9배속 / 1배속(원본) / 1.1배속 ...
- **Codec 증강**
  - mp3, ogg, acc, m4a, ...
- **Large-scale vocoded data**
  - 합성기의 최종 단계에 해당하는 여러가지 vocoder를 이용하여 음성 생성
- **Differently-trained attack models**
  - 예: 합성기에서 파라미터를 여러가지로 변경하여 음성 생성

Adapter Learning from Pre-trained Large Model

- **기존: 사전 학습된 대형 모델을 전체적으로 조정 학습하는 Full-tuning 방법**
  - 학습하는 시간이 많이 걸림
  - 과적합(overfitting) 위험: 학습 데이터에만 강하게 편향됨
  - Catastrophic forgetting 문제 : 새로운 것을 학습하면 이전에 학습한 것을 잊어버림
- **Adapter 라고 부르는 적은 파라미터 추가에도 Full-tuning 보다 뛰어난 일반화 성능 개선 달성**



One-class learning

- **기존 연구의 문제점**
  - 새로운 공격 기법(합성, 변조 등)에 대해서는 오류가 크게 증가함
- **One-class learning**
  - 공격 기법에 관계없이 Bonafide의 분포만을 모델링
  - 기존에는 Bonafide/Spoofed 특성을 학습
  - bonafide 샘플로부터 이상치 탐지 개념으로 접근
    - 데이터 수집과 예측이 어려운 spoofing 샘플보다 유리

정리: 연구 트렌드 분석

- **Self-supervised pre-trained large 모델을 활용하는 연구 증가**
  - Full-tuning을 수행하지 않아도 우수한 성능 달성 가능
- **데이터 증강 기법에 대한 연구 증가**
  - 2024년 ASVspoof5의 대다수 결과가 데이터 증강을 통해 우수한 성능 달성
- **Bonafide 샘플에 집중하여 일반화 성능 향상하는 연구 증가**
  - 다수의 bonafide 샘플 활용 가능
  - bonafide 샘플로부터 이상치 탐지 개념으로 접근
- **Adversarial attack**
  - 사람은 구분할 수 있으나, 인공지능은 구분하지 못하게 하는 공격에 대한 방어

제안: 음성학적 지식을 이용한 음성 딥페이크 탐지

- **신경회로망(딥러닝)을 이용한 탐지 방법은 설명 불가**
  - 탐지 결정 결과에 대한 분석이 어려움.
  - 인간의 인지 방법과 다름.

- **Formant frequency**
- **Acoustic parameters**
  - Pitch, HNR (Harmonic to Noise Ratio), Jitter, Shimmer, Amplitude perturbation quotient (APQ) ?
- **Vocal-tract length?**
- **Self-supervised pre-trained large 모델의 출력 결과 분석**

감사합니다.

Special Thanks to: 신현서, 허정우 학생

일정표로 이동

2024 한국음성학회 가을 학술대회

구두 발표 논문

일정표로 이동

국립국어원의 “일상대화 말뭉치” 설계 및 구축

이용주, 차원철(나라지식정보), 이광수(스마트미디어테크),
강미영, 서샛별(국립국어원)
yjlee@wku.ac.kr

Design and construction of the National Institute of Korean Language’s “Daily Conversation Corpus”

Yongju Lee, Wonchul Cha (Nara Knowledge Information)
Kwangsu Lee (Smart Media Tech.)
Miyoung Kang, Saetbyol Seo (National Institute of Korean Language)

살아있는 생생한 우리말의 변화를 반영하면서 인공지능 학습데이터로도 쓰일 수 있는 “ 일상 대화 말뭉치” 가 국립국어원의 주관으로 2019년부터 현재까지 매년 구축되어 공개되고 있다.

매년 500시간 ~ 1000시간 분량의 일상대화가 2인, 3인, 4인대화의 형태로 수집되어 2023년까지 누적 3630시간분이 구축되어 공개 중이며 2024년에는 700시간분이 구축 중에 있다.

올해는 약 40시간분의 1인발화(독백)도 포함되어 있는데, 공적 독백과 사적 독백으로 나누어 시험적으로 수집되고 있는 중이다, 사적 독백은 주어진 다양한 주제를 대상으로 화자가 혼자 자유롭게 발화한 데이터이고 공적 독백은 예를 들면 유튜브 게임 체널처럼 한사람이 발화하지만 그 대상이 다수인 경우를 의미한다. 다인 대화의 경우는 여행, 방송/연예, 정치, 경제, 건강, 스포츠/레저, 연애/결혼, 음식, 문화, 교통, 학교/회사, 선물, 아르바이트, 가족, 성격 등 다양한 주제를 대상으로 한다. 한 화자 당 발화시간은 합하여 30분을 넘지 않는 범위에서 여러 세션에 참여 가능하다. 성별, 나이, 출신 지역별로 인구통계학적으로 균형을 맞춘 화자들이 참여하고, 16Khz 샘플링과 16비트로 양자화된 디지털 음성자료를 바탕으로 발음 및 철자로 2층의 전사가 이루어졌으며 억양구 단위를 참조하면서 문장단위로 구분하여 정리되어 있다.

사용자가 상세히 참고할 수 있도록 다음과 같은 메타데이터들이 부가되어 있다.

- 말뭉치 파일 메타 정보(파일 제목, 구축 및 배포자, 구축 년도, 분류 등)
- 대화 정보(대화 제목, 녹음일자, 대화 주제 등)
- 화자 정보(연령, 성별, 직업, 출생 및 성장지, 학력 등)
- 환경 정보(화자 간 관계, 친밀도 등)
- 발화 정보(철자 전사, 발음 전사, 발화 시작 및 종료 시간 등)

전사 과정 중에 개인정보는 비식별화 처리하여 제공된다.

2019년 2023년까지 구축된 말뭉치는 아래 사이트에서 배포되고 있다.
국립국어원“ 모두의 말뭉치”
https://kli.korean.go.kr/corpus/main/requestMain.do?lang=ko

일정표로 이동

재위 기간 중 영국 엘리자베스 2세 국왕의 단순 모음 변화

정현성
한국교원대학교 영어교육과
hchung@knue.ac.kr

Monophthongal vowel changes of Queen Elizabeth II during her reign

Hyunsong Chung
Dept. of English Education, Korea National University of Education
hchung@knue.ac.kr

본 연구는 Deterding(1997)과 Harrington 등(2000)의 영국 영어 단순 모음 연구를 재현한 실험 연구이다. 본 연구에서는 엘리자베스 국왕이 권좌에 오른 1952년(26세)과 1953년(27세), 2000년(74세)과 2001년(75세), 타개하기 직전인 2020년(94세)과 2021년(95세)에 송출된 국왕 성탄절 방송에 나타난 단순 모음의 포먼트를 분석하여 재위 기간 중 모음에 어떤 변화가 있었는지 살펴보았다. 모음의 F1과 F2를 Bark로 변환한 후 1950년대(1952년, 1953년), 2000년대(2000년, 2001년), 2020년대(2020년, 2021년)의 평균 포먼트 Bark 값을 이용해 모음도를 그려 비교하고, 평균 값 간에 통계적으로 유의미한 차이가 있는지 분석하였다. 연대별로 전체 모음 F1, F2의 중심 값(centroid)에서부터 개별 모음까지의 유클리드 거리(Euclidean distance) 평균도 분석하였다.

분석 결과 모음의 전체적인 분포에서 1950년대에는 단순 모음을 발음할 때 모음의 조음 폭이 크지 않아 모음 간의 구분이 크지 않았지만, 2000년대와 2020년대에는 모음의 조음 폭이 커지면서, 모음 간의 구분도 더 명확하게 변화되었다. 통계적으로 유의미한 차이가 있는 음들을 기준으로 보면, F1의 경우 /i/와 /ɒ/, /ʊ/, /u/는 연도별로 유의미한 변화가 없었지만, 1950년대와 2000년대에는 /ɪ/, /æ/, /ɑ/, /ʌ/, /ɔ/ 모음에서 유의미한 변화가 있었고, 1950년대와 2020년대에서 모음의 변화가 관찰된 것은 /ɪ/, /æ/, /ɔ/ 모음 이었다. 2000년대와 2020년대 사이에서 모음의 변화가 관찰된 것은 /ɛ/, /æ/, /ɑ/, /ʌ/, /ɜː/ 모음이었다. F2의 경우에는 /ɛ/, /æ/, /ʊ/, /u/의 경우 유의미한 변화가 없었지만, 1950년대와 2000년대에는 /ɪ/, /ʌ/, /ɒ/, /ɔ/, /ɜː/ 모음에서 유의미한 변화가 관찰되었다. 1950년대와 2020년대 사이에서는 /i/, /ɒ/, /ɔ/ 모음에서 변화가 있었고, 2000년대와 2020년대 사이에서는 모음의 변화가 관찰되지 않았다.

유클리드 거리의 경우 /ɪ/, /ɛ/, /æ/, /ɑ/, /ʌ/, /ɒ/, /ɔ/, /ɜː/ 모음에서 유의미한 변화가 관찰되었고, /i/, /ʊ/, /u/ 모음에서는 유의미한 변화가 없었다. 1950년대와 2000년대 사이에서는 /ɪ/, /æ/, /ɑ/, /ʌ/, /ɒ/, /ɔ/, /ɜː/ 모음에서 유의미한 변화가 있었고, 1950년대와 2020년대 사이에서는 /ɑ/, /ɒ/, /ɔ/, /ɜː/

모음에서 유의미한 변화가 있었으며, 2000년대와 2020년대 사이에서도 /ɪ/, /ɛ/, /æ/, /ʌ/ 모음에서 유의미한 변화가 관찰되었다. 1950년대와 2000년대 사이에는 변화가 있는 모음 가운데 /ɜː/만 중심 값에 더 가까워 졌고, 나머지 모음은 중심 값에서 거리가 더 멀어졌다. 1950년대와 2020년 사이에서도 /ɜː/는 중심 값에 더 가까워진 반면, /ɑ/, /ɒ/, /ɔ/ 모음은 중심 값에서 거리가 더 멀어졌다. 2000년대와 2020년 사이에서는 유의미한 변화가 있는 /ɪ/, /ɛ/, /æ/, /ʌ/ 네 모음 모두 2000년대에 비해 2020년대에 유클리드 거리 평균이 중심 값에 더 가까워 졌다. 이러한 분석을 통해 엘리자베스 국왕이 재위 직후에는 모음의 조음 폭이 크지 않았지만, 75세 전후인 2000년대와, 95세 전후인 2020년대에는 모음의 조음 폭이 재위 직후에 비해 커져 있음을 알 수 있었다. 하지만, 2020년대에는 2000년대 비해 모음의 조음 폭이 다소 작아졌음을 관찰할 수 있었다. /ɪ/를 제외한 /i/, /ʊ/, /u/ 등의 고모음에는 큰 변화가 없는 것을 관찰되었다.

참고문헌

Deterding, D. (1997). The formants of monophthong vowels in standard southern British English pronunciation. *Journal of the International Phonetic Association, 27*(1/2), 47-55.

