ISSN 2005-9299

2020 한국음성학회 가을 학술대회

발표 논문집

주제: **speech science for HUMAN**

일시: 2020년 11월 28일(토)

장소: 온라인 (주 송출@서울대학교 인문대학 스마트 강의실)

주최: 사단법인 한국음성학회

주관: 사단법인 한국음성학회/ 연세대학교 언어정보연구원

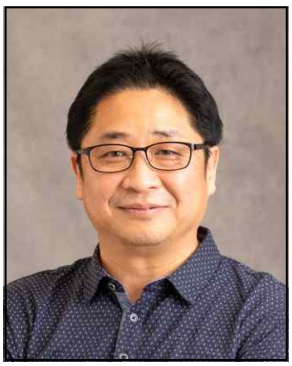
후원: 한국연구재단, 연세대학교 언어정보연구원, ㈜네이버, ㈔한국언어재활사협회, ㈜리드스피커코리아, ㈜셀바스AI, 서울대학교 인문정보연구소

한국음성학회

The Korean Society of Speech Sciences

이 발표논문집은 2020년도 정부 재원(교육부)으로 한국연구재단의 지원을 받아 발간되었음

모시는 글



안녕하십니까, 한국음성학회 모든 회원님들께 건강 안부를 이렇게 지문을 통해 여쭙게 되어 송구스럽습니다. 연초부터 전세계를 강타하고 현재도 그 기세가 여전히 살아있는 COVID-19의 공습에 우선 저희 회원 모든 분들의 건강 안전과 안녕을 진심으로 바라 마지 않습니다. 모든 분들이 각자의 위치에서 참 힘든 한해였음에 틀림없는데, 저희 학회의 운영과 연구 활동도 그러했습니다. 봄 학술대회를 개최하지 못했고, 실험음성학연구회의 하계 워크샵과 교육 분과가 주관하던 말장애 분야의 워크샵도 갖지 못했고, 발성학, 방언음성학을 주제로 야심차게 기획하였던 인문학과 음성공학의 만남 워크샵과도 잠깐 이별할 수밖에 없었습니다.

그러나 저희는 이렇게 가을 학술대회를 시작으로 서서히 그러나 여전히 조심스럽게 저희 본연의 업무인 음성 관련 연구 활동을 본격적으로 재개하며 절대로 잊지 못할 2020년도 늦가을의 하루를 그동안 힘든 여건 하에서도 강건히 이룩하신 연구 결과의 발표의 장으로 가지려고 합니다. 이번 2020년 가을 학술대회는 비대면 온라인 학술대회여서 비록 예년처럼 모두 같이 만나 열띤 토론을 하는 데는 다소나마 지장이 있겠지만 많은 분들의 노고로 치밀히 준비하였기에 참석하시는 모든 분들이 충분히 배우고, 서로의 연구에 관해 열띤 논의로 음성학 지식의 함양에 크게 이바지할 학술대회가 될 것임에 틀림없습니다.

이렇게 힘든 시기에 학술대회를 준비하시고 이끌어주신 성철재 조직위원장님과 훌륭한 학술대회 진행 본부를 마련해 주신 정민화 교수님, 비대면 줌 회의, 이메일 등을 통해 알찬 의견을 주신 학회 모든 임원진분들, 무엇보다도 발표와 토론의 장에 적극적으로 참여해 주시는 모든 회원분들께 이 자리를 빌어 감사의 마음을 전합니다. 아울러 어려운 상황에서도 늘 저희 학회를 물심양면 후원해 주신 한국연구재단, 연세대학교 문과대학, ㈜네이버, ㈔한국언어재활사협회, ㈜리드스피커코리아, ㈜셀바스AI, 서울대학교 인문정보연구소의 응원과 후원에 감사드립니다.

아무쪼록 모두 건강하고 안전하게 본 학술대회에 참여해 주시기 바랍니다. 회원님들의 건승을 빕니다. 감사합니다.

2020년 11월 28일 한국음성학회 회장

이 석 재

2020 한국음성학회 가을 학술대회 준비위원회

학술대회장	회장	이석재(연세대)
조직위원회	위원장	박한상(홍익대)
	위원	성철재(충남대), 윤원희(계명대), 정민화(서울대), 홍기형(성신여대), 최성희(대구가톨릭대), 공은정(항공대), 박상희(대구사이버대)
학술위원회	위원장	성철재(충남대)
	위원	안현기(서울대), 오재혁(건국대), 박기영(ETRI), 남호성(고려대), 김성태(동신대), 박형민(서강대), 김지연(우송대)

2020 한국음성학회 가을 학술대회 일정표

11월 28일(토) 온라인 (zoom Webinar, 주송출@서울대학교 인문대학 스마트강의실)

시 간	발 표 및 내 용
10:30~11:10	**음성학 구두 발표(질문 & 답변) I** 좌장: 윤태진(성신여대)
11:15~11:25	**개회식** 사회: 박한상(홍익대)
11:30~12:00	**말장애 구두 발표(질문 & 답변)** 좌장: 김재옥(강남대)
12:00~13:00	점 심
13:00~13:30	연구윤리교육 강사: 안동형(국가과학기술인력개발원)
13:40~14:30	**음성학 구두 발표(질문 & 답변) II** 좌장: 고언숙(조선대)
14:40~15:30	**음성공학 구두 발표(질문 & 답변)** 좌장: 박정식(한국외대)
15:30~16:00	휴식
16:00~16:40	**초청연사 특강** 음성언어 지능의 현재와 미래 (심층학습 기반 종단간 처리 기술을 중심으로) 강사: 김회린(KAIST) 사회: 성철재(충남대)
16:50~17:20	**총회** 사회: 공은정(항공대)
17:20~17:50	우수발표시상 및 폐회식 사회: 공은정(항공대)

2020 한국음성학회 가을 학술대회 세부 일정표

[포스터 발표]

음성공학

좌장: 박기영(한국전자통신연구원)

구분	제목	저자
P01	음성감정 인식에서 운율 특징의 영향분석	박순찬, 김형순(부산대)
P02	원거리 화자인증 시스템을 위해 선택적으로 특징을 보상하는 분별적 잡음 제거 자기부호화기	김주호, 정지원, 심혜진, 유하진(서울시립대)
P03	Average modeling 및 GST를 이용한 다화자 비병렬 음색변환	엄지섭, 김회린(KAIST)
P04	고속 한국어 종단형 음성인식 시스템	박기영, 오유리, 박전규(한국전자통신연구원)
P05	Frequency-level feature for rapid emotion recognition	Di(Suk Young) Lim(Hanyang Univ.), Hong In Yoon, Hyeong Ju Na, Jeong-Sik Park(Hankuk Univ. of Foreign Studies)
P06	지식증류 기법을 사용한 감정을 포함하는 음성인식의 적응 학습	윤홍인(한국외대), 임석영(한양대), 나형주, 박정식(한국외대)
P07	Transformer 기반 종단형 음성인식 시스템에서의 CTC 활용	오유리, 박기영, 박전규(한국전자통신연구원)
P08	강인 음성인식을 위한 유도 어텐션 기반의 시청각 음성인식	이용혁, 김재빈, 박형민(서강대)
P09	공감형 대화 음성 챗봇 개발을 위한 공감 상황 인지 코퍼스 구축	김종인, 정민화(서울대)
P10	실시간 음성인식을 활용한 임베디드 한국어 음성 대화 시스템	손현수, 박호성, 김지환(서강대)
P11	멀티태스크 U-Net과 채널 어텐션 기반의 다채널 음성 향상	이건우, 김홍국(광주과학기술원)
P12	다화자 동시 발화 환경 diarization을 위한 확장된 personal VAD	이승형, 한민수(KAIST)
P13	범주적 지각 음성인식 실험을 통한 인공신경망 모델의 설명가능성 접근	이호진(서울대)

[포스터 발표]

말장애 및 음성의학

좌장: 이수복(우송대)

구분	제목	저자
P14	말더듬 성인의 취업 및 직장생활 경험에 대한 질적 연구	박소현(충남대), 박진(가톨릭관동대)
P15	좌반구 손상과 우반구 손상 뇌졸중 환자의 의문문 유형에 따른 운율 특성 비교	유영미, 박소현, 성철재(충남대)
P16	과기능적 발성장애 환자의 후두과긴장 특징	김지성(충북대병원), 최성희(대구가톨릭대) 이동욱(충북대의대)
P17	서비스 제공자의 말소리 친절도 평가 및 분석	장나리, 성철재(충남대)
P18	발화속도에 따른 노년층 기준연령 설정	조보라, 성철재(충남대)
P19	5년 이상 한국 거주 외국인 여성 발화에 대한 용인도와 명료도에 관한 연구	박상희(대구사이버대), 이상훈(구미전자정보기술원)
P20	다문화 가정 이주여성들의 한국어 파열음 청지각 평가	박선영, 성철재(충남대)
P21	말소리가 제한된 아동을 위한 말리듬을 이용한 난타 프로그램의 개발과 적용	박영혜(아이&맘 언어심리발달센터), 최성희, 최철희(대구가톨릭대)
P22	수위 센서를 이용한 음성치료용 물컵 장치 개발	최성희, 최혜림, 임금별, 김신혜, 윤종인(대구가톨릭대)

음성학

좌장: 손민정(한남대)

구분	제목	저자
P23	영어 원어민의 영어 어휘 강세 발화에 대한 음향적 분석	유지윤, 이석재(연세대)
P24	한국 초등학생의 L2 영어 낭독 발화 평가에서 분절적・초분절적 자질이 청자 이해도에 미치는 영향	차호윤, 이석재(연세대)
P25	Durations of two English stops across word boundaries	Yungdo Yun(Dongguk Univ.)
P26	응급의료영역 음성대화 데이터 구축	이주영, 최서경, 지승훈, 강지민, 김종인, 김도희, 김보령, 조은기, 김호정, 장정민(서울대), 김준형, 구본혁, 박형민(서강대), 김선희, 정민화(서울대)
P27	영-한 음차표기에서 나타나는 음운현상 분석	전지현, 이석재(연세대)
P28	한국인 학습자 발화 프랑스어 구강 및 비강 모음에 대한 음향적 특징 연구	김도희(서울대), 윤상아(한국외대), 김선희(서울대)

[구두 발표]

음성학 I

좌장: 윤태진(성신여대)

시간	구분	제목	저자
10:30 ~ 11:10	PH1	경북방언 wh-의문문의 작용역에 따른 운율 구조의 변화	윤원희(계명대)
	PH2	발화 상대자의 성별·연령이 발화자의 발화량에 미치는 영향	정희재, 신지영(고려대)
	PH3	Acoustics of noise-adapted and clear speech in individuals with elevated depressive symptoms	Hoyoung Yi(Texas Tech Univ.), Rajka Smiljanic (Univ. of Texas@Austin)
	PH4	한국인 영어학습자의 단어 및 어구 강세-비강세 음절의 모음 길이 구현과 발음 평가 등급과의 관계	박기훈, 박혜숙, 이석재(연세대)
	PH5	프랑스인 한국어 학습자의 비음성 연구 – 초성 비음 /ㄴ/을 중심으로	이보람(Sorbonne-Nouvelle)

말장애 및 음성의학

좌장: 김재옥(강남대)

시간	구분	제목	저자
11:30 ~ 12:00	SD1	심한 말소리 장애 아동에게 적용한 통합치료접근법이 말소리 비일관성과 음운산출능력에 미치는 효과	고유경(바른소리언어치료센터), 김수진(나사렛대)
	SD2	20-30대 베트남 결혼 이주 여성들의 한국어 운율특성	김난숙(충남대 언어치료센터), 성철재(충남대)
	SD3	뇌성마비로 인한 마비말장애 성인의 자음정확도와 말명료도 비교 연구	여은정, 정민화(서울대)
	SD4	공명튜브발성 시 물의 저항에 따른 성대진동특성	채혜림, 최성희(대구가톨릭대)
	SD5	Python을 이용한 아동용 단모음 조음 훈련 프로그램	김하정, 성철재(충남대)

[구두발표]

음성학 II

좌장: 고언숙(조선대)

시간	구분	제목	저자
13:40 ~ 14:30	PH6	후설모음 /ㅡ, ㅜ, ㅓ, ㅗ/에 대한 음향음성학적 연구: 선행 인접 자음과의 상관성을 중심으로	조남민(한국기술교육대), 황미경(이화여대)
	PH7	한국어 단모음의 RNR 평균 비교	윤지현(충남대), Jiayin Gao(Univ. of Edinburgh), Takayuki Arai(Sophia Univ.) 성철재(충남대)
	PH8	L2 영어 말하기 유형별(즉흥자유체 vs. 낭독체) 발화평가 채점 세부 요인이 말하기 평가 총체적 점수에 미치는 영향에 대한 비교	이석재(연세대)
	PH9	한국어 방언 자동 식별을 위한 방언 특징 분석	이주영(서울대), 김경화(대검찰청), 양승희, 정민화(서울대)
	PH10	'동의'의 정도에 따른 운율 실현에 관한 연구	양봉석, 신지영(고려대)

음성공학

좌장: 박정식(한국외대)

시간	구분	제목	저자
14:40 ~ 15:40	SE1	GAN 기반 신경망 보코더의 비교평가 및 음질 개선방안 연구	서영주, 최연주, 엄지섭, 정성희, 김회린(KAIST)
	SE2	합성곱 신경망을 이용한 화자 검증에서 시계열 정보의 활용도 분석	허정우, 심혜진, 정지원, 김주호, 유하진(서울시립대)
	SE3	Including linguistic knowledge in an auxiliary classifier cycleGAN for corrective feedback generation	Seung Hee Yang, Minhwa Chung(Seoul Nat'l Univ.)
	SE4	연속음성에서의 Jitter 측정 방법	조철우(창원대)

특강

사회: 성철재(충남대)

시간	내용
16:00~16:40	음성언어 지능의 현재와 미래 (심층학습 기반 종단간 처리 기술을 중심으로) 강사: 김회린(KAIST)

차 례

특강

포스터발표 1 (음성공학)

포스터 발표 II (말장애 및 음성의학)

포스터 발표 III (음성학)

구두 발표 I (음성학 I)

구두 발표 II (말장애 및 음성의학)

구두 발표 III (음성학 II)

구두 발표 IV (음성공학)

특강

사회: 성철재(충남대)

음성언어 지능의 현재와 미래
(김회린, KAIST)

한국음성학회 가을학술대회 강연

음성언어 지능의 현재와 미래

(심층학습 기반 종단간 처리 기술을 중심으로)

2020. 11. 28

KAIST 전기및전자공학부 김 회 린

Statistical Speech & Sound Computing Lab.

KAIST Korea Advanced Institute of Science and Technology

Table of Contents

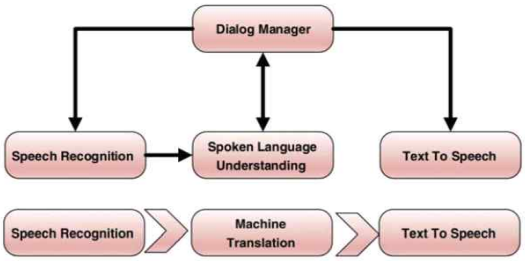
- ASR and End-to-End Approach
- TTS and End-to-End Approach
- Revisit to Speech Chain and End-to-End Approach

2

Statistical Speech & Sound Computing Lab.