Harrington, J., Palethorpe, S., Watson, C. Monophtongal vowel changes in Received Pronunciation: An acoustic analysis of the Queen's Christmas broadcasts. *Journal of the International Phonetic Association, 30*(1/2), 63-78.

일정표로 이동

모음 음향 특징을 활용한 충남-전북 방언 구획 연구

이주영[1], 김선희[2], 정민화[1]
서울대학교 언어학과[1], 서울대학교 불어교육과[2]

Dialect Division Study on Chungcheongnam-do and Jeollabuk-do Dialects using Vowel Acoustic Features

Jooyoung Lee[1], Sunhee Kim[2], Minhwa Chung[1]
Dept. of Linguistics, Seoul National University[1]
Dept. of French Education, Seoul National University[2]
{excalibur12[1], sunhkim[2], mchung[1]}@snu.ac.kr

본 연구는 인접한 전라북도 지역과의 비교를 통해 충청남도 방언의 고유성을 음향적 관점에서 살펴보는 데 목적이 있다. 충남 방언의 구획에 관한 기존 연구는 크게 충남 방언은 독자적인 방언권이 형성되어 있다는 주장(도수희 1965)과 인접 지역 방언권인 경기 방언과 전북 방언의 영향으로 경기 및 전북 방언과의 경계가 모호하다는 주장(한영목 1999)으로 나뉜다. 충남 방언 구획을 비롯한 방언 구획 연구 사례에서는 주로 방언 어휘 정보를 이용하여 인접한 지역이 상이한 어휘를 사용할 경우 서로 다른 방언권으로 나타내었다. 하지만 텍스트 중심의 어휘를 사용한 사례가 많으며 음향 정보를 활용한 연구는 찾아보기 어렵다. 정성훈(2024)은 방언 어휘를 음소 단위에서 비교하여 의미적 유사성과 형태적 유사성을 함께 고려하였으나, 이 또한 음소로부터 음향 정보를 추출하지는 않았다. 음향 특징을 사용한 방언 분석이 중요한 이유는 어휘 텍스트로 표현하기 어려운 연속적인 값을 나타낼 수 있어 정밀한 방언 구획이 가능해지기 때문이다. 음향 특징을 활용한 연구뿐만 아니라 방언 자유발화로 연구한 사례 또한 찾아보기 어렵다. 자료제공인과의 인터뷰로부터 얻은 방언 어휘는 기록하여 사용할 수 있으나 자유발화 자체를 활용하기에는 텍스트를 전사하는 것이 어려울 뿐만 아니라 확인하고자 하는 어휘나 문법을 찾는 시간적 비용 또한 적지 않다. 이와 달리 음향 특징은 음소 단위로 자동 분절하고 나면 대규모 자유발화 데이터라도 원하는 값을 빠르고 정확하게 추출하여 사용할 수 있다.

본 연구는 자유 발화 속 모음의 음향 특징을 기준으로 충남 지역과 인접한 전북 지역까지 포함하여 방언 구획을 하고 충남 방언의 고유성 여부를 살펴본다. 방언 데이터로는 AI 허브에서 제공하는 '중·노년층 한국어 방언 데이터 (충청도, 전라도, 제주도)'의 음성 자료를 활용한다. 충남과 전북 지역 50세 이상 화자의 자유 발화 음성 중 단모음의 음향 특징값을 추출하고 군집화를 통해 충남과 전북 시군별 군집 결과로 충남 지역이 전북 지역과 다르게 묶이는지 확인한다. 모음 구간 분절을 위해 Montreal Forced Aligner 툴킷의 기능인 Grapheme-to-Phoneme 변환과 강제정렬(forced alignment)을 사용하였고 모음 구간 중간

지점의 포먼트(F1, F2)와 모음 길이 각각의 평균과 표준편차를 계산하였다. 7개 모음별 (‘ㅏ’, ‘ㅓ’, ‘ㅔ’, ‘ㅗ’, ‘ㅜ’, ‘ㅡ’, ‘ㅣ’) 3개 음향 특징의 2개 통계치로 구성된 42개(=7x3x2) 음향 특징 값을 사용하였다. 군집화는 최운호(2010)와 정성훈(2016, 2024)에서 채택한 방법과 동일한 계층적 군집화(hierarchical clustering)를 사용하였으며 군집 거리 계산 방식은 지역 군집화에 효과적인 ‘average linkage’로 설정하였다. 성에 따른 포먼트 값이 다르다는 점을 고려하여 성별로 나누어 군집화를 진행하였으며, 덴드로그램(dendrogram) 결과를 충남-전북 지도로 그림 1과 같이 나타내었다.

화자 전체의 경우 (그림 1 왼쪽) 충남-전북 지역을 통틀어 두 지역이 인접한 내부 지역(파랑)과 외부 지역(빨강)의 두 개 군으로 나뉘었다. 이때 내부 지역은 충남과 전북 시군 지역을 걸친 분포로 나타난 것을 확인할 수 있다. 남성 화자의 경우 (그림 1 가운데) 충남-전북 모두에 분포한 지역(빨강), 충남 중부 지역(초록), 전북 중부 지역(노랑), 기타 지역(파랑)의 네 개 분포로 나타났다. 부여군과 보령시의 데이터 결여로 충남-전북의 연속적인 분포는 확인할 수 없으나 두 지역이 하나의 군집으로 묶인 것을 알 수 있다. 여성 화자의 경우 (그림 1 오른쪽) 군집이 다소 산포되어 있어 지역 간 연속성을 찾아보기 어렵다. 그러나 빨강과 파랑으로 표시된 두 군집에 각각 충남과 전북 시군이 포함된다는 점은 여전히 확인 가능하다. 화자의 성에 따른 세 지도 모두에서 충남 지역과 전북 지역이 공유하는 군집이 형성되었다. 다만 이는 인접 지역 부근에 대해서만이며 비인접 지역은 독자적인 방언권을 이루고 있다고 정리할 수 있다.

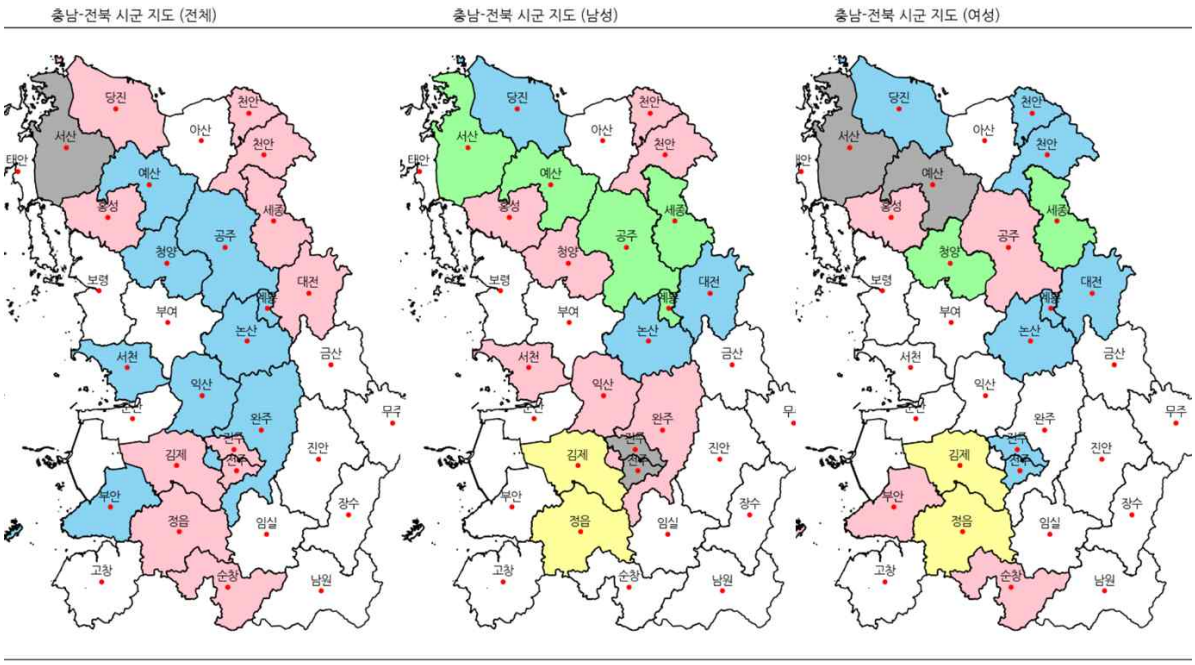


그림 1. 모음 특징에 따른 충남-전북 시군 단위에서의 군집화

(흰색: 해당 지역 데이터 없음, 회색: 어느 군집과도 거리가 멀어 속하지 못함)

본 연구는 자유 발화의 음향 정보를 방언 구획에 적용하였다는 점과 충남, 전북의 두 지역을 함께 고려하여 지도상에서의 방언권 구획을 시도하였다는 점에서 의의가 있다. 다만 제공된 데이터의 한계로 인해 분포 확인이 어려운 시군 지역(하양)도 다소 존재하여 지역 간 연속적인 분포를 분명하게 보기 어려웠다는 점과 이중모음을 연구에 포함하지 않았다는 한계가 있다. 향후 연구에서는 이중모음까지 반영하는 동시에 경기 지역과 충북 지역까지 확장하여 충남 방언권의 독자성 여부를 이어서 밝혀볼 계획이다.

참고문헌

강경원. (2018). 충남의 방언지역 구분. *문화역사지리, 30*(1), 20-32.

도수희. (1965). 충청도방언의 위치에 대하여. *국어국문학, 28,* 237-239.

정성훈. (2016). 네트워크 분석과 하위 방언구획 연구. *한글, 311*, 227-256.

정성훈. (2024). 전라남도 방언구획을 위한 계량적 연구방법 탐색. *우리말연구, 76*, 125-153.

최운호. (2010). 레벤시타인 거리를 이용한 방언거리의 계산과 몽골어의 분석. *알타이학보, 20*, 127-160.

한영목. (1999). *충남방언의 연구와 자료*. 이회.

AI 허브. (2024). 중·노년층 한국어 방언 데이터 (충청도, 전라도, 제주도) (버전 1.2) [데이터세트]. https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71558

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Interspeech, 2017,* 498-502.

일정표로 이동

The Relationship Between Vowel Epenthesis and Stress-Focus Dynamics in L2 Speech Perception

Goun Lee, Dong-Jin Shin
Dankook University, Jeonju University
gounlee@dankook.ac.kr, djshin@jj.ac.kr

This study explores the relationship between L2 learners' ability to perceive epenthetic vowels and their performance in other areas of speech recognition, such as lexical stress, sentence focus, and vowel identification. Twenty-five Korean learners of English participated in perception experiments, including tests on vowel epenthesis, lexical stress, sentence focus, and vowel identification. The findings reveal that performance on the vowel epenthesis task is affected by both lexical stress and sentence focus, suggesting that learners' perception of epenthetic vowels is closely tied to their acquisition of L2 rhythmic structures at both word and sentence levels. Furthermore, the study highlights the role of proficiency, showing that as learners' L2 proficiency increases, the influence of L1 phonotactic constraints on vowel epenthesis recognition weakens. Overall, this study sheds light on how perceptual skills in vowel epenthesis interact with prosodic stress in L2 speech perception.