Automatic Speech Recognition

- Definition
  - From speech signal to word sequence
    - Input : segmented audio signal containing a sentence speech
    - Output : most likely word sequence
  - Front-end module for spoken language system / speech translation system

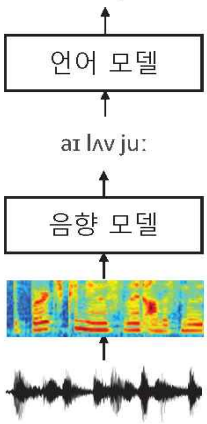


(Figure courtesy of Deng [1])

3

Automatic Speech Recognition

- Classical ASR 구조



단어 단위 문장 출력

Pronunciation dictionary,
N-gram, RNN-LM, ...

음소 단위 출력

GMM-HMM
(Gaussian Mixture Model – Hidden Markov Model)

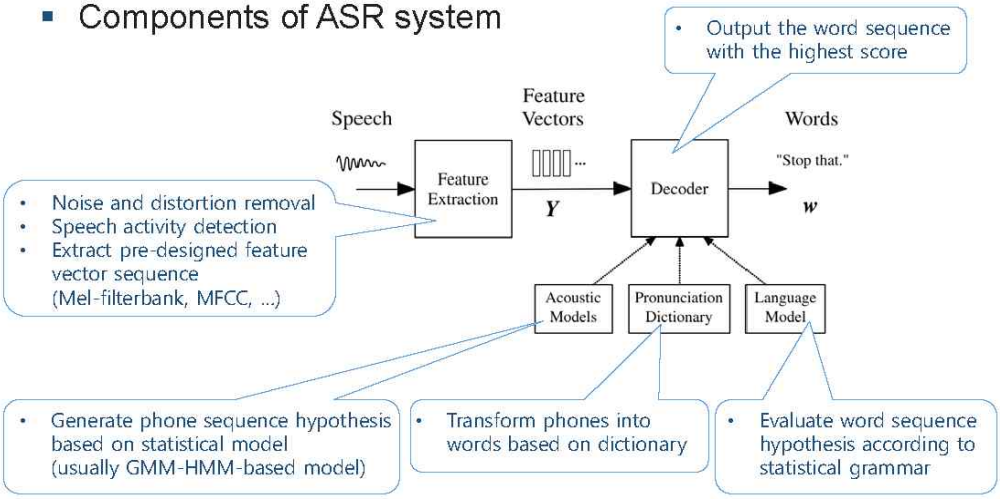
음성 특징 추출 (Mel-filterbank, MFCC)

음성 신호 입력

4

Automatic Speech Recognition

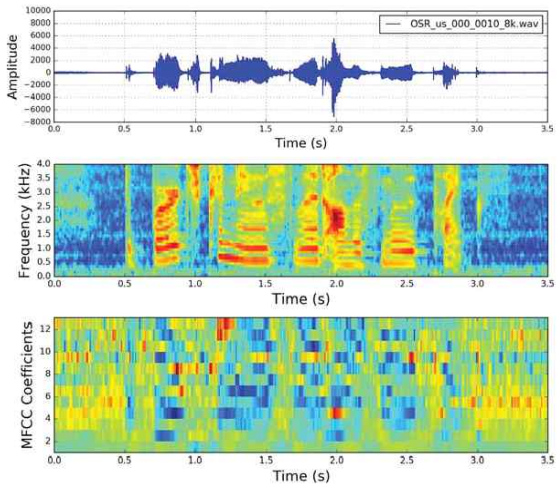
- Components of ASR system



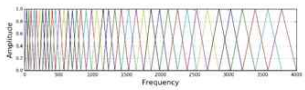
(Figure courtesy of Young [2])

Acoustic Feature Extraction

- Waveform, Mel-filterbank, and MFCC



- Short-time (~10ms) Fourier Transform
- Triangular filter bank with human perception-inspired frequency scale (Mel-scale filter bank, 40-dim.)

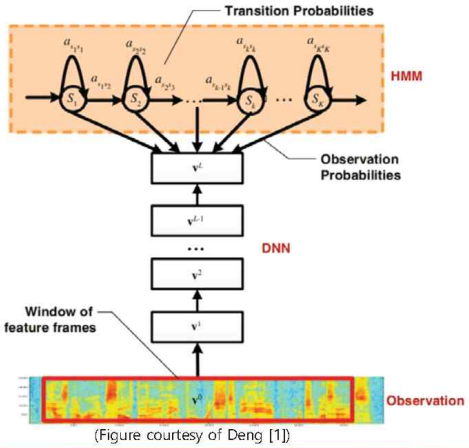


- Discrete Cosine Transform (Type-II)
- To de-correlate the filter bank coefficients and yield a compressed representation (typically 13-dim.)

All image courtesy of haythamfayek.com [15]

Acoustic Model

- DNN-HMM hybrid system
  - Hidden Markov Model (HMM), which has been used from the classical system, still plays an important role in the modern system.

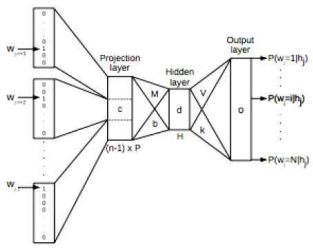


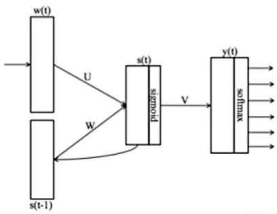
(Figure courtesy of Deng [1])

Language Model

- N-gram language model
  - Example
    - Bigram : $p(\text{hot}|\text{Strike, while, the, iron, is}) \approx p(\text{hot}|\text{is})$
    - Trigram : $p(\text{hot}|\text{Strike, while, the, iron, is}) \approx p(\text{hot}|\text{iron,is})$
  - Context scope problem

- Neural network language model





All image courtesy of shuuki4.wordpress.com [16]

Performance Measure

- What is word error rate?
  - Word Error Rate (WER) is the academic and industrial standard to measure ASR model accuracy.

$$\mathrm{WER} = \frac{\mathrm{ins} + \mathrm{del} + \mathrm{sub}}{N} \times 100(\%)$$

  - N : Total number of words according to true (human-labeled) transcript
  - ins : Insertion error. Number of words that are incorrectly added.
  - del : Deletion error. Number of words that are not detected.
  - sub : Substitution error. Number of words that are substituted between reference and hypothesis.

Performance Measure

- WER counting example

  - True transcript : How $\overbrace{\text{are}}^{\text{del}}$ you today John
  - Decoding result : How you $\underbrace{\text{a}}_{\text{ins}}$ today $\underbrace{\text{Jones}}_{\text{sub}}$

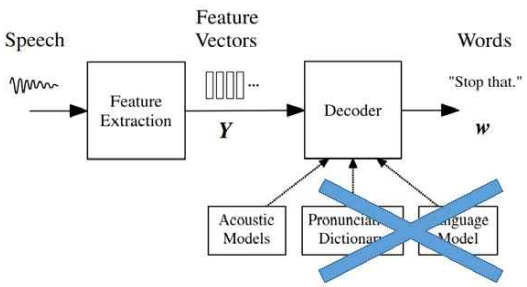
$$\begin{aligned}\mathrm{WER} &= \frac{\mathrm{ins} + \mathrm{del} + \mathrm{sub}}{N} \times 100(\%) \\ &= \frac{1+1+1}{5} \times 100 = 60(\%)\end{aligned}$$

Problem of the ASR System

- Two demerits
  - Conditional independence assumption
    - The HMM-based model assumes that the output of each frame has nothing to do together with each other.
    - This is essential for model construction and training. However, it does not match the actual situation of ASR.
  - Training process is complex and difficult to be globally optimized.
    - Each module should be independently optimized with their own objectives.
    - Lexicon that is essential for overall system requires expert knowledge that must be defined in advance.

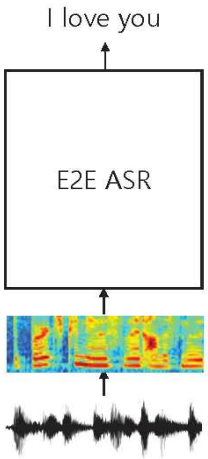
End-to-End Approach

- Main purpose
  - Overcoming the disadvantages of the existing ASR system
  - One system that directly connects audio signal and output sentence (in end-to-end manner) without individual modules
  - Currently, thanks to the explosive increase in speech data and computational power, many ASR systems are known to be implemented with E2E-ASR.



End-to-End Approach

- E2E ASR 구조



문자/단어 단위 문장 출력

CTC-based Model,
Attention-based Model,
...

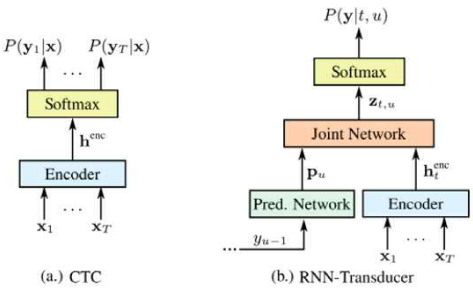
음성 특징 추출 (Mel-filterbank, MFCC)

음성 신호 입력

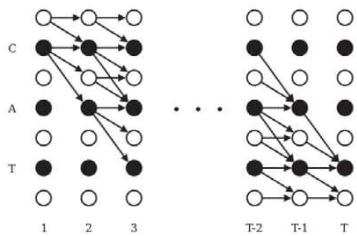
13

End-to-End Approach

- CTC-based model
  - Connectionist Temporal Classification: The first suggestion for E2E-ASR framework.
  - Dynamic programming is also used for I/O length compensation.



(Figure courtesy of Prabhavalkar [4])

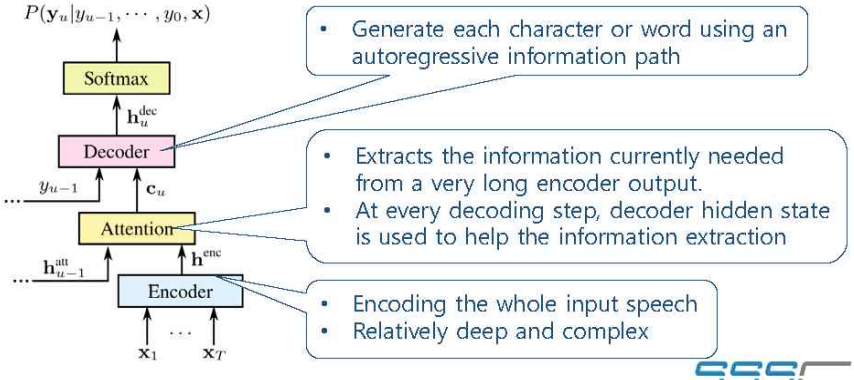


(Figure courtesy of Graves [5])

14

End-to-End Approach

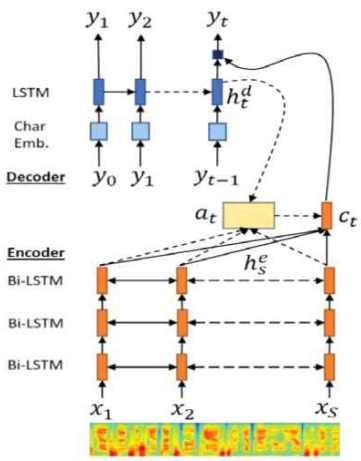
- Attention-based model
  - A method inspired from the neural machine translation
  - Widely used model because of its relatively good performance



(Figure courtesy of Prabhavalkar [4])

End-to-End Approach

- E2E ASR 모델 구현 사례



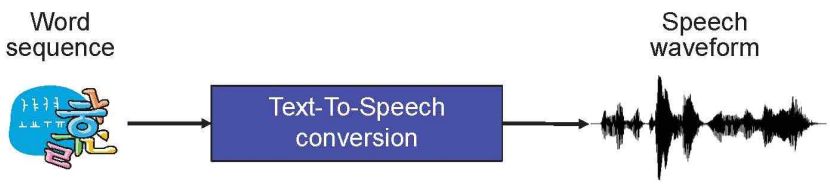
TTS and End-to-End Approach

- Introduction
- DNN-based speech synthesis
- End-to-end speech synthesis

17

Introduction

- 음성합성 기술
  - 문자열 혹은 단어열(word sequence)을 해당 음성파형으로 변환하는 기술
    → 문자음성변환 (Text-To-Speech conversion: TTS) 기술



18

KAIST Korea Advanced Institute of Science and Technology

Introduction

- 전통적 음성합성 방식
  - 포만트 음성합성 (Formant speech synthesis)
    - 전처리부(텍스트분석기) + 포만트 보코더(주파수영역 auditory parameter)
  - 조음 음성합성 (Articulatory speech synthesis)
    - 전처리부(텍스트분석기) + LPC 보코더(시간영역 production parameter)
  - 파형연결 음성합성 (Waveform concatenation speech synthesis)
    - 전처리부(텍스트분석기) + 음편선정/연결(Waveform unit selection/concatenation)
  - 통계모델 (은닉마코프모델) 기반 음성합성 (Statistical model (HMM) based speech synthesis): HTS
    - 전처리부(텍스트분석기) + GMM-HMM + 보코더
      - Merlin tool kit: TN + G2P + CDDT + GMM-HMM + 보코더
      - GMM-HMM → DNN, LSTM
      - 보코더: STRAIGHT [Kawahara, 1999], WORLD [Morise, 2016], WaveGlow [Prenger, 2018]

19

Statistical Speech & Sound Computing Lab.

KAIST Korea Advanced Institute of Science and Technology

Introduction

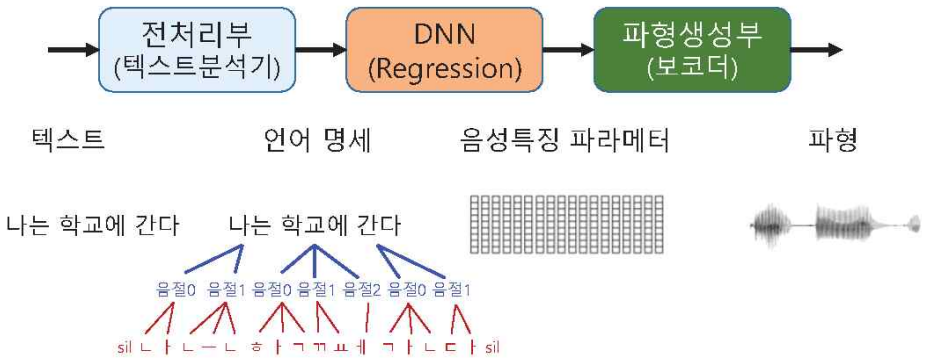
- 최근 음성합성 방식
  - **딥러닝 기반 음성합성 (**Deep learning-based speech synthesis)
    - **고전적 모듈 방식 음성합성** (Classical modular-type speech synthesis)
      - 전처리부(텍스트분석기): DNN (DeepVoice1)
      - Regression: DNN, CNN, RNN, LSTM
      - Vocoder: Griffin-Lim, WORLD, STRAIGHT, WaveNet, WaveGlow, WaveRNN alternative
        - 합성속도: WaveNet (실시간 150배), WaveGlow (실시간 2배), WaveRNN (실시간 10배)
    - **최신 종단간 방식 음성합성** (Recent end-to-end speech synthesis)
      - DeepMind WaveNet: 텍스트분석기 + Regression/Vocoder (WaveNet)
      - Google Tacotron1/2: 텍스트분석기/Regression (CBHG, LSTM) + Vocoder (GL, WaveNet)
      - USTC Transformer TTS: 텍스트분석기/Regression + Vocoder (WaveNet)
      - Baidu DeepVoice2/3: 텍스트분석기/Regression + Vocoder (GL, WORLD, WaveNet)
      - Recent neural vocoders: MelGAN, Parallel WaveGAN, Multi-band MelGAN, …

20

Statistical Speech & Sound Computing Lab.

DNN-based speech synthesis

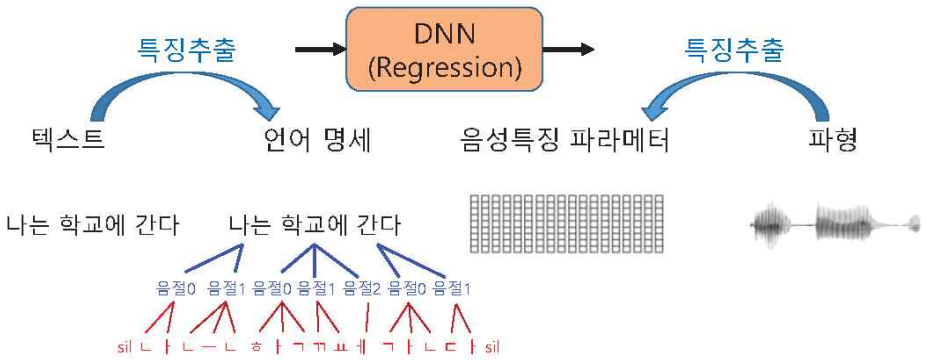
- 기본 구조



21

DNN-based speech synthesis

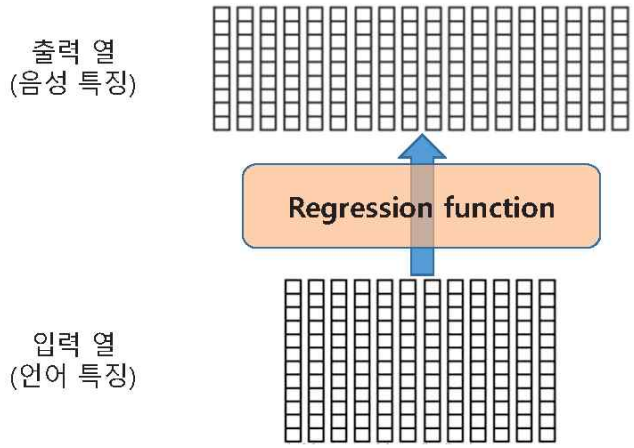
- 기본 구조



22

DNN-based speech synthesis

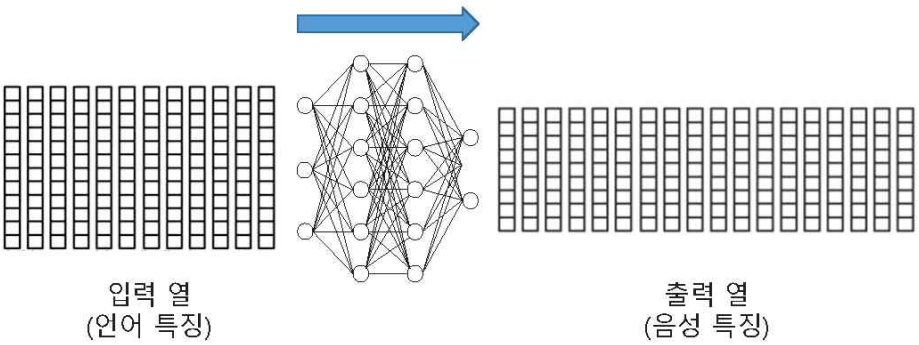
- 음성합성 기술 접근 원리
  - Sequence-to-sequence regression problem



23

DNN-based speech synthesis

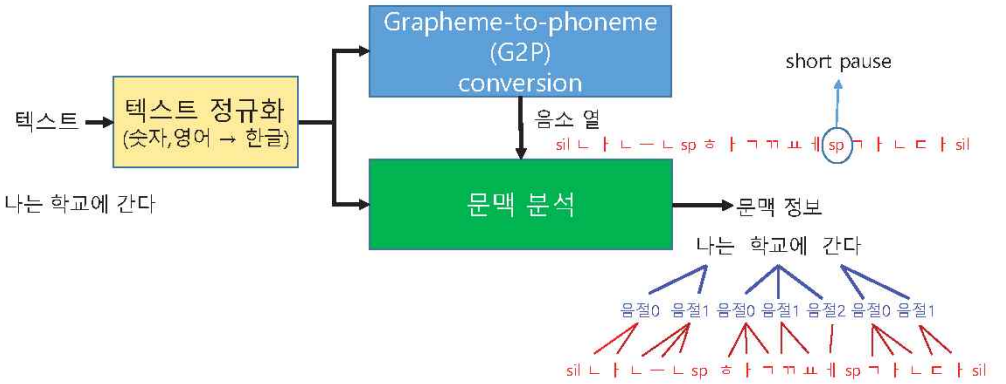
- 음성합성 기술 접근 원리
  - Sequence-to-sequence regression problem



24

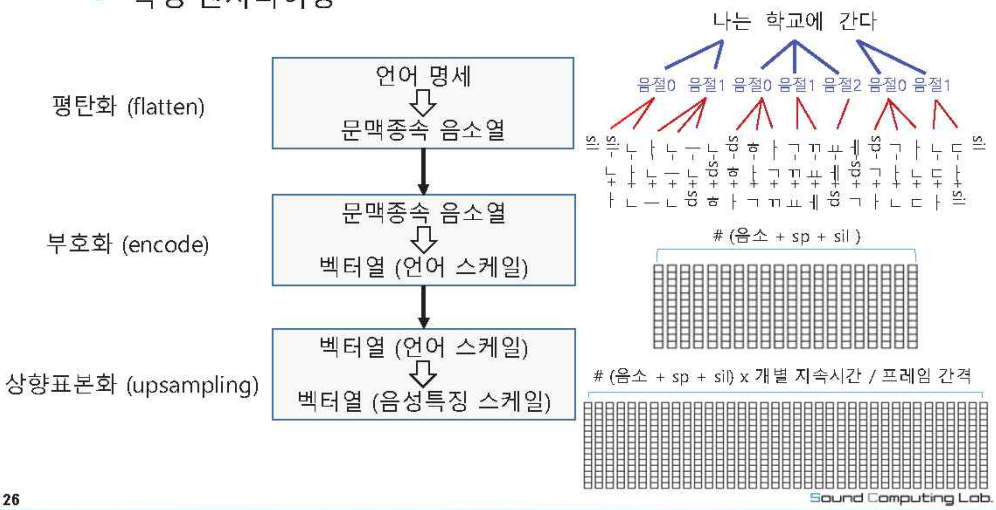
DNN-based speech synthesis

- 전처리부
  - 특징 추출



DNN-based speech synthesis

- 전처리부
  - 특징 엔지니어링



DNN-based speech synthesis

- Regression
  - Two steps in speech synthesis
    - 음소 지속시간 모델링 및 추정
    - 음소 지속시간 내에서의 음성특징 파라메터열 모델링 및 추정

    → 두 개의 독립적인 DNN 으로 regression 모델링

KAIST Korea Advanced Institute of Science and Technology

DNN-based speech synthesis

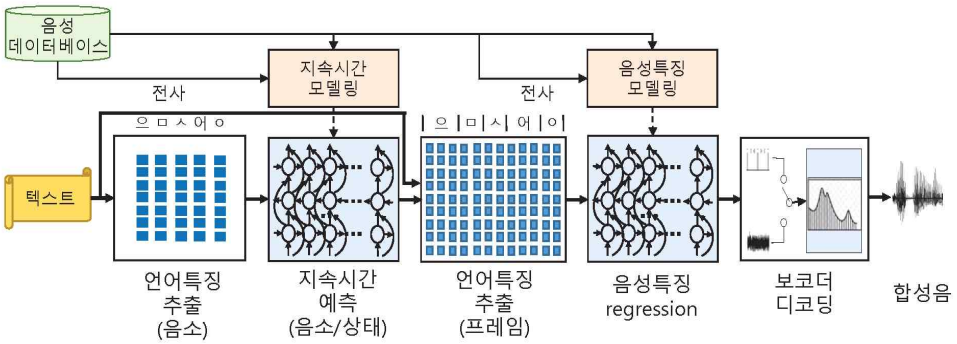
- 파형생성부 (보코더)
  - 기능
    - 음성특징 파라메터열 → 음성파형
  - 종류
    - LPC 보코더
    - WORLD 보코더
    - STRAIGHT 보코더
    - WaveNet 보코더

    저계산량 ⇕ 고품질 합성음

Statistical Speech & Sound Computing Lab.

DNN-based speech synthesis

- 음성합성기 구조



End-to-end speech synthesis

- 배경
  - Classic speech synthesis pipelines are complex.
    - Multi-stages: 전처리부, regression(회귀부), 파형생성부
  - Each component is based on extensive domain expertise.
    - Hard to design

  → Needs substantial engineering effort when building new system
- 개념
  - Synthesizes speech directly from characters

  → Single-stage system

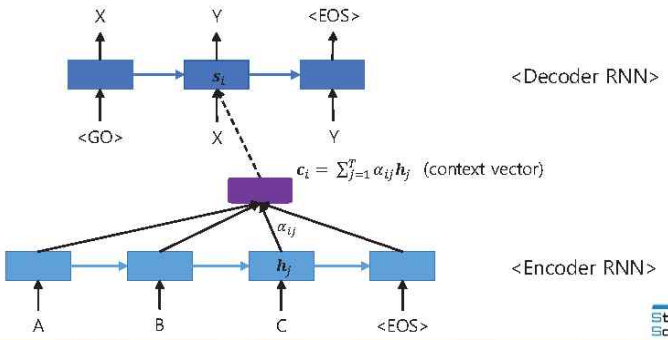
End-to-end speech synthesis

- 장점
  - Given <text, audio> pairs, the model can be trained completely from scratch with random initialization
  - Alleviates laborious feature engineering
  - Allows easily for rich conditioning on various attributes, such as speaker or language, or high-level features like sentiment
  - Single model is likely to be more robust than a multi-stage model where each component's errors can compound
- 단점
  - Needs larger amount of speech data
  - Unexpected synthesis errors can be found in synthesized speech

End-to-end speech synthesis

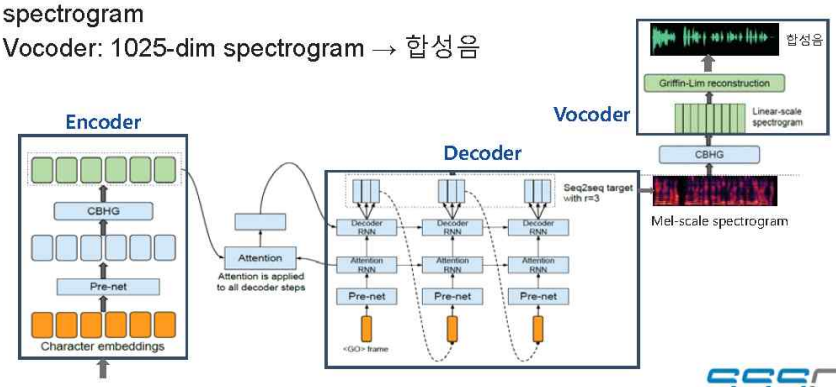
- Tacotron (Wang, 2017, Google)
  - Sequence-to-sequence
    - Encodes character sequence and decodes to speech sequence
    - Shows limited performance
    - Suffers from alignment problem due to input-output length mismatch

  → Sequence-to-sequence with attention approach



End-to-end speech synthesis

- Tacotron1
  - Encoder: 문자열 (grapheme) → 내부 학습된 형태
  - Decoder(+attention): 내부 학습된 형태 → 80-band mel-scale spectrogram
  - Post-processing: 80-band mel-scale spectrogram → 1025-dim linear-scale spectrogram
  - Vocoder: 1025-dim spectrogram → 합성음

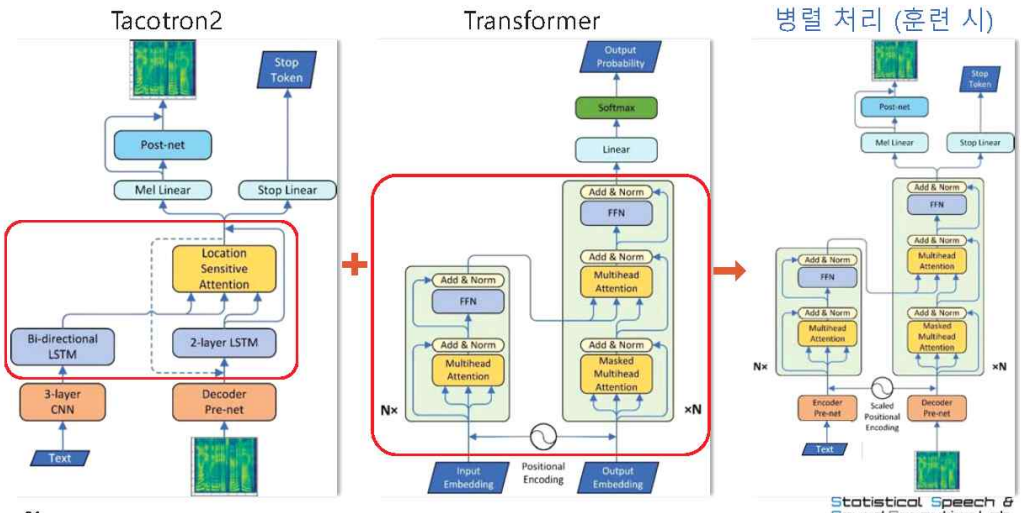


33 CBHG: 1D Convolution Bank + Highway network + bidirectional GRU module

KAIST Korea Advanced Institute of Science and Technology

End-to-end speech synthesis

- Transformer TTS

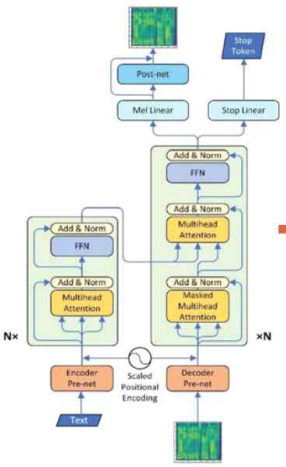


34

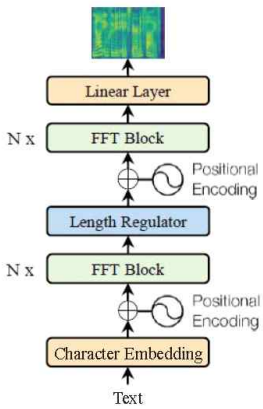
End-to-end speech synthesis

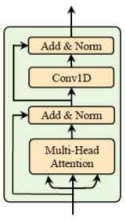
- FastSpeech

Transformer TTS

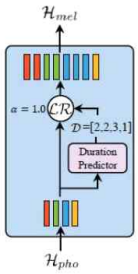


FastSpeech (병렬 처리)