일정표로 이동

한국어 비격식체 종결어미와 억양의 상호작용

김기유
계명대학교 국어국문학과
kkiiyyuu@naver.com

이 글의 목적은 한국어의 비격식체 종결어미(informal style-sentence final ending)과 발화 말 종결 억양(utterance final intonation)의 상호작용(interaction)을 종합적으로 기술하는 데에 있다. 그 목적을 달성하기 위해 우선, 비격식체 종결어미가 쓰이는 다양한 맥락(contexts)을 상정한 뒤, 발화 실험(production test)을 통해 해당 맥락에서 발화되는 발화 말 종결 억양을 시계열 데이터(time series data)로 추출할 것이다. 그 다음 군집 분석(cluster analysis)을 실시하여 도출된 결과를 해석하는 것으로 달성된다.

일정표로 이동

Refining Cuteness: Fine-tuned ChatGPT evaluations of Korean aegyo

Rok Sim[1], Drew Crosby[2]
Linguistics Program, University of South Carolina[1]
Department of Korean Language and Literature, Korea University[2]
rsim@email.sc.edu, drewmcrosby89@gmail.com

The rapid growth of AI in natural language processing has sparked debates about whether large language models (LLMs) can truly capture human linguistic subtleties or merely approximate them (Cai et al., 2023). Cai et al. (2023) showed that ChatGPT exhibits human-like behaviors, such as sensitivity to dialectal variations. This study extends this inquiry by exploring ChatGPT's understanding of culturally-specific communication style, focusing on aegyo—a cute, playful speech style commonly associated with young Korean women (Crosby, 2023).

Previous studies (Moon, 2013; Jang, 2021) have associated aegyo with various segmental alternations, such as resyllabification, /j/-insertion, and affrication. In this study, we replicated Jang's cuteness-rating experiment on the segmental alternations using ChatGPT, asking it to imagine itself as different social categories (age/gender). Results showed that ChatGPT, while recognizing aegyo as cute, ranked the forms differently than humans, reflecting stereotypes (younger women scoring higher) rather than empirical data (older participants rated aegyo forms as cuter).

We further fine-tuned ChatGPT. Fine-tuning ChatGPT with Jang's human data improved its accuracy but revealed limitations: the model still underestimated cuteness compared to humans and no longer displayed socially dependent differences. These findings suggest that while fine-tuning enhances performance, LLMs still struggle with replicating social nuances in language perception.

References

Cai, Z. G., Haslett, D. A., Duan, X., Wang, S., & Pickering, M. J. (2023). Does ChatGPT resemble humans in language use?. *arXiv preprint arXiv:2303.08014*.

Crosby, D. M. (2023). Oppa-Ng Gamsahamnita-Ng~~~: *The Phonetics of Nasal Cuteness in Korean Aegyo* (Doctoral dissertation, University of South Carolina).

Jang, H. (2021). How cute do I sound to you?: Gender and age effects in the use and evaluation of Korean baby-talk register, Aegyo. *Language Sciences*, *83*, 101-289.

Moon, K (2013). Phrase final position as a site of social meaning: Phonetic variation among young Seoul women [Unpublished master's thesis]. Stanford University. https://web.stanford.edu/~eckert/Courses/l1562018/Readings/MoonUnderReview.pdf.

일정표로 이동

비원어민 한국어 발음 평가를 위한 자기 지도 학습 기반 한국어 음소 인식

경나, 나희정, 박지은, 박정식*
한국외국어대학교 언어공학연구소
gengna0324@gmail.com, 920313@naver.com, ppae0216@naver.com, parkjs@hufs.ac.kr*

Self-supervised Learning Based Korean Phoneme Recognition for Non-native Korean Pronunciation Evaluation

Na Geng, Hee-jung Na, Ji-eun Park, Jeong-Sik Park*
Language Technology Research Institute, Hankuk University of Foreign Studies
gengna0324@gmail.com, 920313@naver.com, ppae0216@naver.com, parkjs@hufs.ac.kr*

비원어민의 한국어 발음 평가를 위해서는 음소 인식 능력뿐만 아니라, 발음 오류를 정확하게 탐지할 수 있는 모델이 필요하다. 자기 지도 학습(Self-supervised Learning) 기반 음성인식에서 대량의 음성 자료를 통해 구축된 사전 학습(pretrained) 모델은 정밀한 음성 인식을 가능하게 하는 것으로 알려져 있다. 특히 Wav2Vec2.0 및 Whisper와 같은 모델들은 여러 연구에서 우수한 한국어 음성 인식 성능을 보였으며, Whisper ASR 모델은 특히 뛰어난 성능을 나타냈다. SUPERB 벤치마크를 통해 다양한 사전 학습 모델을 비교한 결과, 음소 인식 분야에서의 성과도 입증되었다. 그러나 비원어*민 화자의 발화 특성을 반영하여 실제 발화에 정교하게 맞춘 라벨링 데이터의 부족으로 비원어민의 한국어 음소 인식 성능을 높이는 데 한계가 존재한다. 따라서, 비원어민 한국어 발음 평가 모델을 구축하기 위해서는 정확한 데이터 처리와 라벨링 과정이 중요하다. 본 연구에서는 사전 학습된 Wav2Vec2.0 및 Whisper 모델을 활용하여 비원어민의 한국어 발음 평가를 위한 한국어 음소 인식 모델을 개발하였다. AIHub에서는 아시아, 중국, 일본, 유럽, 영어권 비원어민의 한국어 교육용 음성 데이터를 대량으로 제공하고 있어 이를 모델의 미세 조정(fine-tuning)을 위한 데이터로 활용하고자 하였다. 그러나, 제공된 음소 레이블의 정확도가 매우 떨어지는 문제가 있어, 한국인이 명료하게 발음한 “뉴스 대본 및 앵커 음성 데이터”를 추가로 활용하여 정확한 한국어 음소 발음을 학습시켰다. 이 두 데이터로부터 구축한 모델을 통해 비원어민 한국어 음성 데이터의 음소 라벨을 실제 발음에 맞게 수정하고, 이를 다시 fine-tuning에 적용하여 비원어민 한국어 음소 인식 모델을 최종적으로 구축하였다. 구축한 모델의

* 이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(2019-0-00004, 준지도학습형 언어지능 원천기술 및 이에 기반한 외국인 지원용 한국어 튜터링 서비스 개발). 또한 한국연구재단의 지원(No. 2020R1A2C1013162)을 받아 수행된 연구임.

유효성을 평가하기 위해 한국어 Wav2Vec2-ASR 모델과 Whisper 모델을 통해 음소 인식 실험을 진행한 결과, 가공한 데이터를 사용하기 이전의 모델에 비해 비원어민 화자의 음소 인식 성능이 유의미하게 향상되었음을 확인하였다.

일정표로 이동

미지의 공격에 강인한 음성 딥페이크 탐지 모델

손지수, 김승빈, 허정우, 신현서, 임찬영, 구교원, 유하진
서울시립대학교

Robust Audio Deepfake Detection Model Against Unknown Attacks

Jisoo Son, Seung-bin Kim, Jungwoo Heo, Hyun-seo Shin,
Chan-yeong Lim, Kyo-won Koo, and Ha-Jin Yu
University of Seoul
hojasu07@naver.com, kimho1wq@naver.com, jungwoo4021@gmail.com,
gustjtls123@naver.com, cksdud585@naver.com, kkwr0504@uos.ac.kr, hjyu@uos.ac.kr

음성 딥페이크 탐지란, 입력된 음성의 위×변조여부를 판단하는 이진 분류 작업이다. 최근, 심층 신경망 기반 탐지 모델이 이 분야에서 우수한 성능을 보임에 따라, 심층 신경망을 활용하여 탐지 작업에 적합한 모델들을 탐색하려는 연구가 활발히 진행되고 있다.

한편, 음성 생성 기술의 발전으로 인해 새로운 유형의 딥페이크 공격을 탐지하기 위해서는 모델의 높은 일반화 성능이 필수적이다. 모델의 일반화 성능 향상을 위해, 몇몇 연구자들은 자기 지도 학습을 통해 방대한 양의 음성을 사전 학습한 거대 모델을 특징 추출기로 활용하였다[1]. Zhang et al.은 이진 분류 학습을 대체하는 One-Class Knowledge Distillation(OCKD)를 제안하여, 단일 클래스 학습 기법이 알려지지 않은 공격에 대한 강인성을 향상시킬 수 있음을 입증하였다[2]. 이에 본 연구에서는 우수한 딥페이크 탐지 성능을 보여주는 HM-conformer에[3], 자기 지도 학습 기반으로 사전 학습된 거대 모델을 결합한 프레임워크를 제안한다. 추가로, HM-Conformer에 OCKD 기법을 적용하여 보다 우수한 일반화 능력을 보유하도록 설계하였다. 본 연구진은 방대한 양의 진짜 음성 데이터를 사용하여 사전 학습된 거대 모델을 특징추출기로 활용하고, OCKD를 통해 진짜 음성의 분포만 밀도 있게 학습함으로써, 미지의 공격에 대한 탐지 능력을 향상시킬 수 있었다.

본 연구의 실험에서는 ASVspoof2019 logical access(LA)를 훈련 데이터로, ASVspoof2021 deep fake(DF)가 평가데이터로 사용하였으며, 동일 오류율을 평가지표로 사용하였다. 실험 결과 제안모델은 4.23%의 동일 오류율을 달성하여, 15.71%를 기록한 HM-Conformer에 비해 상대개선율 73.07%를 달성하였다. 이를 통해 본 연구진은 제안된 방법이 딥페이크 탐지 작업의 일반화 성능 향상에 효과적임을 확인하였다.

References

[1] Tak, H., Todisco, M., et al. (2022) Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation. Proc. The Speaker and Language Recognition Workshop (Odyssey 2022), 112-119

[2] Lu, J., Zhang, Y., et al. (2023). One-Class Knowledge Distillation for Spoofing Speech Detection. I*CASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 11251-11255.

[2] Shin et al., “HM-Conformer: A Conformer-based audio deepfake detection system with hierarchical pooling and multi-level classification token aggregation methods,” Proc. ICASSP 2024, 2024.

일정표로 이동

효율적인 음성 인식을 위한 조기 종료 기법

윤 지 원
중앙대학교 AI학과
jiwonyoon@cau.ac.kr

Early Exit Technique for Efficient Speech Recognition

Ji Won Yoon
Dept. of AI, Chung-Ang University
jiwonyoon@cau.ac.kr

음성 자기 지도 학습 모델은 음성 인식 분야에서 획기적인 성능 향상을 이끌었다. 그러나 이러한 모델들은 뛰어난 성능을 달성하기 위해 대규모 모델 크기와 높은 연산 비용을 요구하기 때문에 추론 속도가 느려지는 한계를 지닌다. 본 논문에서는 음성 인식 모델의 추론 효율성을 높이기 위한 조기 종료 시스템을 제안한다. LibriSpeech 데이터셋을 활용한 실험 결과, 제안된 모델은 기존 음성 자기 지도 학습 모델에 비해 추론 속도를 개선하면서도 성능과 지연 시간 간의 균형을 효과적으로 유지하는 것을 확인하였다.

일정표로 이동

Apple 스마트폰 음성 녹음 파일의 기종, OS 추측 및 위변조 식별을 위한 메타데이터 기반 결정 트리*

문수호[1], 윤나영[1], 이지원[2], 김경화[3], 홍기형[4]
[1]성신여자대학교 AI융합학부,
[2]성신여자대학교 미래융합기술공학과,
[3]대검찰청 과학수사부,
[4]성신여자대학교 서비스디자인공학과

Metadata-Based Decision Tree for Inferring Device Models, OS Versions , and Detecting Forgery in Apple Smartphone Voice Recordings

Suho Moon[1], Nayoung Yoon[1], Jiwon Lee[2], Kyungwha Kim[3], Kihyung Hong[4]
[1]Dept. of AI Convergence, Sungshin W. University
[2]Dept. of Future Convergence Technology Engineering, Sungshin W. University
[3]Forensic Science Division, Supreme Prosecution Service
[4]Dept. of Service Design Engineering, Sungshin W. University
msh1593@honglab.org, na0@honglab.org, jw010208@honglab.org, savoix@spo.go.kr, khhong@sungshin.ac.kr

스마트폰은 일상생활에서 널리 사용되며, 다양한 법적·사회적 상황에서 중요한 증거로 활용될 수 있다. 그러나 디지털 형식으로 저장된 스마트폰 녹음 파일은 위·변조의 가능성이 있어, 증거의 신뢰성에 영향을 미칠 수 있다. 메타데이터 구조 분석을 통해 이러한 녹음 파일의 원본성 및 편집 여부를 구별하려는 연구들이 진행되어 왔다. 기존 연구에서는 주로 삼성 갤럭시 기기를 중심으로 데이터를 수집과 분석이 이루어졌다[1, 2].

본 연구는 iOS 환경에서 녹음 파일의 고유한 특성을 규명하기 위해 iOS 13.2.3~16.6 버전의 iPhone 음성 메모 앱에서 생성된 압축 및 무손실 압축 파일의 메타데이터를 분석하였다. 녹음 파일은 일반 녹음, 일시정지가 포함된 녹음, 자체 편집된 녹음으로 구분하여 분석을 진행하였다.

분석 과정에서 세 가지 타입의 음성 파일(일반, 일시정지 포함, 자체편집본)과 iOS 버전별 메타데이터를 비교하였다. 그 결과, 편집된 파일에서는 메타데이터 구조에 변동이 생기는 경향이 나타났으며, 녹음 방식 및 편집 여부를 고려한 파일 타입과 iOS 버전에 따라 메타데이터 구조가 다르게 나타나는 것을 확인할 수 있었다. 이러한 분석을 바탕으로 녹음 방식 또는 위변조를 구분할 수 있는 결정 트리를 설계하였다.

*) 본 연구는 2023년 대검찰청 연구 용역의 지원으로 수행되었습니다. (과제 명: 스마트폰 녹음 위변조 식별기술 고도화 연구)

참고문헌

[1] 안서영, 유세희, 김경화, 홍기형. (2022). 법과학적 활용을 위한 삼성 스마트폰 음성 녹음 파일의 메타데이터 구조 및 속성 비교 분석 연구. 말소리와 음성과학, 14(3), 103-112.

[2] 김경화. (2021). 스마트폰 녹음 데이터베이스의 법과학적 활용에 대한 연구. 디지털포렌식연구, 15(1), 26-42.

일정표로 이동

웨어러블 청각 장치 적용을 위한 종단간 음성 디노이징 신경망 연구

김선만
한신대학교 AI·SW학과
smkim@hansin.ac.kr

End-to-End Speech Denoising Neural Network for Wearable Hearing Device Applications

Seon Man Kim
School of Computing and Artificial Intelligence, Hanshin University
smkim@hansin.ac.kr

주변 소음 문제로 인해 청각 장애가 있는 사람 중 실제로 보청기를 사용하는 사람은 약 23%에 불과하다. 이러한 잡음 문제를 해결하기 위해 주로 사용되었던 위너필터와 같은 통계모델 기반 음성향상 방법은 비정상 특성의 잡음환경에서 성능이 제한되는 문제점이 있다. 일반적인 음성향상 기술에 관해서는 지난 10년 동안 심층신경망(DNN)을 활용한 음성 잡음 제거 방법이 광범위하게 연구되어왔으며, 현재는 DNN이 기존 방법에 비해 음성 명료성과 품질을 크게 향상할 수 있다는 것이 보편적으로 인정되고 있다. 특히, 잡음이 있는 파형 입력에서 대상 신호 파형을 직접 예측하려고 시도하는 종단간 구조 전략이 현재 큰 주목을 받고 있으며, 대표적으로 여러 Conv1D 레이어를 기반으로 인코더와 디코더 간의 스킵(Skip) 연결을 통합한 U자형 구조의 Wave-U-Net이 현재 대부분 음성 디노이징 모델의 기본 구조로 사용되고 있다. Wave-U-Net 구조에서는 입력된 잡음 음성신호가 여러 인코딩 레이어를 거쳐 병목 지점에서 잠재 대표 정보로 압축되고, 이 잠재 정보는 디코더 레이어를 통해 점차적으로 목표 음성으로 복원된다. Wave-U-Net 구조의 잡음 제거 성능을 더욱 향상시키기 위한 노력의 일환으로 병목 단에서 잠재적 대표 정보의 시간적 맥락을 유지하기 위해 Long Short-Term Memory (LSTM)과 Gated Linear Unit(GLU)를 활용한 디노이저(Denoiser)가 제안되었으며, 이 모델은 현 종단간 구조의 기본 모델로 보편적으로 활용되고 있다.

본 연구에서는 Denoiser모델을 보청기와 같은 착용형 청각기기에 적용하기 위한 내용을 다룬다. 음성향상 모델이 착용형 청각기기에 적용되기 위해서는 10ms이하의 시간지연을 필수적으로 만족하여야 하며, 이를 위해 먼저, 디노이저 모델의 시간 지연 메커니즘을 분석하고, 이로부터 10ms이하를 만족하는 수정된 디노이저 구조를 제안한다.

일정표로 이동

시선추적 기반 인공와우이식 아동의 청각 단어재인 특성

이영미[1], 윤해나[1], 최병윤[2]
[1]이화여자대학교 대학원 언어병리학과
[2]서울대학교 서울대학교 의과대학 서울대학교병원 이비인후과학교실
youngmee@ewha.ac.kr, hamm_123@ewhain.net, choiby2010@gmail.com

Real-time Spoken Word Recognition Processing in Young Children with Cochlear Implants:

Youngmee Lee[1], Haena Yoon[1], Byung Yoon Choi[2]
[1]Department of Communication Disorders, Ewha Womans University, Seoul;
[2]Department of Otorhinolaryngology-Head and Neck Surgery, Seoul National University Bundang Hospital, Seoul National University College of Medicine, Seongnam
youngmee@ewha.ac.kr, hamm_123@ewhain.net, choiby2010@gmail.com

Background and Objectives: Spoken word recognition is crucial for language processing, as it links auditory perception to meaning and syntax. Efficient word recognition is essential for successful language development. This study used eye-tracking technology to investigate real-time spoken word processing in young children with cochlear implants (CIs).

Methods: Twelve children with CIs, aged 19 to 45 months, were divided into two groups based on their language development rate: faster language learners (FLL) and slower language learners (SLL). The children completed the intermodal preferential-looking task, and their responses were recorded online.

Results: The FFL group showed significantly less time looking at both the targets and distractors compared to the SLL group. The SLL group tended to fixate more on the distractors, while the FLL group focused more on the targets. However, these differences were not statistically significant. Negative correlations were observed between language age and overall proportion of looking time to the targets and distractors in children with CIs.

Conclusion: These findings suggest that slower language learners may have less efficient spoken word processing. Eye-tracking measurements could serve as a valuable clinical tool for predicting language skills in children with CIs.

일정표로 이동

인공와우이식 유아의 자발화 음성 분석에 대한 종단 연구: 예비 보고

이승진[1], 이영미[2]
[1]한림대학교 자연과학대학 언어청각학부 및 청각언어연구소, [2]이화여자대학교 언어병리학과
[1]sjl@hallym.ac.kr, [2]youngmee@ewha.ac.kr

A Longitudinal Study on the Voice Analysis of Spontaneous Speech in Children with Cochlear Implants: A Preliminary Report

Seung Jin Lee[1], Youngmee Lee[2]
[1]Div. of Speech Pathology and Audiology, Research Institute of Audiology and Speech Pathology, College of Natural Sciences, Hallym University
[2]Department of Communication of Disorders, Ewha Womans University
[1]sjl@hallym.ac.kr, [2]youngmee@ewha.ac.kr

목적: 양측 고도난청 아동은 조기에 인공와우이식(Cochlear Implantation, CI)을 통해 정상적인 말언어발달을 위한 청각자극을 받을 수 있으나, CI 후에도 운율, 말 속도, 명료도 차원에서 건청(typical hearing, TH) 아동과 차이가 있으며, 유아 시기의 음성 특징 변화에 대한 연구는 부족하다. 따라서 이 연구에서는 CI 유아와 TH 유아의 자발화 음성 특성 변화를 종단적으로 추적하여 비교하고자 하였다.

방법: 연구대상자는 양측 CI 수술을 받은 양측고도난청 유아 6명과 건청 유아 8명이었다. 월령 20.9±4.6개월 시점(T1), 27.6±5.1개월 시점(T2), 34.1±4.9개월 시점(T3)에 아동과 보호자와의 자연스러운 놀이 상황에서 포터블레코더와 유선 핀마이크를 이용하여 자발화 음성을 녹음하였다. 녹음된 음성 중 잡음 구간이나 보호자의 음성과 겹치는 구간을 제외한 후, 각 시기별 음성을 결합하여(concatenate) 만든 음성 샘플에 대하여 음도, 강도, 음질 차원의 음향학적 분석을 수행하였다.

결과: 음도 차원에서는 Hz와 반음단위의 음도범위가 T1보다 T3에서 넓어졌으며, TH보다 CI가 더 급격한 증가폭을 보였다. 또한 두집단 모두 F0의 최솟값이 T1에서 더 높았던 반면, F0와 반음 단위의 음도 표준편차는 T1시기에 더 낮았다. 강도 차원에서는 DB 최솟값이 CI가 TH보다 더 컸으며, T1이 T3보다 컸고, 강도의 범위도 T1이 T3보다 좁았다. 강도의 최댓값은 T1에서 CI가 TH에 비해 훨씬 크다가 T2에서 유사해지는 양상이었다. 음질의 경우 SR과 SR의 최댓값이 T2와 T3보다 T1에서 더 낮았다. 그외 평균 CPP, F0, DB 등의 변수에서는 유의미한 차이가 없었다.