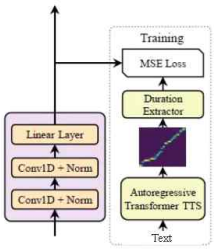




FFT Block



Length Regulator

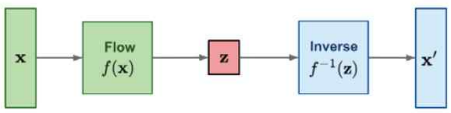


Duration Predictor

KAIST Korea Advanced Institute of Science and Technology

End-to-end speech synthesis

- DNN 기반 보코더 개선
  - WaveGlow (Prenger et al., 2018)

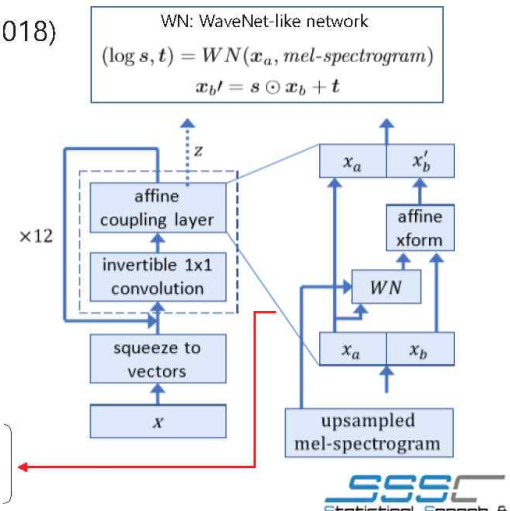


$$\mathrm{z} = f(\mathrm{x}) = f_k \circ f_{k-1} \circ \cdots f_1(\mathrm{x})$$

with $p_Z(Z) = \frac{p_X(x)}{|J(x)|}$, $z \sim \mathcal{N}(z; 0, \boldsymbol{I})$

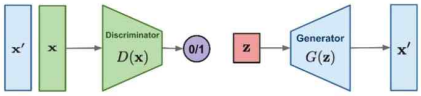
$$\log p_\theta(\mathrm{x}) = \log p_\theta(\mathrm{z}) + \sum_{i=1}^{k} \log|J(f_i(x))|$$

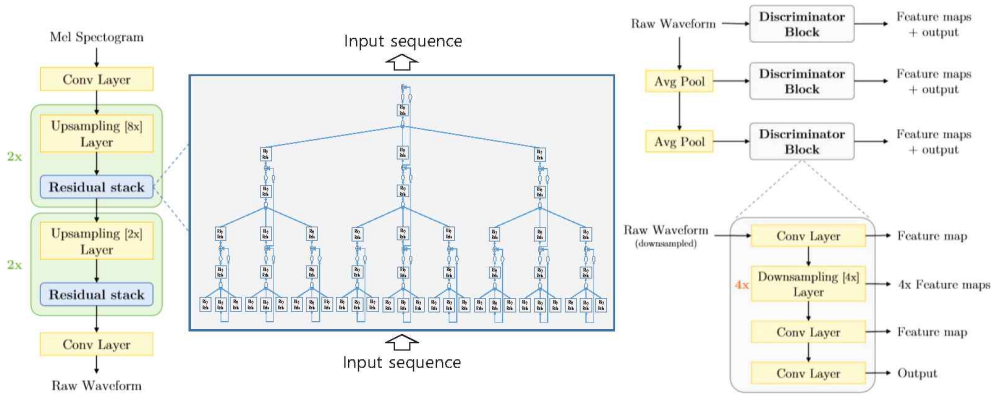
$$\begin{pmatrix} x_a \\ x_b' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \mathrm{WN}(\cdot) & \mathrm{s} \end{pmatrix} \begin{pmatrix} x_a \\ x_b \end{pmatrix} + \begin{pmatrix} 0 \\ t \end{pmatrix}$$



End-to-end speech synthesis

- DNN 기반 보코더 개선
  - MelGAN (Kumar et al., 2019)



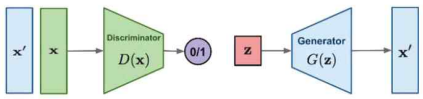


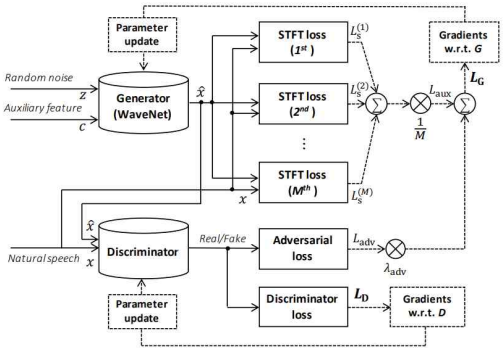
37 (a) Generator (b) Discriminator

End-to-end speech synthesis

- DNN 기반 보코더 개선
  - Parallel WaveGAN (Naver, 2019)
    - GAN 기반 실시간 보코더
    - 고음질
    - 돌발잡음

→ GAN 네트워크 구조 변경 및 조율(tuning)로 돌발잡음 제거 시도





38

End-to-end speech synthesis

- Multi-band MelGAN (Northwestern Polytechnical University, 2020)
  - Multi-band MelGAN
    - Smaller footprint
    - Faster decoding with CPU

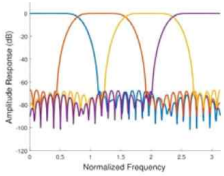


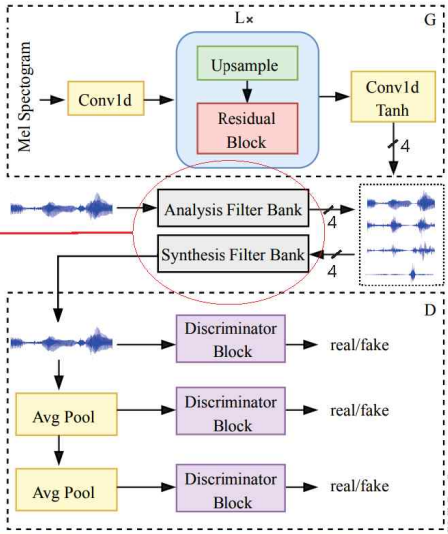
Figure 5: Frequency response of 4-band Pseudo-QMF filter banks.

Pseudo QMF

Index	Model	Loss	MOS
M1	MB-MelGAN	L_{full} (Eq. (7))	4.22±0.04
M2	MB-MelGAN	$L_{full} + L_{sub}$ (Eq. (9))	**4.34±0.03**

Table 5: *Model complexity.*

Index	Model	GFLOPS	#Paras. (M)	RTF
F0	MelGAN [20]	5.85	4.27	0.2
F3	FB-MelGAN	7.60	4.87	0.22
M2	MB-MelGAN	**0.95**	**1.91**	**0.03**



39

End-to-end speech synthesis

- DNN 기반 보코더 음질 시연

Real Speech	MelGAN Stacks 3	MelGAN Stacks 4	Multi-band MelGAN	Parallel WaveGAN Gen layers 30	Parallel WaveGAN Gen layers 33

40

End-to-end speech synthesis

- End-to-End 방식 음성합성 음질 시연
  - 음성합성기 성능 (MOS)
    - 평가 대상: 4 종류의 음성합성기
    - 평가 스크립트: 훈련 데이터에 없는 25문장
    - 평가자: 10명

((char)Transformer
+ WaveGlow)

Seq2seq model	(char) Transformer	(char) FastSpeech	(G2P) Transformer	(phn) Tacotron2
Neural vocoder	ParallelWaveGAN			WaveGlow
MOS	**4.55**	4.20	4.49	3.97

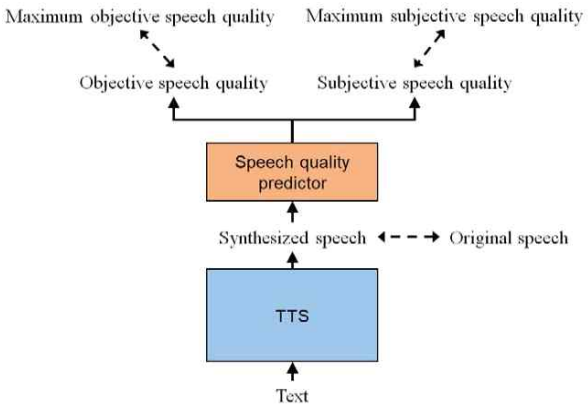
(cf.)

(원음)

41

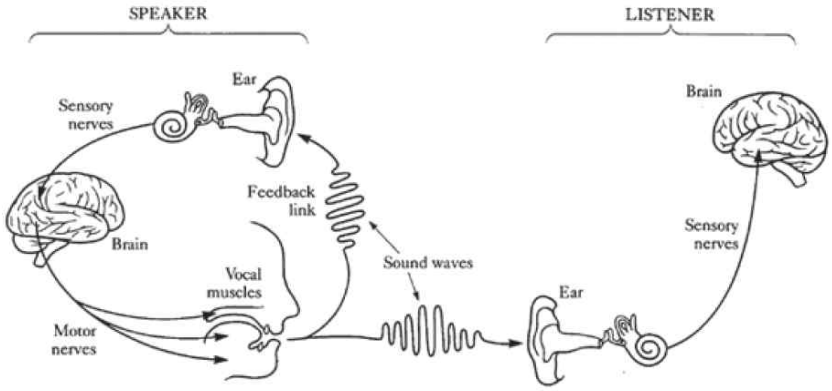
End-to-end speech synthesis

- Improvement of E2E TTS
  - E2E TTS model combined with speech quality predictors



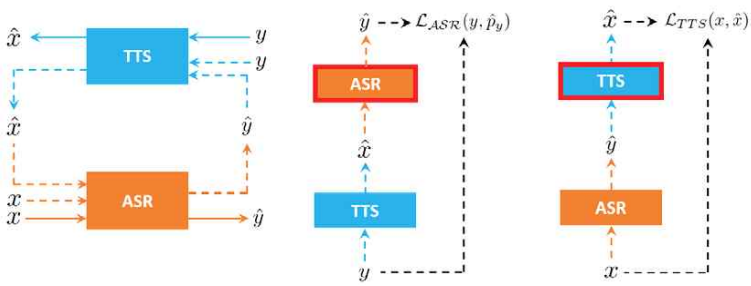
42

Revisit to Speech Chain and End-to-End Approach



Revisit to Speech Chain and End-to-End Approach

- Unified model for E2E ASR and TTS using speech chain interaction



(a) (b) (c)

Thanks

Q & A

45

Statistical Speech & Sound Computing Lab.

포스터 발표 I

음성공학

좌장: 박기영(한국전자통신연구원)

음성 감정 인식에서 운율 특징의 영향 분석

박 순 찬, 김 형 순
부산대학교 전자공학과

Analysis of the Effect of Prosodic Feature in Speech Emotion Recognition

Sunchan Park, Hyung Soon Kim
Dept. of Electronics Engineering, Pusan National University
sunchanpark@pusan.ac.kr, kimhs@pusan.ac.kr

기본 주파수, 에너지의 변화와 같은 운율 특성은 음성을 통한 감정 인식에서 중요한 정보로 여겨진다. 본 연구에서는 대표적인 음성 특징인 Mel-Frequency Cepstral Coefficient(MFCC)와 Log Mel-Filterbank Energy(LMFE)에 운율 특징을 추가로 적용했을 때 감정 인식 성능에 미치는 영향을 분석하였다. 충분히 많은 수의 필터를 통해 얻은 LMFE에는 포먼트 주파수와 함께 기본 주파수 F0가 드러나지만, MFCC에서는 F0의 정보가 제거되고 포먼트 주파수의 정보만 남는다. 일반적으로 운율 특징을 추가할 경우 성능이 개선될 것으로 기대되나, 신경망이 LMFE로부터 F0에 대한 정보를 추출할 수 있다면 운율 특징 적용에 의한 성능 개선 효과가 미미할 것이다. 성능 비교를 위해 각 특징을 입력으로 하는 convolution layer, long-short term memory, attention module로 구성된 심층 신경망 기반 감정 인식 시스템을 구현하였다. 운율 특징은 log F0, log 에너지와 그 델타 파라미터로 구성되며, 보다 정확한 F0를 추정하기 위해 신경망 기반의 CREPE [1] 모델을 사용하였다. 훈련 및 평가에는 The Interactive Emotional Dyadic Motion Capture (IEMOCAP) 데이터베이스에서 기쁨, 슬픔, 화남, 중립의 네 가지 감정 범주의 음성을 추출하여 이를 통해 훈련 및 평가하였다. 실험 결과 운율 특징을 MFCC에 추가한 경우 정확도가 3.95% 향상되었지만, LMFE에 추가한 경우 정확도는 0.42% 향상에 그쳤다. 그러나 LMFE를 단독으로 사용한 경우에도 MFCC와 운율 특징을 함께 사용한 경우보다 정확도가 0.76% 높았다. 위 실험 결과를 통해 충분히 높은 해상도의 LMFE로부터 신경망이 운율 정보를 학습할 수 있음을 확인할 수 있었다.

감사의 글

이 논문은 2019년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2019S1A5A2A03045884)

참고문헌

[1] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, “CREPE: A convolutional representation for pitch estimation,” in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Feb. 2018, pp. 161 - 165.

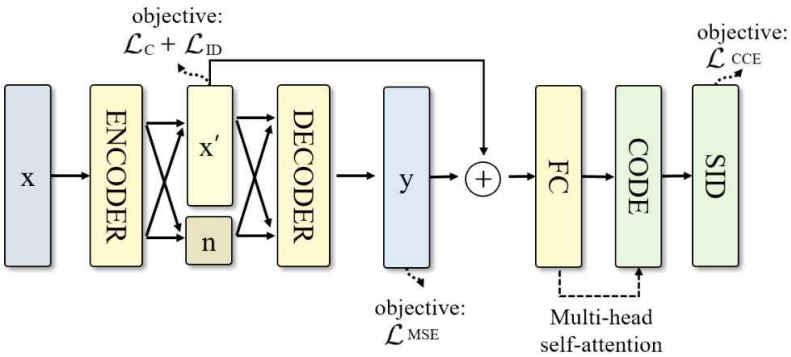
원거리 화자인증 시스템을 위해 선택적으로 특징을 보상하는 분별적 잡음 제거 자기부호화기

김주호, 정지원, 심혜진, 유하진
서울시립대학교 컴퓨터과학과

Discriminative Denoising Autoencoder for Selective Embedding Compensation for Long-distance Speaker Verification

Ju-ho Kim, Jee-weon Jung, Hye-jin Shim, Ha-Jin Yu
School of Computer Science, University of Seoul
wngh1187@naver.com, jeewon.leo.jung@gmail.com, shimhz6.6@gmail.com, hjyu@uos.ac.kr

원거리에서 발화된 발성은 화자 인증 시스템의 신뢰도 저하 요인 중 하나이며, 성능 저하를 완화하기 위해 다양한 보상기법들이 제안되었다. 그러나 기존 원거리 발성 보상시스템에 다양한 거리에서 발화된 발성이 입력될 경우, 원거리 입력 발성에 대한 성능은 향상되지만, 근거리 입력 발성에 대한 성능 저하 문제가 나타날 수 있다. 이는 보상된 화자 특징 추출기가 원거리 발성에 최적화하도록 학습되어, 근거리 발성에 대해 분별력이 낮은 화자 특징을 추출하기 때문이라고 가정할 수 있다.[1] 본 논문은 입력 발성의 거리에 따른 성능 하락을 완화하고자, 발성의 거리에 따라 화자 특징을 선택적으로 보상하는 분별적 잡음 제거 자기부호화기를 제안한다.