결론: 연령이 증가함에 따라 음도와 강도 차원에서 CI 유아와 TH 유아 공히 더 작고 낮은 소리를 산출하는 경향이 있었다. 또한 일부 변수에서는 CI 유아의 인공와우 사용기간이 길어짐에 따라 청각 정보 입력이 충분히 이루어질수록 음성의 음도와 강도에 있어 정상 발달 아동의 음성 발달을 점차적으로 따라잡게 되는 경향이 관찰되었다. 이는 난청아동의 CI 중재 필요성과 함께 유아 시기 음성 발달을 장기적으로 추적하여 중재효과를 확인할 필요성을 시사하는 것으로 여겨진다.

일정표로 이동

시선추적장비를 활용한 일반 아동의 단어재인 양상

김지연, 이나경, 이영미
이화여자대학교 언어병리학과
jiykim103@ewhain.net, lna0705@ewhain.net, youngmee@ewha.ac.kr

Real-time Spoken Word Processing in Children with Typical Hearing

Jiyeon Kim, Nakyoung Lee, Youngmee Lee
Dept. of Communication Disorders, Ewha Womans University
jiykim103@ewhain.net, lna0705@ewhain.net, youngmee@ewha.ac.kr

Spoken word recognition refers to the process of listening to spoken words and understanding their meaning. It requires the conversion of auditory signals into linguistic forms and the retrieval of phonological information from the lexicon and is an important predictor of children's language processing and acquisition (Dollaghan, 1998). This study aimed to find out the patterns of spoken word recognition process in children using eye-tracking technology. Twenty children, aged between 2 to 6 years old, were presented with two pictures (a target and a distractor) on a screen, along with auditory stimuli containing a target word. The correlations between the collected eye tracking variables(overall proportion of looking time, fixation counts) in response to spoken words and the children's chronological and language age were then analyzed using SPSS 29.0. The results showed that with increasing chronological age, fixation counts on distractors significantly decreased. Furthermore, as language age increased, fixation counts on distractors decreased, while the overall proportion of looking time on targets increased. According to Baddeley(1986)'s study, children takes less time and energy to process words they already know compared to uncoded information. This result indicates that the older the child's chronological and language age gets, the more words they've already learned, so they can easily find out targets without looking at distractors multiple times.

References

Dollaghan C. (1998). Spoken word recognition in children with and without specific language impairment. *Applied Psycholinguistics, 19*(2), 193-207.

Baddeley, A. (1986). *Working memory.* Oxford: Oxford University Press.

2024 한국음성학회 가을 학술대회

포스터 발표 논문

일정표로 이동

Acoustic cue weighting in lexical stress perception of real and nonce words by Korean learners of English

Eunkyung Sung, Seulgi Oh
Dept. of English, Cyber Hankuk University of Foreign Studies
Dept. of Korean Language and Literature, Hankuk University of Foreign Studies
eks@cufs.ac.kr, jyoua501@gmail.com

In speech perception, listeners must identify which acoustic cues are relevant and evaluate the importance of each cue. While listeners rely on various acoustic cues simultaneously to distinguish sound contrasts, the weighting of these cues varies across languages (Holt and Lotto 2006, Zhang and Francis 2010, Chrabaszcz 2014). This paper has two main objectives. First, it examines how listeners' first language (L1) affects their use of acoustic cues in perceiving stress. Second, it examines how the use of these cues differs when listeners encounter English nonce words and real words. We manipulated three acoustic cues－duration, pitch(f0), and intensity－ based on three levels in both nonce words (i.e., /kɑtɑ/, /kətə/) and real words (i.e., console, converse, object). A total of 42 listeners participated in a lexical stress judgment task, consisting of 21 native Korean speakers learning English and 21 native English speakers. In order to test the effects of four acoustic cues on the perception of English lexical stress, the participants completed a forced-choice auditory identification task implemented through the PsychoPy program.

Figure 1. Screen Layout for the Identification Experiment (L: nonce words, R: real words)

The results revealed a significant difference between the two listener groups in their processing of nonce words, but not real words. The influence of the three acoustic cues (duration, f0, and intensity) was more clearly shown in nonce words than in real words. When examining each group separately, Korean listeners primarily relied on f0 and

intensity to identify

lexical stress in both nonce and real words. However, their reliance on intensity was stronger for nonce words than for real words. In contrast, English listeners used all three cues when processing nonce words, but relied only on duration and intensity for real words. Like the Korean listeners, the English participants exhibited a stronger reliance on duration and intensity when dealing with nonce words compared to real words.

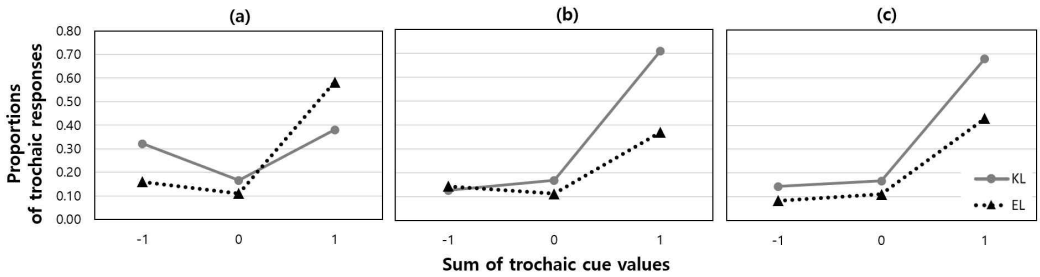


Figure 2. Two Listener Groups' Proportions Trochaic Responses (Responses of 1 or 2) in Nonce Words as a Function of Duration (a), f0 (b), and Intensity (c) Manipulation

Table 1. The Effects of Listener Group and Acoustic Cues in Nonce words

Variable	Estimate	Std.Error	z value	*p*-value
(Intercept)	5.47020	0.24956	21.920	<.001 ***
Language	0.31088	0.06394	4.862	<.001 ***
Duration	-0.50705	0.06040	-8.395	<.001 ***
f0	-0.90041	0.06345	-14.190	<.001 ***
Intensity	-1.16556	0.06650	-17.528	<.001 ***

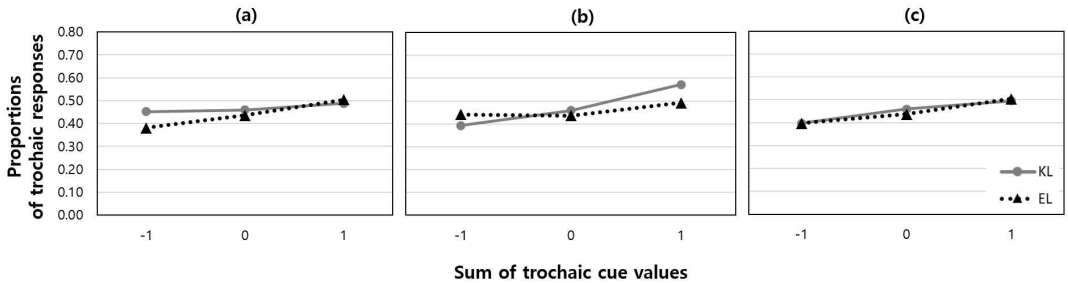


Figure 3. Two Listener Groups' Proportions of Trochaic Responses (Responses of 1 or 2) in Real Words as a Function of Duration (a), f0 (b), and Intensity (c) Manipulation

Table 2. The Effects of Listener Group and Acoustic Cues in Real Words

Variable	Estimate	Std.Error	z value	*p*-value
(Intercept)	1.35041	0.22461	6.012	<.001 ***
Language	0.06623	0.05984	1.107	0.27
Duration	-0.16038	0.06345	-2.527	<.05 *
f0	-0.23309	0.06368	-3.661	<.001 ***
Intensity	-0.20880	0.06359	-3.283	<.01 **

References

Chrabaszcz, A., Winn, M., Lin, C. Y., & Idsardi, W. J. (2014). Acoustic cues to perception of word stress by English, Mandarin, and Russian speakers. *Journal of speech, language, and hearing research 57*(4), 1468-1479.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America 119*(5), 3059-3071.

Zhang, Y., & Francis, A. (2010). The weighting of vowel quality in native and non-native listeners' perception of English lexical stress. *Journal of Phonetics 38,* 260-27l.

일정표로 이동

Prospective English teachers' cue-weightings in L2 stop perception

Hyunjung Lee[1], Eun Jong Kong[2]
Dept. of English Education, Incheon National University[1]
School of Liberal Arts and Sciences, Korea Aerospace University[2]
hjlee123@inu.ac.kr[1], ekong@kau.ac.kr[2]

Research on individual variability in speech processing has revealed that non-linguistic factors including cognitive abilities and personality traits systematically influence one's linguistic behaviors. In particular, studies on L2 learning have sought to identify how these individual traits correlate with learners' ability to process unfamiliar sounds and whether these factors contribute to more effective language learning. This study investigates how individual differences manifest in prospective English teachers' perception of voiced-voiceless stops, focusing on acoustic cue weighting (VOT, F0). Specifically, it examines the relationship between response patterns to perceptual distractors and personal traits as measured by the Big Five. Twenty-five university students majoring in English Education participated in the 2AFC perception task, judging English /d/ and /t/ while solving distracting math calculations. In addition, participants also completed an English proficiency cloze test and the Big Five personality survey. All the experimental tasks were designed and performed in the online platform. We will present findings on how individuals adjust VOT and F0 in stop identification under cognitive distraction and discuss how these patterns relate to their personality traits. The results provide insights into developing a customized L2 teaching model that addresses individual differences in cognition, personality, and aptitude.

일정표로 이동

일상에서의 자가 인지기능 평가 문항개발을 위한 예비연구

박상희, 이옥분
대구사이버대학교 언어치료학과

A Preliminary Study for the Development of Self-Assessment Tool for Cognitive Function in Daily Life

Sanghee Park, Okbun Lee
Dept. of Speech and Language Pathology, Daegucyber University
psh4292@dcu.ac.kr, oblee@dcu.ac.kr

I. 서론

정상 성인 및 노인이 일상생활에서 자신의 인지 기능이나 기억력의 변화를 스스로 체크할 수 있는 간단한 평가 도구의 개발은 국내외적으로 계속해서 연구보고 되고 있다. 노화에 따른 병리적 감퇴에 대한 최대한의 빠른 확인은 여러 가지 측면에서 개인과 가족 그리고 공동체의 삶의 질에 중요한 의미를 가진다. 즉, 초기 인지 저하 발견은 정상 성인이라도 나이가 들거나 스트레스, 수면 부족 등의 요인으로 인해 인지 기능이나 기억력에서 미묘한 변화를 경험할 수 있기 때문에 이러한 변화를 조기에 감지하는 것은 치매나 경도 인지장애(MCI)와 같은 질환의 초기 단계에서 적절한 대응을 할 수 있게 한다. 그리고 간편하고 접근 가능한 방식의 평가 도구는 복잡한 임상 검사가 아닌 간단한 평가 도구는 누구나 쉽게 사용할 수 있어, 시간과 비용을 절약하면서도 꾸준히 자신의 인지 상태를 모니터링할 수 있는 장점이 있기 때문에, 보다 많은 사람들이 자신을 돌보고 건강한 노화를 유지하는 데 기여할 수 있다는 이점이 있다. 이는 건강한 노화 및 예방 차원에서 인지 기능이 나빠지기 전에 건강한 습관을 유지하거나 기억력 훈련을 통해 인지 저하를 예방하는 것이 중요하기에, 간단한 평가문항과 방식의 평가도구를 통해 스스로 상태를 점검하면 이러한 다방면 차원에서의 노화 관리에 중요한 정보를 제공할 수 있게 된다. 이러한 차원에서 이 연구에서는 국외 연구보고된 일상에서의 인지기능을 간편하게 평가할 수 있는 프로토콜(Annunziata, et. al, 2012)을 번안 및 수정 그리고 검토하여 국내 정상 성인을 대상으로 한 반응 결과들을 알아보고자 한다.