<그림 1. 제안한 시스템의 구조>

제안한 시스템은 베이스라인으로 사용한 RawNet[2]으로부터 추출한 화자 특징을 입력으로 한다. 그림 1은 제안한 자기부호화기 시스템의 구조를 나타낸다. 먼저, RawNet으로부터 추출된 화자 특징(x)은 화자 정보를 강화하기 위해 인코더에 입력되어, 정제된 화자 특징(x')과 나머지 정보(n)로 분리된다. 분리한 정보들은 잡음 및 잔향이 제거된 화자 특징으로 재구성하기 위하여 디코더에 입력된다. 인코더 및 디코더는 각각 2개의 전결합층으로 구성된다. 잡음 및 잔향이 제거되어 재구성된 화자 특징(y)과 정제된 화자 특징(x')은 하나의 벡터로 연결되어 전결합층에 입력된다. 이로부터 출력된 벡터는 multi-head self attention 기법이 적용되어, 최종적으로 입력 화자 특징 x의 보상된 화자 특징(CODE)으로 사용된다.

제안한 모델은 4가지의 손실함수를 사용하여 학습된다. 먼저, x'계층에서 center

loss(L_c)[3]와 inter dispersion loss(L_{ID})[4]를 사용하며, y계층에는 mean squared error (L_{MSE})를 사용한다. 마지막 계층에는 categorical cross entropy loss(L_{CCE})를 사용하여 화자 식별을 수행한다.

제안한 모델의 학습 및 평가에는 원거리 발성과 근거리 발성이 모두 포함된 VOiCES[5] 데이터 셋을 사용하였다. 실험 결과, 베이스라인은 7.70%의 동일 오류율, 베이스라인에 제안한 화자 특징 보상 시스템을 적용한 경우 6.34%의 동일 오류율을 얻을 수 있었다. 이 결과로부터 제안한 분별적 잡음 제거 자기부호화기가 다양한 거리에서 녹음된 발성 환경에 강인함을 확인하였다.

감사의 글

이 연구는 2018년도 산업통상자원부 및 산업기술평가관리원(KEIT) 연구비 지원에 의한 연구임 (10076583).

참고문헌

[1] Jung, Jee-weon, et al. "Selective Deep Speaker Embedding Enhancement for Speaker Verification." Proc. Odyssey 2020 The Speaker and Language Recognition Workshop, 2020.

[2] Jung, Jee-weon, et al. "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification." Interspeech, 2019.

[3] Yandong Wen, et al. "A discriminative feature learning approach for deep face recognition," in European conference on computer vision. Springer, 2016.

[4] Hung-Shin Lee, et al. "Discriminative autoencoders for speaker verification," ICASSP 2017.

[5] C. Richey et al., "Voices obscured in complex environmental settings (voices) corpus.", Interspeech, 2018.

Average modeling 및 GST를 이용한 다화자 비병렬 음색변환[1)]

엄 지 섭, 김 회 린
KAIST 전기및전자공학부

Multi-speaker & non-parallel voice conversion using average modeling and GST

Jisub Um, Hoirin Kim
School of Electrical Engineering, KAIST
{twiz0311, hoirkim}@kaist.ac.kr

본 연구는 average modeling 방식으로 훈련을 진행한 음소 인식기 기반의 음색변환 시스템과 MelGAN 보코더를 결합하는 연구를 제안한다. 앞선 여러 연구에서 비병렬 음색변환을 위해 음소 인식기를 활용하여 추출한 phonetic posteriogram(PPG)를 사용하는 방법들이 제안되어왔다. 본 연구 또한 해당 방식을 사용하여 타겟 화자의 mel-spectrogram을 생성하도록 음색변환 모델을 feature average modeling 방식으로 훈련을 진행하고 이후 MelGAN 보코더와 결합하여 음색이 변환된 음성을 생성했다.

기존의 음소 인식기를 활용한 음색변환 시스템은 훈련 과정에서 사용된 하나의 화자에 대해서만 음색변환이 가능하다는 한계를 가지고 있었다. 하지만 본 논문에서는 average modeling 방식을 적용하여 다화자에서도 음색변환이 가능하도록 하였다. Average modeling이란 화자와 독립적인 linguistic 정보를 화자와 연관성을 가지는 acoustic 정보와 매칭을 시켜주도록 다화자의 평균값으로 훈련되도록 하는 방식으로 여러 화자의 음성을 생성하기 위해서 화자별 embedding을 average model에 추가 input으로 사용한다. 화자별 embedding은 화자마다 학습이 가능한 파라미터로 구성되어 음색변환모델과 함께 훈련이 진행된다. 또한 Global Style Token(GST)을 활용하여 음색에 대한 추가정보를 추론 과정에서 사용할 수 있도록 하여 성능을 높였다.

18명의 한국어 여성 화자와 17명의 남성 화자를 이용하여 음소 인식기, 음색변환 모델, MelGAN 보코더 훈련을 진행하였다. 이후 남녀 각각 1명의 화자를 소스 화자로 또 다른 남녀 각각 1명의 화자를 타겟 화자로 설정하여 총 4가지의 음색변환을 진행하였다. 변환이 완료된 음성을 대상으로 음소 인식기를 통해 음소 인식 오류율과 화자 인식기를 활용하여 얻은 화자 embedding을 바탕으로 유사도를 측정하였다. 해당 값들을 객관적 평가 지표로 사용한 결과 2가지 측면 모두에서 성능이 향상됨을 보였다.

1) 본 연구는 산업통상자원부의 산업기술혁신사업으로부터 지원을 받아 수행된 연구임 (No. 10080667, 음원 다양화를 통하여 로봇의 감정 및 개성을 표현할 수 있는 대화음성합성 원천기술 개발).

고속 한국어 종단형 음성인식 시스템

박 기 영, 오 유 리, 박 전 규
한국전자통신연구원 복합지능연구실

Fast End-to-end Korean Speech Recognition

Kiyoung Park, Yoo Rhee Oh, Jeon Gyu Park
Artificial Intelligence Research Lab, ETRI
{pkyoung,yroh,jgp}@etri.re.kr

기계학습 및 인공지능 분야의 꾸준한 발전과 함께 음성인식기술도 발전을 거듭하고있다. 특히 최근 들어서는 종단형 음성인식기가 종래의 인식기에 비하여 매우 높은수준의 음성 인식 성능을 보여주고 있다. 종단형 음성인식기는 주어진 음성 신호에대한 단어열의 조건부 확률을 구하기 위하여 언어모델, 발음사전, 음향모델의 구분 없이 하나의 학습 규칙으로 훈련된 신경망을 이용하여, 기존 인식기에 비하여 단순한 구조로 높은 성능을 보인다.

본 발표에서는 오픈소스로 널리 이용되는 ESPNet 기반의 종단형 음성인식 툴킷을 이용하여, 한국어 종단형 음성인식 모델을 학습하고 베이스라인 음성인식 디코더 구성하였다. 실험에 사용한 모델은 트랜스포머 기반의 인코더-디코더 구조를 가지며, 훈련 과정의 안정성 및 빠른 수렴을 위하여 Connectionist Temporal Classification(CTC) 조건이 같이 사용되었다. 몇 가지 구조의 모델의 성능 속도를 비교하였으며, 기존의 DNN/HMM 방식의 인식기와도 성능을 비교하여 성능 개선을 확인하였다.

긴 발화 인식 및 인식 속도 개선을 위하여 발화 분할, 다중 GPU 사용, 배치 디코딩, 반정밀도(half precision) 디코딩 등의 기법이 적용되었다. 이러한 속도 개선 기법들은 성능 저하가 거의 없이 고속의 디코딩이 가능하게 하였으며, 최종적으로 8개의 GPU 카드를 장착한 고성능 서버에서 100시간의 한국어 자유발화 음성을 90% 이상의 음절 인식률을 내도록 인식하는데 2분 이하의 시간이 소요되었다.

* 이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2019-0-01376, 다중 화자간 대화 음성인식 기술개발)

Frequency-level Feature for Rapid Emotion Recognition

Di (Suk Young) Lim[1], Hong In Yoon[2], Hyeong Ju Na[2], Jeong-Sik Park[2]
[1]Data Intelligence Lab, Hanyang University
[2]Language Technology Research Institute, Hankuk University of Foreign Studies
offlim@outlook.com, gnlenfn@gmail.com, skgudwn34@gmail.com, parkjs@hufs.ac.kr

As Automatic Speech Recognition, ASR has achieved fine performance, enhancing the quality of the recognition emerged as a next step. To fully understand human intention of the spoken language, recognizing speaker's emotion added as another feature of the data. Indeed, emotion recognition requires precise model as well, yet considering its change frequency within a single speech, it is expected to be able to be handled with low resolution or relatively small data compared those require for ASR. In this paper, we suggest usage of data extracted at frequency level. By using the data extracted at each frequency bands, the size can be decreased significantly while maintaining the emotional characteristics. Since its time dimension is cut, only single dimension, which is by frequency level, is left with the size of less than 1% of the original full MFCC feature data from up to 2-second-long speeches. We conducted experiment comparing the accuracy between the full MFCC data and the proposed frequency level data with the same environment, simple 3-layered DNN and confirmed maintaining of its performance. Especially for 3 emotion recognition, the accuracy dropped only by 0.84%. Then by optimizing the parameters for proposed data extraction, we were able to improve its accuracy, especially at speaker dependent environment.

1.

지식 증류 기법을 사용한 감정을 포함하는 음성인식의 적응학습

윤홍인, 임석영, 나형주, 박정식
한국외국어대학교 영어학과

Adaptation of Emotional Speech Recognition based on Knowledge Distillation

Hong In Yoon, Suk Young Lim, Hyeong Ju Na, Jeong-Sik Park
Dept. of English Linguistics, Hankuk University of Foreign Studies
gnlenfn@gmail.com, offlim@outlook.com, skgudwn34@gmail.com, parkjs@hufs.ac.kr

최근, End-to-End 음성인식은 기존의 HMM-DNN과 같은 음성인식 시스템과 비교했을 때 더 나은 성능을 보이고 있다. End-to-End 방식은 기존의 음성인식 시스템의 각 단계에서 필요한 음향모델, 언어모델 등을 만들지 않더라도 한 번에 통합하여 처리할 수 있다는 장점을 가지고 있다. 하지만 딥러닝을 활용한 End-to-End 시스템은 훨씬 많은 연산과 데이터를 요구하는 단점을 가지고 있다.

이것을 극복하기 위해 Geoffrey Hinton은 지식 증류 기법 (Knowledge Distillation)을 제안했다. 지식 증류 기법은 딥러닝 모델을 덜 복잡하게 만들어 준다. 딥러닝 모델에는 매우 많은 양의 파라미터를 가지고 있지만 지식 증류를 통해 파라미터 수를 줄이더라도 기존의 모델 성능을 유지할 수 있도록 하는 방법이다. 이때, 지식 증류 기법은 모델 압축 뿐 아니라 적응 학습에도 사용될 수 있다. 지식 증류의 수식이 쿨백-라이블러 발산과 유사하기 때문에 적응학습에도 사용될 수 있는 것이다.

이처럼 딥러닝을 활용한 End-to-End 음성인식은 큰 성과를 거뒀지만 아직 감정과 같은 비언어적 요소를 포함하는 음성의 경우 일반적인 음성인식에 비해 부족하다. 왜냐하면 감정을 포함하는 대규모 음성 데이터를 구축하는 것이 어렵기 때문이다. 이 연구에서는 지식 증류를 활용한 적응 학습을 통해 기존의 음성인식 시스템을 감정을 포함한 음성에서도 높은 인식률을 보일 수 있도록 했다. 그 결과 선생 모델과 학생 모델의 크기가 같을 때 성능의 차이가 거의 없었지만, 학생 모델의 크기가 작아지면 적응학습의 효과가 두드러져 더 나은 결과를 얻을 수 있었다.

Transformer 기반 종단형 음성인식 시스템에서의 CTC 활용

오 유 리, 박 기 영, 박 전 규
한국전자통신연구원 인공지능연구소 복합지능연구실

On the use of connectionist temporal classification (CTC) for a Transformer-based end-to-end speech recogintion

Yoo Rhee Oh, Kiyoung Park, and Jeon Gyu Park
Artificial Intelligence Research Laboratory
Electronics and Telecommunications Research Institute (ETRI)

{yroh,pkyoung,jgp}@etri.re.kr

하드웨어 및 딥러닝(deep learning) 기술의 발전과 함께 음성인식(automatic speech recognition) 성능도 향상되고 있다. 특히, 대용량 음성데이터베이스로 학습된 종단형(end-to-end) 음성인식기는 기존의 음성인식기와 비교하여 높은 음성인식 성능을 보인다. 본 발표에서는 학습 과정에서의 빠른 수렴과 음성인식 성능 개선을 위하여 connectionist temporal classification (CTC)가 결합된 Transformer 기반 종단형 음성인식기에 초점을 두며, 음성인식 디코딩 과정에서 계산되는 CTC score의 활용들을 보인다. 먼저, 음성인식 디코딩 과정의 CTC score를 이용하여 음성인식 문자열의 시간 정보를 획득하는 방법을 보인다. 그 후, CTC score를 기반으로 획득된 시간 정보를 활용하여 음성 끝점 검출(end-of-speech detection)을 개선하는 방법을 보인다. 또한, CTC score를 기반으로 획득된 시간 정보를 활용하여 CTC score 계산 구간을 제한하는 방법을 보인다.

본 발표에서는 음성인식을 위하여, 종단형 음성인식 툴킷인 ESPNet [1]로 학습된 Transformer 기반 한국어 종단형 음성인식기를 사용한다. 먼저, 음성인식 디코딩 과정에서의 CTC score 분포를 보이고, 이를 기반으로 한 음성인식 문자열에 대한 시간 정보 추출 결과를 보인다. 다음으로, CTC score 기반 시간정보를 활용한 음성 끝점 검출 방법 및 CTC 계산구간 제한 방법을 적용함으로써, 음성인식 성능은 최대한 유지하면서 음성인식 디코딩 속도를 개선함을 보인다.

감사의 글

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2019-0-01376, 다중 화자간 대화 음성인식 기술개발)

참고문헌

[1] S. Watanabe et al., "ESPnet: End-to-end speech processing toolkit," in Proceedings of Interspeech, 2018, pp.2207-2211.

강인 음성인식을 위한 유도 어텐션 기반의 시청각 음성인식

이용혁, 김재빈, 박형민
서강대학교 전자공학과

Audio-Visual Speech Recognition Based on Guided Attentions for Robust Speech Recognition

Yong-Hyeok Lee, Jae-Bin Kim, Hyung-Min Park
Dept. of Electronic Engineering, Sogang University
eug92@sogang.ac.kr, rlawoqls7@naver.com, hpark@sogang.ac.kr

기계 번역분야에 어텐션(attention)이 도입된 이후, 어텐션은 음성인식과 같은 sequence-to-sequence 문제를 해결하기 위하여 Long Short-Term Memory (LSTM)와 함께 사용되었다. 일반적인 음성인식과 달리 시청각 음성인식은 오디오와 비디오사이의 상관 관계를 학습함으로써 향상된 성능을 제공할 수 있다. 하지만 오디오 음성인식은 립리딩으로 알려진 시각정보 음성인식보다 더 쉽게 훈련될 수 있기 때문에 시청각 음성인식은 균형잡힌 학습이 어렵다. 이를 위해 시각정보와 청각정보의 균형잡힌 학습을 위하여 두 정보가 선형적으로 시간 정렬된 것을 이용한 guided attention loss를 제안한다. Guided attention loss를 통하여 어텐션에서 시청각 정보간 정렬이 보다 잘 이루어지게 되며 BBC-LRS2와 TED-LRS3 데이터셋에 대하여 성능 향상을 확인하였다.

키워드: 시청각 음성인식, 강인 음성인식, 어텐션, sequence-to-sequence

사사: 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2017R1A2B4009964, NRF-2020R1A2B5B01002398).

공감형 대화 음성 챗봇 개발을 위한 공감 상황 인지 코퍼스 구축

김종인, 정민화
서울대학교 인지과학 협동과정

Jongin Kim, Minhwa Chung
Dept. of Cognitive Science, Seoul National University
prows12@gmail.com, mchung@snu.ac.kr

최근 대화 모델은 인공지능 스피커의 발달과 더불어서 다양한 분야에 응용되고 있다. 일반적으로 대화 모델은 자연언어처리 분야의 하나로, 다양한 머신러닝 기법을 적용하여, 연구되어왔다. 대화 모델은 크게 Task-Bot과 Chit-Chat Bot로 분류된다. Task-bot은 특정 업무를 수행하는 챗봇이다. 호텔 예약이나 비행기 예약 등의 특정 업무를 수행하기 위해서 개발된 챗봇이다. 일반적으로는 Natural Language Understanding, Dialog Manager, Natural Language Generation으로 구성되어 있고, 각각의 모듈을 특정 업무에 맞게 모델링한다. 최근에는 이 세 가지 모듈을 통합하는 end-to-end 모델이 활발하게 연구되고 있다. Chit-Chat 모델은 단순한 잡담이나, 일상적인 대화를 수행하는 챗봇을 말한다. 최근에 딥러닝의 발전으로 Chit-Chat 모델이 Seq2Seq, Transformer 등의 다양한 뉴럴넷을 이용하여 연구되고 있다. 대화 모델의 발전으로 인해, 최근에는 소셜챗봇에 대한 관심도 높아지고 있다. 본 연구는 소셜 챗봇의 하나인 공감형 대화 음성 챗봇 개발을 최종 목표로 하고 있다.

공감은 타인의 감정, 의견, 주장에 대하여 마치 자신이 그렇게 느끼는 것처럼 동일하게 느끼는 것을 말한다. 공감은 실제 커뮤니케이션에서는 중요한 요소로 인정받고 있지만, 음성대화시스템의 대화 모델에 공감을 반영하는 연구는 최근에 와서야 이루어지고 있다. 공감형 챗봇의 연구는 감정을 인지(Emotion-Aware)하고 특정한 감정에 대해 다양한 방식을 발화에 반영하는 형태(Emotion-Expressive)로 연구되어 왔다.[1] 예를 들면 하나의 발화에 대하여 기쁨, 슬픔, 화남, 혐오 등의 형태로 감정을 분류하고, 이를 기반으로 하여, 기쁨, 슬픔, 화남, 혐오 등의 감정을 표현하는 발화를 생성하는 것이다. 이외에도 페르소나를 인지하거나, 외부 세계의 지식을 반영하고자 하는 연구들도 진행되고 있다. [2],[3]

하지만 기존의 공감형 대화 모델에는 몇 가지 한계점이 있다. 첫째, 기존의 공감형 모델은 공감의 정의에 대해서 제대로 반영하지 못하고 있다. 공감은 알포트(1961)[4]에 따르면 ”자신을 다른 사람의 생각, 느낌 및 행동 속에 상상적으로 전위시키는 것“이다. 이러한 공감의 정의를 고려해볼 때, 기존의 공감형 대화 모델은 공감을 ‘감정의 이해‘의 의미로 제한시켜, 실제로 공감의 일부만을 대화모델에 반영하고 있다.

둘째, 기존의 공감형 대화 모델은 인지적 공감에 대해서는 거의 대응하지 못하고 있다. 통상적으로 공감은 크게 정서적 요소와 인지적 요소로 구분할 수 있다. 정서적 공감은 상대의 감정 상태를 인지하고, 상대의 감정을 고려하여 적절한 발화를 생성하는 것을 의미한다. 현재 공감형 대화 모델에서는 정서적 요소에 대해서는 어느 정도 반영하고 있다. 이는 발화의 감정 분류를 통해 감정을 인지하는 형태로 가능하다. 즉 공감을 반영할 때는 상대방의 발화를 통해 감정을 인식하고 이에 따른 적절한 공감 발화를 생성하는 것을 목표로 한다.

인지적 공감은 미드(1934)의 정의 따르면 '타자의 역할을 취해 보고 대안적인 조망을 취해 볼 수 있는 능력'을 의미한다.[5] 실제 대화에서는 자신의 감정을 단어로 명확하게 표현하기보다는 뉘앙스로만 들어내는 경우가 많으며, 이는 인지적인 공감을 통해서만 파악이 가능하다. 그러나 기존의 공감형 챗봇에서는 인지적 공감이 필요한 상황에 대해서는 거의 대응하지 못하는 약점을 지닌다. 예를 들면 다음의 대화를 가정하자.[6]

A : 오빠 그래도 우리는 꽤 친한 편이야 그치
B : 너랑 내가 아니 난 전혀 그렇게 생각한 적이 없는데 단 한번도
A : 왜 그래 갑자기 오빠 답지 않게
B : 나답지 않다니 나 다운 게 뭔데

A는 본인이 B와 친하다는 것을 어필하고 싶어하는 상황이다. B는 본인이 생각하기에는 자신과 A는 전혀 친분이 없다는 걸 얘기하고 싶어한다. A가 오빠 답지 않다고 서운함을 표시하자, B는 자신 다운 게 뭔지 묻고 있다. 이 대화에서는 직접적으로 감정을 나타내는 단어는 거의 나타나지 않는다. 그러므로 정서적 공감 혹은 감정 분류로 상황을 파악하기는 어렵다.

기존의 공감형 챗봇 시스템에서는 이러한 상황에 대해서, 단순히 한 가지의 감정으로 분류하여 반영할 뿐 제대로 대응하기 힘들다. 본 연구에서는 이러한 약점을 극복하기 위해서 정서적 공감 뿐만 아니라 인지적 공감에 대응하고자 했다. 이를 위해, 공감이 필요한 상황을 인지하는 소프트웨어를 개발하는데 필수적인 학습 코퍼스 구축하였다. 기존의 공개된 Aihub 멀티모달 영상 데이터셋을 바탕으로 정서적 공감과 인지적 공감에 대응하는 메타 정보를 추가하는 방식으로 진행하였다.

Aihub 멀티모달 영상 데이터셋은 100시간 분량의 영상 데이터셋이며, 6,000개의 영상클립과, 85,077개의 발화로 구성되어 있다. 공감 태깅은 각각의 발화에 대하여 공감이 필요한 상황인지를 판단한 후 인지적 공감과 정서적 공감을 구분하여 태깅하였다. 기존의 공감형 대화 모델이 대부분 텍스트 기반의 챗봇인 것과 달리, 본 연구는 Aihub 멀티모달 영상 데이터셋을 이용하여, 텍스트 기반이 아닌 음성 기반의 공감형 챗봇을 개발하고자 한다.

본 연구에서는 구축한 확장 가공한 코퍼스를 통하여, 공감형 대화 음성 챗봇 개발 시 정서적 공감과 인지적 공감이 필요한 상황에 모두 대응을 할 수 있게 하는 것을 목표로 하고 있다. 이와 함께, 기존의 텍스트 기반의 챗봇이 아닌 음성 기반의 챗봇 소프트웨어 개발을 최종 목표로 하고 있다.

참고 문헌

[1] Shum, H.-Y., He, X., & Li, D. (2018). From Eliza to XiaoIce: Challenges and Opportunities with Social Chatbots. http://arxiv.org/abs/1801.01957
[2] Ma, Y., Nguyen, K. L., Xing, F. Z., & Cambria, E. (2020). A survey on empathetic dialogue systems. Information Fusion, 64(June), 50-70. https://doi.org/10.1016/j.inffus.2020.06.011
[3] Pamungkas, E. W. (2019). Emotionally-Aware Chatbots: A Survey. http://arxiv.org/abs/1906.09774
[4] 박성희, 『공감함: 어제와 오늘』, 학지사(2017), p22
[5] 박성희, 『공감학: 어제와 오늘』, 학지사(2017), p48
[6] Aihub 멀티 모달 데이터셋의 대화 일부 발췌.

감사의 글

본 연구는 대학ICT연구센터지원사업의 '의료 빅데이터 융합전문가 인력양성을 위한 비정형 빅데이터의 정형화 기술 및 분석 플랫폼 개발' 과제의 연구결과로 수행되었습니다.

실시간 음성인식을 활용한 임베디드 한국어 음성 대화 시스템

손 현 수, 박 호 성, 김 지 환[2)]
서강대학교 컴퓨터공학과

Embedded Korean Spoken Dialogue System Using Real-time Speech Recognition

Hyunsoo Son, Hosung Park, Ji-Hwan Kim
Dept. of Computer Science and Engineering,
Sogang University
sonhyunsoo@sogang.ac.kr, hosungpark@sogang.ac.kr, kimjihwan@sogang.ac.kr

본 논문은 핵심어 검출, 음성인식, 질의응답 대화모델, 음성합성을 결합한 인공지능 비서 시스템을 제안한다. 스마트폰 기반의 비대면 금융 서비스는 어플리케이션 실행, 메뉴 탐색 등의 사용자의 능동적인 접근을 요구한다. 이에 대한 대안으로 AI 스피커를 통한 금융 업무가 제안되었으나, 하드웨어 제조사에 사용자의 개인정보가 전달되는 문제가 발생하여 환율 정보, 금융 상품 추천 등의 개인정보를 사용하지 않는 한정된 범위에서의 서비스만 제공되고 있다. 본 논문에서 제안하는 인공지능 비서 시스템은 개인정보 유출 문제가 없는 인공지능 비서 시스템을 제안한다. 시스템은 핵심어 검출, 음성인식, 대화모델, 음성합성으로 구성 되어있다. 핵심어 검출은 CMU pocketsphinx와 Google WebRTC VAD를 사용하여 임베디드 기기에서 동작 가능한 GMM-HMM 기반 단어 인식 모델을 사용하였다. 핵심어가 인식되면 사용자로부터 음성으로 명령을 받는다. 음성인식 모델은 음향모델, 언어모델, 디코딩 네트워크로 구성되어 있다. 음향모델은 음성신호를 입력받아 가장 확률이 높은 발음열을 생성하며, Kaldi를 활용한 TDNN(Time Delay Neural Network) 기반 음향모델로 구현하였다. 언어 모델은 음향모델의 발음열의 결과로 출력된 단어 간의 확률을 계산하여 가장 확률이 높은 문장을 출력하며, n-gram 기반으로 구현하였다. 음향모델과 언어모델을 WFST(Weighted Finite State Transducer) 기반 디코딩 네트워크를 통해 음성인식 결과를 출력하며 대화처리 모델의 입력으로 전송된다. 대화처리 모델은 사용자의 발화를 분석하여 사전 지식을 바탕으로 적합한 시스템 응답을 제공한다. 대화처리 모델은 ALBERT(A Lite Bidirectional Encoder Representation from Transformers)기반 언어모델을 질의 응답형 대화모델로 KorQuAD 1.0에 fine-tuning 하여 사용하였다. 음성합성은 Amazon Web Service의 음성합성 서비스인 Polly를 사용하여 입력된 텍스트에 대해 출력 오디오 스트림을 실시간으로 재생한다. 본 시스템을 통해 계좌 이체, 조회 등의 개인정보가 필요한 서비스와 사용자 맞춤형 인공지능 비서 서비스 제공이 가능하다.

감사의 글

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임. (NO.2020R1F1A1076562)

* 교신저자: 김지환(kimjihwan@sogang.ac.kr)

멀티태스크 U-Net과 채널 어텐션 기반의 다채널 음성 향상

이 건 우, 김 홍 국
광주과학기술원 전기전자컴퓨터공학부

Multi-channel Speech Enhancement Based on Multi-task U-net and Channel Attention

Geon Woo Lee, Hong Kook Kim
Gwangju Institute of Science and Technology (GIST)
{geonwoo0801, hongkook}@gist.ac.kr

다채널 마이크로폰 어레이를 활용한 음성 향상 기술은 자동 음성 인식 등 다양한 음성 처리 응용 분야의 전처리 과정에 광범위하게 적용되고 있다. 하지만 음성 응용 시스템의 성능은 배경 잡음이 큰 환경에서는 그 성능이 저하가 되는 경향이 있으며, 이를 극복하기 위해 다양한 다채널 음성 향상 기술이 연구되고 있다. 본 논문에서는 다채널로 입력되는 신호로 부터 각 채널의 가중치를 활용할 수 있는 채널 어텐션을 적용한 딥러닝 기반의 음성 향상 기술을 제안한다. 제안된 딥러닝 모델의 구조는 합성곱층으로 구성된 1개의 부호기와 음성과 잡음 신호의 마스크를 추정하는 2개의 복호기의 멀티태스크 U-net으로 구성되며, 부호기와 복호기의 동일한 위치의 합성곱층은 접합층으로 연결된다. 이와 같이 구성된 모델의 각 합성곱층 사이에는 채널 어텐션을 적용한다. 제안된 딥러닝의 입력 특징은 각 채널 신호의 스펙트로그램을 채널 단위로 쌓아서 3-D로 구성되고, 출력 특징은 타겟 음성과 배경잡음의 두가지 스펙트로그램으로 구성된다. 제안된 딥러닝 모델의 음성 향상 성능을 평가하기 위해, 6개의 마이크 채널로 구성된 CHiME3의 훈련 데이터를 사용하여 딥러닝 모델을 학습하고 평가 데이터를 사용하여 PESQ와 SDR를 측정하여 비교하였다. 실험 결과, 제안된 음성 향상 기법은 PESQ와 SDR를 각각 1.18점과 11.95 dB 만큼 개선하였다.

감사의 글

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2019-0-00330, 영유아/아동의 발달장애 조기선별을 위한 행동·반응 심리인지 AI 기술 개발)

참고문헌

[1] J. Barker, *et al.*, “The third 'CHIME' speech separation and recognition challenge: analysis and outcomes,” *computer speech & language*, vol. 46, pp. 605–256, 2017.