II. 연구방법

본 연구에서는 국내 정상 성인(인지 및 의사소통 기능에 문제가 없는 자) 50명을 대상으로 Annunziata, et al, 2012의 연구 보고된 인지기능 평가 문항을 번안 및 수정하고 연구자들이 검토하여 구글 설문으로 작성하고 이 설문에 반응한 결과들을 분석하였다. 대상자 특성은 연령은 20대 1명, 30대 22명, 40대 19명, 50대 4명, 60대 4명이었고, 성별은 남성 7명, 여성 43명이었다.

학력은 고졸 6명, 대졸 36명, 대학원졸 8명이었다.의사소통능력에 문제에 대한 자가평가에서 문제가 없다고 대답한 경우는 13명으로 26%였고, 문제가 있다고 느끼는 경우가 37명으로 74였다. 문제는 단어 및 상황이 전혀기억이 나지 않는 경우가 있었고 조리있게 말하는 것이 부족하다고 느끼는 경우가 있었다. 인지기능의 관련 문항은 총18개이며, 5점 척도로 평가하도록 하였다. 설문 참여자들은 일상생활에서 해당하는 인지기능(설문 문항)에 대해 심리적으로 자각하는 정도(5점 척도)로 체크할 수 있도록 하였다. 자가인지 능력은 주의력, 공감시간지남력, 기억으로 나누어 각 영역과 구두유창성에 대해서는 상관분석을 실시하였다.

III. 연구결과

평소 생활에 대한 심리적으로 느끼는 정도에 대해서 집중력 부족에 대해서는 절대없다고 답한경우는 10%였고 자주나타나는 경우는 14%, 가끔 34%, 거의 드물게 42%가 나타난다고 하여 집중력 부족에 대해서는 연령에 상관없이 전반적으로 느끼고 있었다.

주의 산만에 관해서는 44%가 정도가 주의 산만에 대해서 느끼고 있었다. 기억 중 구체적인 단어 대신 완곡하거나 일반적인 용어사용에 대해서는 50%정도가 가끔씩 이상으로 느끼고 있었다. 주의력에 관한 지적/인지 활동 중 잠시 마음이 떠나는 현상에 대해서는 44%가 가끔씩 이상으로 느끼고 있었다. 주의력 중 정도 상기나 건만증에 대해서도 40% 이상에서 어려움이 있다고 반응하였다. 기억 중 적절한 단어를 찾는데 어려움 점에 대해서는 50%가 어려움이 있다고 답하였다. 특히 주의력 독서 중 집중력 부족에 대해서는 자주 26%, 가끔씩 40%, 거의드물게 26%로 나타났다. 공간-시간-지남력 중 운동 협응 부족에 대해서는 절대없음이 40%, 공간-시간-지남력 중 움직인 실행 속도 느림에 대해서 절대없음이 40%로 운동 부분에서는 어려움이 있다고 답하지는 않았다. 위 18개의 항목 중 6개월 동안 변화가 있다고 답한 항목은 건만증이나 기억부분에서 특히 변화를 많이 느꼈고, 없다고 대답한 경우도 19명이었다.

구두 유창성에서 동물이름말하기는 평균 20.43이었고, /ㄱ/음소로 된 단어 말하기는 18.06으로 동물이름 말하는 능력이 더 높게 나타났다. 구두 유창성 영역과 자가인식 능력과의 상관에 대해서는 상관이 없는 것으로 나타났다. 자가 인지능력간에는 주의력과 공간시간지남력, 주의력과 기억, 공간시간 지남력과 기억에서 모두 .01수준의 유의한 상관이 있는 것으로 나타났다.

IV. 결론 및 논의

나이가 듦에 따른 인지기능의 다차원적인 변화는 일상에서의 삶의 질에 중요한 영향을 미친다. 따라서 일상 속에서 좀더 다차원적으로 인지기능의 변화를 감지할 수 있는 간편한 자가 평가도구의 개발은 국내 다양한 치매감별 검사도구와 주관적인지변화 관련 평가들과 함께 활용할 때 보다 유의미한 정보를 제공할 것으로 사료된다. 자신의 총체적인 생활을 고려해서 인지 기능을 스스로 모니터링할 수 있는 다변적인 인지검사도구의 개발은 성인 및 노인의 스스로의 상태를 잘 인식하고 필요한 경우 전문가의 도움을 받거나 생활 방식을 조정하는 등 적극적으로 관리할 수 있는 기회를 제공하므로써, 정신적 건강을 유지하고 웰빙을 증진시키는 데 중요한 역할을 할 것으로 기대된다.

참고문헌

Annunziata, M. A., Muzzatti, B., Giovannini, L., & Lucchini, G. (2012): Cognitivefunctioning self-assessment scale (CFSS): Preliminary psychometric data, *Psychology, Health & Medicine, 17*:2,

일정표로 이동

인공와우이식 영유아 어머니의 아동지향어 변화에 관한 종단연구

이누리[1], 박희선[2], 이영미[1]
이화여자대학교 언어병리학과[1], 김포연세언어치료교육원[2]

A longitudinal study on the changes in Korean mother's child-directed speech to young children with cochlear implants

Nuri Lee[1], Heesun Park[2], Youngmee Lee[1]
Dept. of Communication Disorders, Ewha Womans University, Seoul, Korea[1]
Gimpo Yonsei Speech-Language Therapy Center, Gimpo,Korea[2]
shoona85@ewhain.net, parkhs96@naver.com,youngmee@ewha.ac.kr

1. 서론

아동지향어는 전형적으로 음도는 높고 천천히 말하는 특징이 있다. 청각장애 영유아는, 부모의 발성에 대해 덜 반응적이고 수동적인 것을 고려할 때(Lee & Lee, 2023), 건청 부모가 청각장애 영유아에게 적절한 아동지향어를 사용할지에 대해서 불확실하다. 따라서, 본 연구에서는 시간의 흐름(초기 평가, 6개월 후, 12개월 후)에 따라 인공와우이식 영유아 부모의 아동지향어 관련 변수(평균 음도, 발화 길이, 발화 수, 말 속도)의 변화를 살펴보고자 하였다.

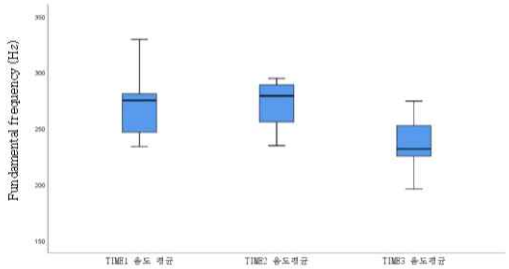
2. 연구방법

본 연구의 대상자는 인공와우이식 영유아 어머니의 8명을 대상으로 초기평가,6개월 후,12개월 총 세 시점의 어머니의 발화를 수집하였다. 어머니 발화 분석은 Praat 프로그램(ver.6.1.24)을 사용하여 평균기본주파수(average fundamental frequency, Fo: Hz),발화 수 및 표준편차 ,발화길이(sec), 말속도(초당 평균 음절수)를 분석하였다. 유의한 차이가 있는지를 살펴보기 위하여 일원배치 반복측정분산분석(One-way repeated Measures ANOVA)을 실시하였다.

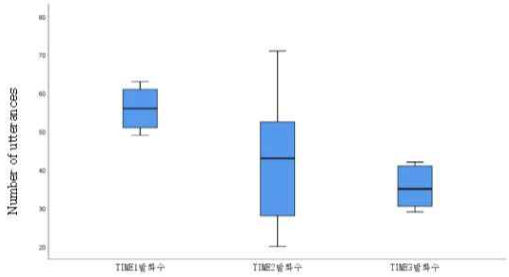
3. 연구결과

연구결과, 평균 음도에서는 시점에 대한 주 효과는 유의하였다. ($F(2,14)=12.224$, $p=.001$). 주 효과에

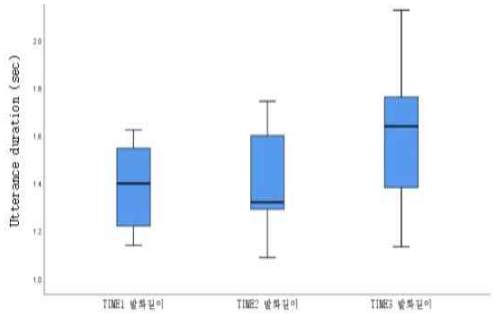
대한 Bonferroni 사후검정 결과, 초기와 12개월간(p =.036), 6개월과 12개월간(p =.005)의 음도 차이가 유의하였다. 연구 결과는 <그림 1>과 같다. 발화 수에서도 시점에 대한 주 효과가 유의하였다(F(2,14)=4.770, p =.026). 주 효과에 대한 Bonferroni 사후검정 결과, 초기와 12개월간(p =.046) 발화 수가 유의하였다. 연구 결과는 <그림 2>와 같다. 발화 길이에서는 시점에 대한 주 효과가 유의하지 않았다(p =.291).말속도에서도 시점에 대한 주효과가 유의하지 않았다(p =.145). 연구 결과는 <그림3>과 <그림4>와 같다.



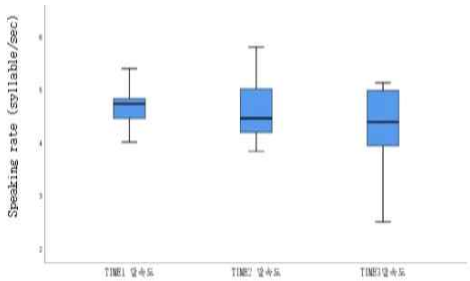
<그림 1> 평균 음도 종단별 비교



<그림 2> 발화 수 종단별 비교



<그림 3> 발화 길이 종단별 비교



<그림 4> 말 속도 종단별 비교

4. 논의 및 결론

본 연구에서는 시점에 따른 인공와우이식 영유아 어머니의 평균 음도와 발화수의 유의한 차이가 났다. 평균 음도에서는 시점에 대한 주 효과는 유의하였다. 초기와 12개월 간 6개월 후와 12개월간 음도 차이가 유의하였다. 발화 수에서도 시점에 대한 주 효과는 유의하였다. 초기와 12개월간 발화 수가 유의하였다. 이러한 결과는 아동의 연령이 발달하면서 아동의 발화 수가 증가하고 엄마의 발화 수는 감소하고 있다고 볼 수 있다.

References

이다예, 이영미, 이유진 (2023). 인공와우이식 영유아 부모의 양적 및 질적 언어입력변화에 관한 종단연구, *Communication Sciences & Disorders , 23*(3), 669-688.

일정표로 이동

생후 12개월 이전에 인공와우이식을 받은 선천성 농 영유아의 발성 발달

이나경[1], 전예솔[2], 이영미[1]
[1]이화여자대학교 언어병리학과[1], [2]보아스 목소리 언어 연구소
lna0705@ewhain.net, yesol0811@naver.com, youngmee@ewha.ac.kr

Vocal Development in Congenitally Deaf Children Receiving Cochlear Implants in the First Year of Life

Nakyoung Lee[1], Yesol Jeon[2], Youngmee Lee[1]
[1]Dept. of Communication Disorders, Graduate School of Ewha Womans University
[2]BOAZ Voice & Language Research Institute
lna0705@ewhain.net, yesol0811@naver.com, youngmee@ewha.ac.kr

1. 서론

영유아의 초기 발성 발달은 말과 언어발달과 밀접한 관련이 있어 매우 중요한 과정이다. 발성은 생리적인 소리에서 시작해 옹알이로 발전하며, 말의 기본 요소를 습득하는 기초가 된다. 난청 영유아는 청력 손실로 인해 발성 발달이 지연될 수 있으며, 이는 말과 언어 습득에 부정적인 영향을 미칠 수 있다. 이를 보완하기 위해, 고도(severe) 이상의 난청 영유아는 인공와우이식(cochlear implantation)을 받으며, 조기 청각언어중재의 중요성이 강조된다.

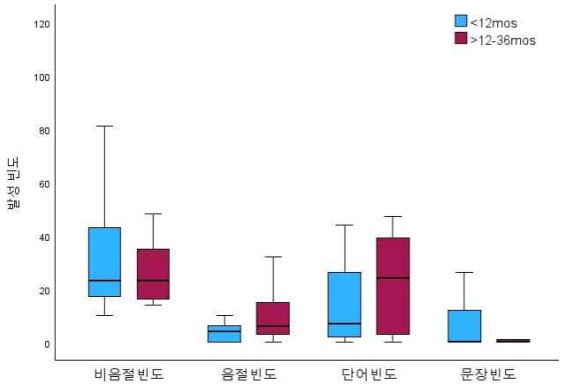
12개월 이전에 인공와우이식을 받는 경우, 청각 신경망이 제대로 형성되어 다양한 언어 영역에서 정상 범위 내 발달이 가능해진다(Dettman & Dowell,2010). 조기 인공와우이식은 초기 발성 발달을 촉진하며, 이후의 언어 발달 속도를 크게 향상시킨다고 보고된다(Oller, 1999; Svirsky et al., 2000). 본 연구에서는 인공와우이식 연령이 12개월 이하인 영유아와 12-36개월인 영유아의 발성 발달을 발성 유형(비음절 발성, 음절 발성, 단어, 문장)별로 비교분석하고, 인공와우이식 영유아의 인공와우이식 관련 변수와 발성 관련 변수 간의 관련성을 탐색해보고자 하였다.

2. 연구방법

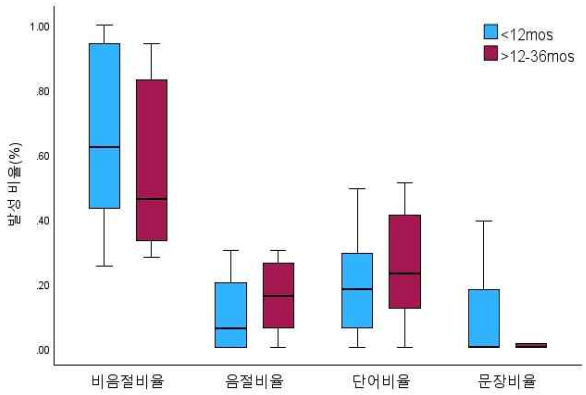
본 연구는 생후 12개월 이전과 12-36개월 사이에 동시적 양측 인공와우이식을 시행 받은 영유아 30명(남 17, 여13)을 대상으로 부모-영유아 상호작용 데이터를 수집하였다. 수집한 데이터는 10분으로 편집하여 Praat 6.3.19으로 분석하였다. 발성의 구분은 한 호흡 단위로 하였으며(김호, 하승희, 2022), 비음절성 발성, 음절성 발성, 단어, 문장으로 나누어 분석하였다.

3. 연구결과

인공와우이식연령이 12개월 이하인 집단과 12-36개월인 집단 간에 발성 유형에 따른 발성 빈도와 비율의 차이를 살펴본 결과, 12개월 이하 집단의 문장 산출 비율이 12-36개월 집단에 비해서 유의하게 높았고(p<.05), 결과는 다음 <그림 1>, <그림2>와 같다. 그리고 인공와우이식 영유아의 생활연령이 증가할수록 음절성 발성, 단어, 문장 산출 비율이 유의하게 증가하였다. 언어연령이 증가할수록 단어와 문장의 산출 빈도와 비율이 모두 유의하게 증가하였다.



<그림 1> 발성 유형별 발성 빈도



<그림 2> 발성 유형별 발성 비율

4. 논의 및 결론

본 연구 결과로 12개월 이하에 인공와우이식을 받은 영유아들이 12-36개월에 이식을 받은 영유아들보다 발성 발달에서 더 높은 성취를 보이며 특히, 12개월 이하에 이식을 받은 영유아의 문장 산출 비율이 유의하게 높았다. 이는 12개월 이하의 인공와우이식이 난청 영유아의 발성 발달에 긍정적인 영향을 미친다는 점을 보여준다. Svirsky et al.(2000)은 조기인공와우이식 아동들이 청각 자극에 대한 반응이 더 빠르고, 청각적 피드백을 통해 자신의 발성을 조절하는 능력이 향상되어 문장 발달로의 전환이 빨라진다고 언급한 바 있다. 또한, 인공와우이식 영유아는 생활연령과 언어연령이 증가할수록 더 발달된 형태의 발성 산출 빈도와 비율이 모두 증가함을 확인하였다. 이를 통해 난청 영유아의 구어 의사소통 발달을 위한 발성 평가와 촉진 전략을 조기에 제공하는 것이 중요함을 확인하였다.

References

김호,하승희. (2022). 초기 발성과 낱말과의 관계. *말소리와 음성과학, 27*(1), 1-13.

Svirsky, M. A., Robbins, A. M., Kirk, K. I., Pisoni, D. B., & Miyamoto, R. T. (2000). Language development in profoundly deaf children with cochlear implants. *Psychological Science, 11*(2), 153-158.

Dettman, S., & Dowell, R. (2003). Language acquisition and critical periods for children using cochlear implants. In M. Marschark & P. E. Spencer (Eds.), *The Oxford handbook of deaf studies, language, and education*(Vol. 2, pp. 331-344). Oxford University Press..

Oller, D. K., Eilers, R. E., Neal, A. R., & Schwartz, H. K. (1999). Precursors to speech in infancy: The prediction of speech and language disorders. *Journal of Communication Disorders, 32*(3), 223-245.

일정표로 이동

멀티모달 접근법을 통한 L2 영어 말하기 평가 모델 개발

오세현, 김선희, 정민화
서울대학교 협동과정 인공지능, 서울대학교 불어교육과, 서울대학교 언어학과
ohsehyun12@snu.ac.kr, sunhkim@snu.ac.kr, mchung@snu.ac.kr

본 연구는 L2 영어 말하기 평가에서 발화의 최종 점수를 예측하기 위해 음성 및 텍스트 임베딩을 결합한 멀티모달 프레임워크를 제안합니다. 이 프레임워크는 음성과 텍스트 데이터를 통합하여 학습할 수 있도록 설계되었으며, 이를 통해 더 풍부하고 다층적인 언어 정보를 활용할 수 있습니다. 음성 신호의 음향적 특성과 텍스트 데이터의 의미론적 정보를 동시에 학습함으로써, 발화 평가의 정확성을 향상시키고자 합니다. 연구의 주요 목표는 음성과 텍스트 임베딩의 조합이 L2 말하기 평가에 미치는 영향을 확인하는 것입니다. 이를 위해 MFCC와 GloVe, MFCC와 Word2Vec, wav2vec 2.0과 BERT, 그리고 wav2vec 2.0과 GPT-3 임베딩을 사용하여 다양한 실험을 수행했습니다. 특히, Transformer 기반의 모델이 생성하는 임베딩과 그렇지 않은 임베딩 구조를 비교함으로써 두 접근 방식의 성능 차이를 비교하여 평가하였습니다. 실험 결과, wav2vec 2.0과 GPT-3 임베딩을 결합한 모델이 가장 높은 예측 성능을 보였으며, Pearson 상관계수(PCC) 0.728을 기록하였습니다. 이러한 결과는 최신의 멀티모달 접근법이 자동화된 말하기 평가 분야에서 큰 가능성을 가지고 있음을 시사합니다. 또한, 본 연구는 향후 더욱 발전된 말하기 평가 모델을 개발하기 위한 기초 연구로서, 멀티태스크 학습 기법과 joint loss function을 도입하여 영어 말하기 음성에서 여러 평가 점수를 동시에 예측할 수 있는 모델로의 확장을 목표로 하고 있습니다. 이와 같은 접근법은 추후 더 높은 성능과 정확성을 제공하는 자동화된 언어 평가 시스템의 개발에 중요한 기여를 할 것으로 기대됩니다.