다화자 동시 발화 환경 diarization을 위한 확장된 personal VAD

이 승 형, 한 민 수
한국과학기술원 전기 및 전자공학부

Extended Personal VAD for Overlapped Speaker Diarization

Seunghyung Lee, Minsoo Hahn
School of Electrical Engineering, Korea Advanced Institute of Science and Technology
shlee92@kaist.ac.kr, mshahn2@kaist.ac.kr

본 연구는 동시 발화 상황을 포함하는 다화자 환경에서의 화자 diarization을 위해 확장된 personal VAD를 제안한다. 화자 diarization은 'who speak when?', 즉 어떤 화자가 어느 시점에 발화했는지를 사전 정보 없이 화자별로 할당해주는 알고리즘으로, 음성 인식, 화자 인식을 위한 각종 전처리 및 화자 tagging 등에 사용되고 있다.

화자 diarization의 성능을 저하하는 큰 요소로 발화 겹침 현상을 뽑을 수 있는데, 이는 여러 화자가 동시에 발화하는 상황에서 화자 특성이 크게 왜곡되고, 화자 분류 및 군집화 과정의 오류가 커지는 문제로 이어지기 때문이다. 또한, 발화 겹침이 발생한 시점에서 발화 중인 여러 화자를 동시에 할당해주어야만 올바른 diarization을 수행한 것인데, 대부분의 기존 연구들은 아직 이를 해결하지 못하는 상황이다.

본 논문에서는 타겟 화자의 임베딩을 미리 등록해 타겟 화자의 발화가 아니라면 설령 음성이더라도 non-target speech로 간주하는 기존의 personal VAD를 개선한, 확장된 personal VAD (extended personal VAD)를 제안하였다. 확장된 personal VAD는 발화 겹침이 포함된 상황에서도 검출하고자 하는 화자를 추적할 수 있도록 이용해 다양한 발화 겹침 상황을 시뮬레이션 한 데이터베이스로 훈련되었으며, 타겟 화자 임베딩과 겹침 구간 음성에서 추출한 임베딩, 그리고 mel spectrogram을 이용해 프레임 단위로 해당 구간에 타겟 화자의 존재 여부를 검출한다.

실제 diarization 과정에서는 사전 정보를 얻을 수 없으므로 미리 화자 등록을 하는 것이 불가능하다. 따라서 preliminary diarization을 통해 1차적으로 얻은 화자 클러스터들로부터 임베딩을 추출하고, 이를 확장된 personal VAD에 등록하여 사용하였다. 제안된 알고리즘을 통해 다 화자 동시 발화 환경을 포함하는 환경에서의 화자 diarization을 수행하였고, diarization error rate (DER)의 감소를 확인할 수 있었다.

범주적 지각 음성인식 실험을 통한 인공신경망 모델의 설명가능성 접근

이 호 진
서울대학교 자유전공학부

Assessing Artificial Neural Network Explainability via Categorical Speech Perception Experiment

Hojin Lee
College of Liberal Studies, Seoul National University
nyx1371@snu.ac.kr

최근 발전하고 있는 심층신경망 기반 종단간(end-to-end) 음성인식 모델은 전통적 처리모듈 없이도 높은 정확도를 보이고 있으며, 훈련 과정과 구조 또한 간단하다는 장점이 있다. 하지만 종단간 모델은 내부 정보처리 과정을 알 수 없다는 단점을 지니고 있으며, 이러한 인공신경망 모델의 설명가능성(explainability) 문제는 음성인식뿐만 아니라 모든 인공지능 연구의 중요한 화두 중 하나이다. 본 연구는 종단간 음성인식 인공신경망의 내부 연산과정을 이해하기 위해서 심리언어학에서 이루어지는 음성실험의 도입을 제안한다. 이러한 접근방법은 크게 두 가지의 장점을 지니고 있다. 첫째로 음성코퍼스에 존재하지 않는 조작된 음성자극을 인공신경망이 어떻게 처리하는지 탐구할 수 있으며, 이는 기존 연구들에서는 다루어지지 않았던 측면이다. 둘째로 음성실험과 관련된 심리언어학적 이론들을 통해 인지적 관점에서 인공신경망의 언어처리를 이해할 수 있다. 이 연구가 제시하는 새로운 접근방식은 향후 음성인식 인공신경망의 다양한 특징을 이해하는 데에 응용될 수 있을 것이며, 특히 범주적 지각(Categorical perception) 실험을 통해 그 구체적 예시와 장점이 무엇인지 밝힐 것이다.

범주적 지각이란, 연속적으로 변하는 일련의 음성자극 집합들을 불연속적이고 범주적으로 인식하는 현상을 의미한다. 구체적으로 이 연구에서는, 영어의 어두 /d-t/ 대립을 가지는 최소대립쌍 단어를 기준으로 연속적인 VOT(voice onset time)를 가지도록 조작된 음성자극 연속체를 인공신경망이 어떻게 처리하는지 살펴보았다. 그 결과 불연속적으로 변하는 최종 출력층의 확률값을 통해 인공신경망에서도 범주적 지각 현상이 나타남을 확인하였으며, 은닉층 출력의 시계열을 분석함으로써 이러한 범주적 지각이 단어 층위 정보에 의한 하향효과일 것이라 예측하였다. 본 연구를 발판삼아 인공신경망의 설명가능성을 탐구하기 위한 인지과학적 접근이 활발해지기를 기대한다.

포스터 발표 II
말장애 및 음성의학

좌장: 이수복(우송대)

말더듬 성인의 취업 및 직장생활 경험에 대한 질적연구

박 소 현* 박 진**
***충남대학교 대학원 언어병리학과**
****가톨릭관동대학교 언어재활상담학과**

A Qualitative Analysis on Work-related Experiences of Adults who Stutter

Sohyun Park, Jin Park
Department of Interdisciplinary Program of Communication Disorders, Chungnam National University
Department of Speech, Rehabilitation, and Counseling, Speech Rehabilitation Clinic Center, Catholic Kwandong University
sohpark@cnu.ac.kr, gatorade70@cku.ac.kr

서론: 현재까지 말더듬 성인(AWS)의 취업 및 직장생활 관련해서는 직업 추천 양상이나 직무수행 능력에 대한 일반인의 인식 관련 연구가 주를 이루어왔다(Gabel et al., 2004, Park et al., 2020). 이러한 연구들은 AWS가 실제 취업 과정이나 직장생활에서 얻게 되는 구체적이고 개별적인 경험, 인식, 느낌, 태도 등을 보여주기에는 분명 한계가 있다. 본 연구에서는 구직 및 취업 과정 그리고 직장생활과 관련해 질적 연구 방식을 통해 AWS의 좀 더 개별적이고 내밀한 경험과 이를 통한 인식 및 태도 양상에 대해 알아보고자 하였다. **연구방법**: AWS 남성 5명을(평균연령: 32.6세) 대상으로 개별 심층면담을 통해 말더듬이 구직 및 취업 과정 그리고 직장생활에 어떠한 영향을 미쳤는지를 알아보았다. 이를 위해 관련 주제에 대한 질문지를 미리 작성, 반구조화된 상황에서 개별 심층면담을 실시하였다. 면담자료는 전사(verbatim)하였으며 연속적 비교법(Shin et al., 2004)을 통해 부호화하고 범주 주제를 도출하는 방식으로 분석하였다. **연구결과**: 분석결과, '구직 및 취업 과정에서의 말더듬의 부정적 영향', '직장생활에서의 말더듬의 부정적 영향', '부정적 경험이 AWS의 인식 및 정서에 미치는 영향', '말더듬으로 인한 문제 해결 모색', '문제 해결 모색의 결과'를 포함한 5개의 주제(main theme)가 도출되었다. **결론 및 논의**: 구직 및 취업 과정, 그리고 직장생활에서 말더듬이 주로 부정적인 영향을 미치고 있지만 문제 해결을 위한 개별적 방안들도 모색하고 있음을 알 수 있었다. 또한, 임상현장에서 언어치료사가 AWS를 지원해 줄 방안도 제시하였다.

참고문헌

Gabel, R. M., Blood, G. W., Tellis, G. M., & Althouse, M. T. (2004). Measuring role entrapment of people who stutter. *Journal of Fluency Disorders*, *29(1)*, 27-49.

Park, H. Y., Park, S. H., & Park, J. (2020). Perceptions of Vocational Counselors and Their Career Advice for Individuals Who Stutter. *Audiology and Speech Research*, *16(2)*, 147-156.

Shin, K. L., Jo, M. O., & Yang, J. H. (2004). *Qualitative research methodology*. Seoul, Korea: Ewha Womans University Press.

좌반구 손상과 우반구 손상 뇌졸중 환자의 의문문 유형에 따른 운율 특성 비교

유 영 미, 박 소 현, 성 철 재
충남대학교 대학원 언어병리학과

Comparison of Prosodic Characteristics according to the Question Types in Patients with Left Hemisphere Damaged and Right Hemisphere Damaged

YoungMi Yu, Sohyun Park, Cheoljae Seong
Dept. of Speech & Language Pathology,
Chungnam National University
dudal7178@hanmail.net, sohpark@cnu.ac.kr, cjseong49@gmail.com

대뇌편재화 관점에서 좌반구와 우반구 손상 경도 뇌졸중 환자의 발화 산출시 한국어의 언어적 운율 차이를 살펴 보았다. 의사소통에 있어 중요한 역할을 하는 운율은 기능에 따라 언어적 운율과 정서적 운율로 구분한다. 대뇌 편재화 관점에서 각각의 연구를 살펴보면 정서적 운율 처리에 있어 우반구가 우세하다는 결과가 일반적으로 받아들여지지만 언어적 운율에 대한 연구를 살펴보면 연구 간 방법적인 차이로 인해 결과가 상이하게 나타난다. 또한 국외에 비해 국내는 아직 연구가 부족하다. 이에 본 연구자들은 세 가지 의문문 조건(의문사, 예-아니오, 선택)에서의 발화 과제를 이용하여 좌·우반구 손상 경도 뇌졸중 환자와 정상화자의 운율 관련 음향변수(지속시간, 음도, 강도)의 특성을 살펴보고 청지각 평가와 함께 그들 발화의 차이점을 분석하였다.

연구 결과, 첫째 예-아니오 의문문과 의문사 의문문에서 좌반구 손상 환자가 느린 말속도로 조음하였다. 둘째, 마지막 낱말 음도 범위에서 의문사 의문문과 예-아니오 의문문에서 좌반구 손상 환자가 좁게 나타났으며, 첫 낱말 변화량 기울기는 의문사 의문문에서 좌반구 손상 화자가 낮게 나타났다. 셋째, 예-아니오 의문문에서 문장 강도 표준 편차가 우반구 손상 환자가 유의하게 작게 나타났다. 넷째, 선택 의문문의 첫 번째 운율구(Q1) 음도 회귀 기울기와 마지막 운율구(Q1) 음도 회귀 기울기에서는 유의한 차이를 보이지 않았다. 마지막으로, 억양 자연성에 관한 평가자들(5명의 언어재활사)의 청지각적 분석 결과가 앞서 서술한 음향변수에서 전반적으로 좌반구 손상 화자들의 운율 특성 산출의 결함을 나타낸 것과 일치한다.

결과적으로, 통계적으로 유의한 주요변수들이 좌반구 손상 환자의 언어적 운율을 산출하는데 결함을 보였으며, 예-아니오 의문문과 선택 의문문보다 의문사 의문문에서 더욱 어려움을 나타냈다고 할 수 있다. 이런한 결과는 한국어 사용자의 의문사 사용에 있어서 어휘-의미론적, 통사론적 정보와 같은 언어학적으로 관련 있는 운율 처리의 경우 우반구보다 좌반구에서 우세하다는 점을 시사한다.

과기능적 발성장애 환자의 후두과긴장 특징

김지성[1], 최성희[2], 이동욱[3]

충북대학교병원 이비인후과[1], 대구가톨릭대학교 언어청각치료학과[2], 충북대학교 의과대학[3]

Characteristic of laryngeal muscle in patients with hyper-functional dysphonia

ChungBuk National University Hospital[1], Dept of Audiology & Speech-Language Pathology, DaeguCatholic University[2], Dept of Otorhinolaryngology, ChungBuk National University[3]

slp2046@naver.com, shgrace67@gmail.com, dwlee@chungbuk.ac.kr

본 연구는 정상과 과기능적 발성장애(성대결절, 성대폴립, 근육긴장성발성장애) 환자의 후두외근 과긴장 유무를 비교하기 위한 것이다. 연구자는 후두외근을 설골상근(Suprahyoid muscle), 설골하근(Infrahyoid muscle), 흉쇄유돌근(Sternocleido-mastoid muscle)으로 구분하고, 연구자의 임상적 판단과 후두외근 촉진 시 환자가 호소하는 통증이나 환자의 보고를 참고하여 과긴장 여부를 있음(presence)과 없음(Nothing)으로 구분하였다. SPSS를 이용하여 정상군(23명)과 성대결절(23명), 성대폴립(21명), MTD(17명)로 진단받은 환자의 후두외근 과긴장 빈도를 카이제곱 검정을 통해 비교하였다. 그 결과, 설골상근(p >.000), 설골하근(p >.000), 흉쇄유돌근(p >.001) 유의한 차이가 나타났다. 설골상근에서 성대결절(17.4%)과 MTD(17.6%)집단의 빈도가 높았으며, 성대폴립과 정상집단의 과긴장은 없었다. 설골하근에서는 MTD(76.5%), 성대결절(73.9%), 성대폴립(52.4%), 정상(8.7%)순으로 과긴장 비율이 높았으며 흉쇄유돌근에서도 MTD(64.7%), 성대결절(43.5%), 성대폴립(23.6%), 정상(4.3%)순이었다. 후두외근의 전체의 과긴장 비율은 MTD(52.9%), 성대결절(44.9%), 성대폴립(27%), 정상집단(4.3%)이었다. 이러한 결과는 MTD뿐만 아니라 성대 결절과 성대폴립의 음성평가와 치료에 있어도 후두촉진을 통한 후두외근의 과긴장 평가와 후두이완을 치료가 시행될 필요가 있음을 시사한다.

서비스 제공자의 말소리 친절도 평가 및 분석

장 나 리, 성 철 재*
충남대학교 언어병리학과, 충남대학교 언어학과*

Evaluation and Analysis on the Utterance of Service Provider

Nari Jang, Cheoljae Seong*
Speech-Language Pathology, Chungnam National University
Linguistics, Chungnam National University*
jangzzang22@gmail.com, cjseong49@gmail.com

언어재활사는 상담사의 역할을 하는 서비스 제공자이면서 의사소통 문제를 해결해주고 언어를 다루는 직업이기 때문에 음성을 효율적으로 다룰 수 있어야 한다. 우리는 대화 상황에서 화자의 의도보다 청자가 지각한 감정을 더 중요하게 생각한다. 따라서 발화자가 표현한 친절도와 청자가 받아들인 친절도라는 두 가지 관점에서 음향음성학적 분석을 진행하였다.

발화자는 대전에 거주하는 만 20-29세의 남자 12명, 여자 12명 총 24명으로 ① 현재 서비스직에 근무 중인 자, ② 서비스직 근무 경력이 3개월 이상인 자, ③ 근무 중에 '어서 오세요', '주문 도와드릴까요?', '맛있게 드세요', '감사합니다. 안녕히 가세요.' 문장을 사용하는 자, ④ 후두 병력이 없는 자를 대상으로 하였다. 위의 4가지 문장을 각각 '친절한 말소리', '보통의 말소리', '불친절한 말소리'의 3가지 경우로 산출하여, 발화자 별로 12문장을 청지각 평가에 사용하였다. 청지각 평가자는 충남대학교 대학원 언어병리학과 석사과정 여학생 9명으로 산출문장에 대해 6점 척도로 친절도를 평가하였다.