일정표로 이동

인공지능 기반 말더듬 반복과 연장에 대한 자동 식별: 합성곱층신경망(CNN)을 이용하여

박 진, 이창균
가톨릭관동대학교 언어재활학과, 가톨릭관동대학교 경영학과

AI-based Automatic Detection of Repetitions and Prolongations in Stuttered Speech: Using a Convolutional Neural Networks (CNN)

Jin Park, Chang-Gyun Lee
Dept. of Speech-Language Rehabilitation, Catholic Kwandong University
Dept. of Business Administration, Catholic Kwandong University
gatorade70@cku.ac.kr, kdmis@cku.ac.kr

본 연구는 인공지능 기술을 기반으로 말더듬 비유창성(즉, 반복, 연장, 막힘)을 자동적으로 식별할 수 있는 방법을 개발하고자 하는 목적으로 진행하였다. 특히, 막힘의 경우, 비교적 짧은 '끊김' 또는 단순한 시각적 긴장(tension)으로도 나타날 수 있기에 음성 데이터만을 가지고 정확한 파악이 어려울 수 있다(Guitar, 2019). 따라서 막힘의 정확한 식별을 위해서는 말더듬 화자의 음성 데이터뿐 아니라 턱이나 입술 등과 같은 조음기관의 긴장을 보여주는 시각적 단서를 포함한 종합적 판단이 필요하다(Altinkaya & Smeulders, 2020). 이러한 점을 고려해 본 연구에서는 막힘을 제외한 말더듬 비유창성, 즉 반복(음소 반복, 단어 반복, 구 반복을 포함)과 연장에 대한 자동 식별 방식을 개발하고자 하였다. 또한 비교적 적은 음성 데이터셋을 통한 이전 연구(Park & Lee, 2023)와는 달리 일정량 이상의(50명의 말더듬 화자의 20시간 길이의 음성 데이터) 공개된 데이터셋(즉, LibriStutter)을 활용해, 수립하고자 하는 말더듬 자동 식별 모델의 견고성(robustness)이나 일반화 가능성도 증대시키고자 하였다. 이를 위해 공개 LibriStutter ASR 말뭉치에서 제공하는 메타데이터를 기반으로 음성데이터를 MFCCs 이미지 데이터로 전처리하여 학습 및 검증 데이터셋을 7:3으로 구성하였다. 딥러닝 알고리즘으로 CNN 알고리즘을 활용하여 음소 반복, 단어 반복, 구 반복, 연장을 구분하는 인공지능 자동 식별기를 설계하였다. 또한 식별기 성능을 최적화하기 위해 Grid Search방식을 적용하고 Python KERAS 라이브러리를 활용하였다. 본 연구를 통해 보다 식별 정확도가 높고 신뢰성있는 말더듬 자동 식별 기술의 개발과 함께 이를 통해 고도화된 말더듬 평가 및 중재 관련 서비스가 창출되기를 기대한다.

References

Altinkaya, M., & Smeulders, A. W. M. (2020). A dynamic, self supervised, large scale audiovisual dataset for stuttered speech. Proceedings of the 1st International Workshop on Multimodal Conversational AI (pp. 9-13). Seattle, WA.

Barrett, L., Hu, J., & Howell, P. (2022). Systematic Review of Machine Learning Approaches for Detecting Developmental Stuttering in IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 1160-1172. doi: 10.1109/TASLP.2022.3155295.

Guitar, B. (2019). Stuttering: An integrated approach to its nature and treatment. Baltimore, PA: Lippincott Williams.

Kourkounakis, T. (2021). “LibriStutter”, Borealis, V1. https://doi.org/10.5683/SP3/NKVOGQ

Park, J., & Lee, C. G. (2023). AI-based stuttering automatic classification method: Using a convolutional neural network. Phonetics and Speech Sciences, 15(4), 71-80.

일정표로 이동

BSRNN과 HiFi-GAN을 이용한 디클리핑 방법

최승운, 문승현, 최승호
서울과학기술대학교 전자공학과
seungun9275@gmail.com, jameskyer@seoultech.ac.kr, shchoi@seoultech.ac.kr

A Declipping Method Using BSRNN and HiFi-GAN

Seung-Un Choi, Seung-Hyun Moon, Seung-Ho Choi
Dept. of Electronic Engineering, Seoul National Univ. of Sci. and Tech.
seungun9275@gmail.com, jameskyer@seoultech.ac.kr, shchoi@seoultech.ac.kr

디지털 오디오 시스템에서 클리핑(clipping)은 신호의 진폭이 임계 값을 초과할 때 발생하며, 이는 청취자에게 불쾌한 소음을 유발하는 신호 왜곡을 일으킨다. 따라서 클리핑 된 부분을 복구하고 신호를 재구성하는 디클리핑(declipping) 과정이 필요하다. 기존의 심층신경망 기반 음성 향상 알고리즘에서는 주로 크기 스펙트럼을 향상시키는 방식이 사용되었고, 최근 연구에서는 위상 향상을 위한 complex ideal ratio mask와 같은 마스크 기법이 개발되었다. 하지만, 이 기법은 디클리핑을 위한 위상 재구성에 한계를 보인다. 본 논문에서 제안하는 디클리핑 방법은 이러한 마스크 기법을 각 밴드에 활용하는 BSRNN(Band-Split Recurrent Neural Network) 모델을 기반으로 하며, 위상 재구성 성능이 우수한 HiFi-GAN(Generative Adversarial Networks for efficient and High Fidelity speech synthesis)을 사용하여 음성의 품질을 향상시킨다. 실험 결과, 제안한 방법이 기존의 크기 스펙트럼 향상 방식 및 BSRNN 모델 기반 디클리핑 방식에 비해 P.835 OVRL 점수에서 더 우수한 성능을 보였다.

In digital audio systems, clipping occurs when the amplitude of a signal exceeds a threshold, causing signal distortion that produces unpleasant noise for the listener. Therefore, a declipping process is required to recover the clipped parts and reconstruct the signal. Existing deep neural network-based speech enhancement algorithms mainly enhance the magnitude spectrum, and recent research has developed masking techniques such as complex ideal ratio masks for phase enhancement. However, this technique is limited in phase reconstruction for declipping. The declipping method proposed in this paper is based on a BSRNN (Band-Split Recurrent Neural Network) model that utilizes these masking techniques for each band and uses HiFi-GAN (Generative Adversarial Networks for efficient and High Fidelity speech synthesis) with decent phase reconstruction performance to improve speech quality. Experimental results show that the proposed method outperforms the conventional magnitude spectrum enhancement method and BSRNN model-based declipping method in terms of P.835 OVRL score.

일정표로 이동

후처리 필터를 이용한 HiFi-GAN 생성 음성의 음질개선

손세현, 전민서, 황규찬, 최승호
서울과학기술대학교 전자공학과
sehyun6307@gmail.com, shchoi@seoultech.ac.kr

Audio Quality Enhancement of HiFi-GAN Generated Speech Using Post-Processing Filters

Sehyun Sohn, Minseo Jeon, Gyuchan Hwang, Seungho Choi
Dept. of Electronic Engineering, Seoul National University of Science and Technology
sehyun6307@gmail.com, shchoi@seoultech.ac.kr

TTS에 사용되는 HiFi-GAN은 멜 스펙트로그램을 입력받아 파형을 출력하지만, 일부 유성음 구간에서 고조파가 손상되며 거슬리는 잡음이 발생할 수 있다. 이러한 현상은 주로 기본 주파수가 높은 프레임에서 발생하며, 스펙트럼 분석을 통해 고조파 손상을 확인할 수 있다. 이를 해결하기 위해, voicing, pitch, correlation map sum을 사용하여 후처리할 프레임을 선택하고 고조파 복원 강도를 조절한다. 연구에서는 LP 분석을 통해 HiFi-GAN에서 생성된 음성의 잔차 신호를 고조파 강화 필터에 통과시킨 후, 다시 합성하여 손상된 고조파를 복원하였다. 이러한 과정을 거친 음성은 원래 HiFi-GAN 출력보다 더 주기적이고 선명한 음질을 가지게 된다. 음질 평가는 객관적 지표로 PESQ를 사용하여 진행되었으며, 필터 적용 후 음성의 주관적 품질이 향상됨을 확인할 수 있었다.

HiFi-GAN used in TTS receives mel-spectrogram as input and outputs a waveform, but in some voiced sound sections, harmonics are damaged and annoying noise may occur. This issue arises when the harmonic structure of certain frames is compromised, as observed in spectral analysis. To address this, the frames requiring post-processing are identified using voicing, pitch, and correlation map sum, while the extent of harmonic enhancement is controlled accordingly. LP (Linear Prediction) analysis is applied to the HiFi-GAN-generated audio, and the residual signal is passed through a harmonic enhancement filter, followed by re-synthesis of the audio. This process partially restores the damaged harmonic structure, resulting in a more periodic and cleaner signal compared to the original HiFi-GAN output. Audio quality evaluation is conducted using the PESQ as an objective metric, which shows an improvement in the perceptual quality of the processed audio.

일정표로 이동

잡음 모델링을 이용한 잡음 인지 음성 향상 기법

김장연, 신의협, 김경화, 박형민
서강대학교 인공지능학과, 서강대학교 전자공학과, 대검찰청 과학수사부, 서강대학교 전자공학과
jykim97@sogang.ac.kr, dmlguq456@naver.com, savoix@spo.go.kr, hpark@sogang.ac.kr

본 논문은 잡음 정보를 활용하는 새로운 음성 향상 기법을 제안한다. 잡음에 대한 모델링을 수행하는 모듈과, 모델링된 잡음 정보를 음성 향상 모델에 통합하는 두 단계 학습 방식을 통해, 학습 과정에서 보지 못한 잡음으로 인한 성능 저하와 음성 왜곡을 최소화한다. 실험 결과, 제안된 방법은 기존 모델에 비해 음성 품질과 잡음 억제 성능에서 우수한 결과를 보였다.

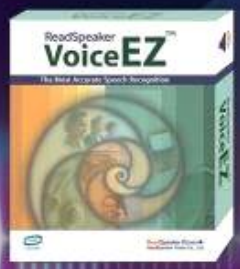
* 본 연구는 2024년 대검찰청 연구용역 “스마트폰 녹음 음성 식별을 위한 오디오 신호처리 기술개발”의 일환으로 진행되었음

글로벌 음성기술 전문회사 (주)리드스피커코리아

25년 이상의 음성기술을 향한 노하우와 기술력의 신뢰를 바탕으로
국내는 물론 해외 다양한 분야에서 인정받고 있습니다.

음성합성 ReadSpeaker™

인공지능(AI) 기술을 적용하여 음질은 한층 더 높이고
음성합성기의 개발기간은 단축시킨 음성합성기,
Standard TTS 49개 언어, 133개 음색 보유



음성인식 ReadSpeaker VoiceEz™

단어 인식부터 비정형 자연어 음성인식 강점을
모두 갖춘 최고의 음성인식 솔루션
콜센터용 STT, 음성인식자막, 음성인식회의록

(주)리드스피커코리아
www.readspeaker.co.kr
sales@readspeaker.co.kr
tel 02-3016-8500

selvas AI

음성기술

인공지능 전문 기업 **셀바스 AI**가 이끌어가겠습니다.

음성합성 / 음성인식 분야의 원천기술을 보유한 국내 대표 AI 기업 셀바스 AI,
자체 개발하여 다양한 환경에 최적화가 가능한 AI 기술로 모든 고객의 성공적인 사업을 돕겠습니다.

기업	selvasai.com
채용	selvasai.career.greetinghr.com
제품	speech.selvasai.com
이메일 문의	support@selvasai.com
전화 문의	02.852.7788

Selvy deepTTS

감정표현, 숨소리까지 재현하는 음성합성 솔루션

odiro

편리한 음성 콘텐츠 (오디오북 등) 제작 서비스

Selvy deepTTS On-Device

Android, iOS에서 네트워크 연결 없이 사용하는 고품질 음성합성 솔루션

kakaobank

ISSN 2005-9299

Proceedings

of the 2024 Autumn Conference of the Korean Society of Speech Sciences

October 19, 2024

Hankuk University of Foreign Studies

The Korean Society of Speech Sciences