발화자의 친절도를 중심으로 말소리의 음향음성학적 특성을 분석한 결과, 화자가 친절하게 산출하는 말소리의 특성은 불친절하게 산출한 말소리에 비해 조음속도 및 발화속도가 느리고 문장의 발화시간이 길어지는 경향을 보였다. 그리고 음도가 높고 억양이 풍부해지며 강도의 변화량이 많다는 결과를 확인할 수 있었다. 발화자의 친절도와 청자의 지각이 일치하는 경우를 살펴본 결과, 남·녀 모두 문장의 음도계열 변수가 발화의 친절도를 전달하는데 중요한 역할을 하는 것으로 나타났다. 남성의 경우, 문장 마지막 어절의 강도평균이, 여성의 경우에는 발화속도가 친절도를 전달하는데 중요한 음향학적 변수로 나타났다. 또한 발화자의 친절도와 청자의 지각이 불일치하는 경우, 남성은 주로 강도관련 요소가 청자에게 친절도를 잘못 전달할 수 있는 운율정보가 될 수 있고, 여성은 음도의 산출 범위와 문두에서의 음도가 청자에게 친절도를 잘못 전달할 수 있는 운율정보가 될 수 있는 것으로 나타났다.

발화자의 친절도를 종속변인으로 하고 정규화된 피치, 정규화 강도, 발화속도 등의 운율변수를 독립변인으로 하는 선형판별분석(LDA) 결과, 남성 발화는 문장의 음도중앙값, 여성 발화는 음도변화량, 문장의 음도중앙값, 그리고 조음속도 등이 높은 판별력을 보였다. 또한 청자의 지각을 예측해 집단을 분류한 결과, 남성 발화는 음도변화량, 문장의 마지막 어절 강도중앙값, 여성 발화는 음도변화량이 높은 판별력을 보였다.

우리는 의사소통 과정에서 화자가 어떤 의도로 말하였느냐보다 청자가 어떻게 받아들였는가에 더 주목하게 되고 그것을 중심으로 피드백을 요구한다. 때문에 본 연구에서 나타난, 화자의 의도가 잘못 전달되는 경우의 운율적 요소, 청자의 지각을 예측하는데 중요한 운율적 요소를 확인하고 활용할 수 있을 것이다.

발화속도에 따른 노년층 기준연령 설정

조 보 라, 성 철 재
충남대학교 언어병리학과, 충남대학교 언어학과

Establishment of Standard Age for the Elderly According to the Speech Rate

Bora Cho, Cheoljae Seong
Speech-Language Pathology, Linguistics, Chungnam National University
1231supia@gmail.com, cjseong49@gmail.com

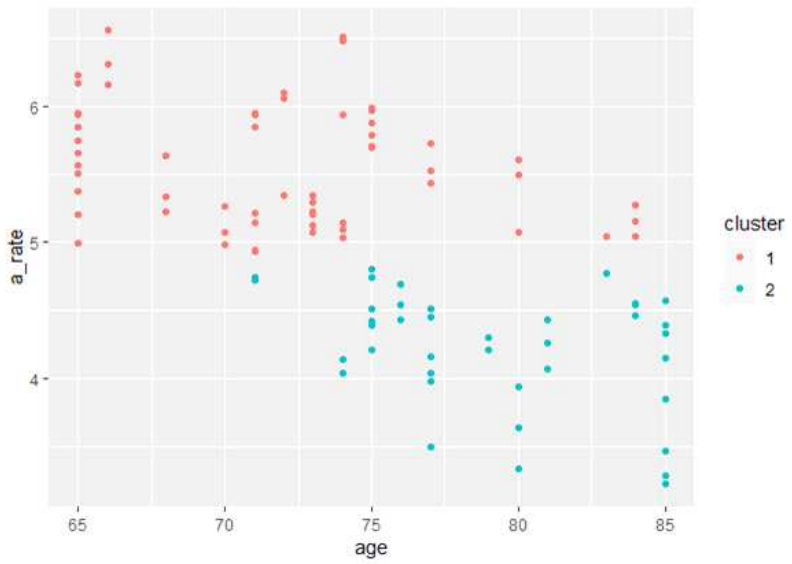
발화 속도(speech rate)는 의사소통 과정에서 말 명료도(speech intelligibility)와 연관이 깊은 것으로 알려져 있다. 발화 속도가 지나치게 빠르면 청자가 화자의 의도를 명확히 이해하기 어려워지며, 지나치게 느린 경우 청자의 주의 집중이 분산되어 발화 내용의 온전한 수용이 어려워진다. 따라서 적절한 발화 속도는 원활한 의사소통을 위한 필수적 요소이다. 그러나 나이가 들면 말 산출에 관여하는 하부 기관 기능의 감소로 인해 점차 발화 속도가 느려지는데, 연령에 따른 발화 속도 변화를 살펴보는 것은 정상 노화과정에서 나타나는 일반적인 말속도 변화 양상에 대한 구체적 정보를 제공한다. 이에 본 연구는 음향학적 분석 및 청지각적 평가의 통합적 정보를 통하여 노년층의 연령에 따른 발화 속도 변화와 그것이 발화 속도 지각에 어떤 영향을 주는지 알아보고자 하였다.

실험은 읽기 과제와 청지각 평가 두 가지로 진행되었다. 읽기 대상자는 대전 지역에서 30 년 이상 거주한 만 65세 이상 85세 이하(M=75(±6.5))의 여성 정상 노인이었다. 그 중 18 명은 청지각 평가에도 참여하였다. 읽기 과제는 우화 <바람과 해님>을 연구자가 목적에 맞게 재구성하였으며, 청취 자료는 읽기 과제의 일부를 일반 성인 여성(만 29세)이 녹음한 뒤 속도별(빠름, 보통, 느림)로 편집하여 준비하였다. 청자는 연구자가 제공한 말자료(총 15 개)를 듣고, 그 자극이 듣기에 속도가 어떠한지 응답하는 방식으로 진행하였다.

연구 결과 첫째, 연령이 증가할수록 노년층의 발화 속도는 느려지는 것으로 나타났다. 즉, 말속도(speaking rate)와 조음 속도(articulation rate)에서 모두 연령과 음의 상관이 있는 것으로 나타났다. 둘째, 통계 프로그램 R에서 k means 군집화 분석을 가능하게 하는 caret 패키지를 이용하여 상대적으로 느리고, 빠른 두 가지 발화속도 집단을 구분하였다. 각 집단 경계 부분에 해당하는 발화 속도를 선형회귀 방정식(age=-4.6277*speechRate + 95.8439; age=(-4.8045* articulation rate) + 98.9959)에 넣어 발화 속도가 느려지는 기준 연령을 구할 수 있었는데 말속도와 조음 속도에서의 기준 연령은 각각 75.3 세, 76.07 세로 차이가 크지 않았다(그림 1). 이것은 정상 노화 과정에서 쉼(pause)보다 말 운동 조절(speech motor contro)이 미치는 영향이 크다는 해석을 가능케 한다. 셋째, 노년층의 발화 속도 지각(speech rate perception)을 알아본 결과, 노년층은 느린 속도에 대하여 '느리다'라고 느끼고, 빠른 속도 역시 '빠르다'라고 느낀다는

것을 알 수 있었다. 그러나 말 속도 집단 간 비교를 해 보았을 때, 말 속도가 빠른 집단은 보통보다 '느리다'라고 응답한 비율이 더 높았고, 말 속도가 느린 집단은 '보통이다'라고 응답한 비율이 가장 높았다. 이것은 같은 속도의 발화에 대한 노년층 내 속도 지각 차이가 있음을 나타내며, 속도 지각은 개인의 말 속도와 연관이 깊다는 것을 의미한다.

발화 속도는 다양한 장애군의 진단과 중재에서도 중요하게 다루어지는 변수다. 환자의 발화 속도가 정상인에 비해 얼마나 느려졌는지를 가늠하는 것은 중요한 문제이기 때문에 환자와 비교하기 위한 정상 규준이 마련되어야 한다. 이에 본 연구는 정상 노년층 집단을 세분화하여 발화 속도에 대한 정밀한 지표를 제공하고 있다는 점에서 의의가 있다. 또한, 청자로서의 노년층을 설정하여 정상 노인의 보다 수준 높은 의사소통의 질을 확보하는데 기여했다고 볼 수 있을 것이다. 향후 남성을 포함한 더 많은 대상자를 확보하고, 다양한 과제 및 통제 기준이 적용된 정상 노년층 발화 속도에 대한 후속 연구가 필요할 것으로 생각된다.



<그림 1> caret 패키지를 이용한 k means clustering 결과. 조음 속도 4.77 syl/sec과 5.04 syl/sec 부근에서 경계를 형성하며 연령은 대략 75세 내외(74.78 < * < 76.08)에서 집단 분할됨을 보여준다.

5년 이상 한국 거주 외국인 여성 발화에 대한 용인도와 명료도에 관한 연구

박상희, 이상훈
대구사이버대학교 언어치료학과, 구미전자정보기술원

A study on the naturalness and intelligibility of utterances of foreign women residing in Korea for more than 5 years

Sanghee Park[1], Sanghun Lee[2]
[1] Dept. of Speech-Language Pathology, DaeguCyber University, Professor.
[2] Gumi Electronics & Information Technology Research Institute, Principal Researcher
psh4292@dcu.ac.kr, legotech@gmail.com

본 연구의 목적은 한국 거주 5년 이상의 외국인 여성에 대해서 영어교육 전공자들의 청지각적 평가에 관해서 알아보고자 실시하였다. 청지각적인 평가는 용인도와 명료도 2가지로 하였다. 명료도는 청자가 화자의 말을 얼마나 알아들었는가에 관해서 평가하는 것이고, 용인도는 청자가 화자의 얼마나 자연스럽다고 판단하였는가를 보는 것이다. 외국인들이 발화하였을 때 우리가 발화의 내용을 이해는 하지만 한국인들의 억양과 다르다고 생각하는 경우가 있다. 이러한 억양의 다름으로 외모는 한국인이라고 생각이 들더라도 한국인이 아님을 알 수 있게 된다. 일반적으로 억양은 한국인이 외국어를 배울 때도 각 지역의 방언의 톤이 남아 있다. 억양은 고치기 힘든 부분이기도 하지만 한국어를 너무나 자연스럽게 산출하는 외국인들도 있다.

영어를 전공하는 학생들이 외국인들의 발화에 대해서 어떻게 평가하는 가를 알아보는 것은 일반인들과 언어치료 전공자들과 인식의 차이를 연구하기 위한 기초자료로 사용 될 수 있다고 본다. 따라서 최소 5년 이상 거주한 외국인 여성들의 발화에 대해서 평가하였다.

연구방법에서 평가자는 영어교육전공자이며 20대였다. 외국인과 발화 경험이 있는 평가자 4인과 경험이 없는 평가자 6명이었다. 발화자는 한국거주 5년이상되는 외국인 여성이며 8명이었으며 제시되는 그림을 설명하도록 하였다. 발화의 길이를 따로 제한을 두지는 않았다. 평가자는 8명의 발화에 대해서 용인도와 명료도를 5점 척도로 판단하였다.

연구 결과 외국인과 대화경험에 따라서는 5번 화자의 명료도 발화($t=-1.239$, $p<0.01$)를 제외하고는 용인도와 명료도 판단에서 차이를 보이지 않았다. 용인도와 명료도에 대해서는 명료도는 1번화자를 제외하고 평균 4점 이사의 점수를 얻은 반면, 용인도는 1번화자는 평균 1.7이었고, 3번과 8번 화자만 평균 4점대를 넘었다. 즉, 화자의 말을 알아는 들었지만 자연스럽지는 않다고 판단한 것이다. 화자 3번과 8번은 명료도도 높고 용인도도 높게 나타났고 2, 4, 5, 6번 화자는 명료도는 4점이상이었지만 용인도는 2점과 3점대였다.

실제 영어교육을 전공한다고 하여도 한국어의 용인도에 대해서는 일반인과 같은 판단을 하고 있다고 보여진다. 추후 언어치료 전공자와 일반인들이 외국인 발화에 대해서 어떻게 청지각적인 판단을 하는지 비교연구를 할 필요성이 있다고 판단된다. 또한 음성인식 기기가 외국인의 발화에 대해서 어떻게 인식하는가에 대해서도 추가적으로 연구하여 기기개발의 기초자료로 활용되기를 바란다.

다문화 가정 이주여성들의 한국어 파열음 청지각 평가 - 중국, 일본, 베트남, 필리핀을 중심으로-

박선영, 성철재
충남대학교 대학원 언어병리학과

Auditory-Perceptual Assessment of Korean Plosives by Multicultural Immigrant Women

Sunyoung Park, Cheoljae Seong
Dept. of Speech & Language Pathology, Chungnam National University
psy6252@naver.com, cjseong49@gmail.com

한국어 파열음의 조음위치(양순음, 치조음, 연구개음)와 발성유형(평음, 경음, 격음)에 따른 청지각 특성을 다문화 이주민의 출신 국가(중국, 일본, 베트남) 별로 비교하였다. 실험에 참여한 다문화가정 이주여성은 중국인 30명, 일본인 30명, 베트남인 31명, 베트남인 31명으로 총 122명이었다. 한국인 여자 대학생 5명이 녹음한 파열음{바, 빠, 파}, {다, 따, 타}, {가, 까, 카}을 듣고 자극음과 일치하는 항목 하나를 선택하도록 하였다.

다문화가정 이주여성이 파열음 {바, 빠, 파}, {다, 따, 타}, {가, 까, 카}를 인식한 자극과 반응에 대한 표출 값은 contingency table로 정리하였다. 이 contingency table을 토대로 하여 조음위치 (양순음, 치조음, 연구개음)와 발성유형 (평음, 경음, 격음) 별 정반응 백분율 값을 계산하였다.

파열음 청지각 평가 결과 조음위치(양순음, 치조음, 연구개음)에서는 출신 국가별 정반응률이 다양하게 분포하였다(그림 1). 양순음에서 중국인은 /바/와 /빠/, 베트남인은 /파/에서 정반응율이 가장 높았으며, 치조음에서 일본인은 /다/, 중국인은 /따/와 /타/에서 가장 높게 나타났다. 연구개음에서 일본인은 /가/, 중국인은 /까/와 /카/에서 가장 높았다. 발성유형(평음, 경음, 격음)의 경우, 일본인은 평음, 중국인은 경음, 중국인과 베트남인은 격음에서 정반응률이 가장 높았다(그림 2).

조음위치와 발성유형에 따라 중국인은 정반응율이 고르게 분포되어 있는 반면, 일본과 베트남, 필리핀인은 편차가 있었다. 필리핀인은 정반응률 자체가 가장 낮게 나타났다. 청지각적 평가의 정반응률이 높을수록 한국어 말소리 지각 능력에 유리하게 작용할 수 있다고 예상할 수 있으므로 한국어의 특정 음소를 습득하는데 유리하다는 결론을 내릴 수 있다.

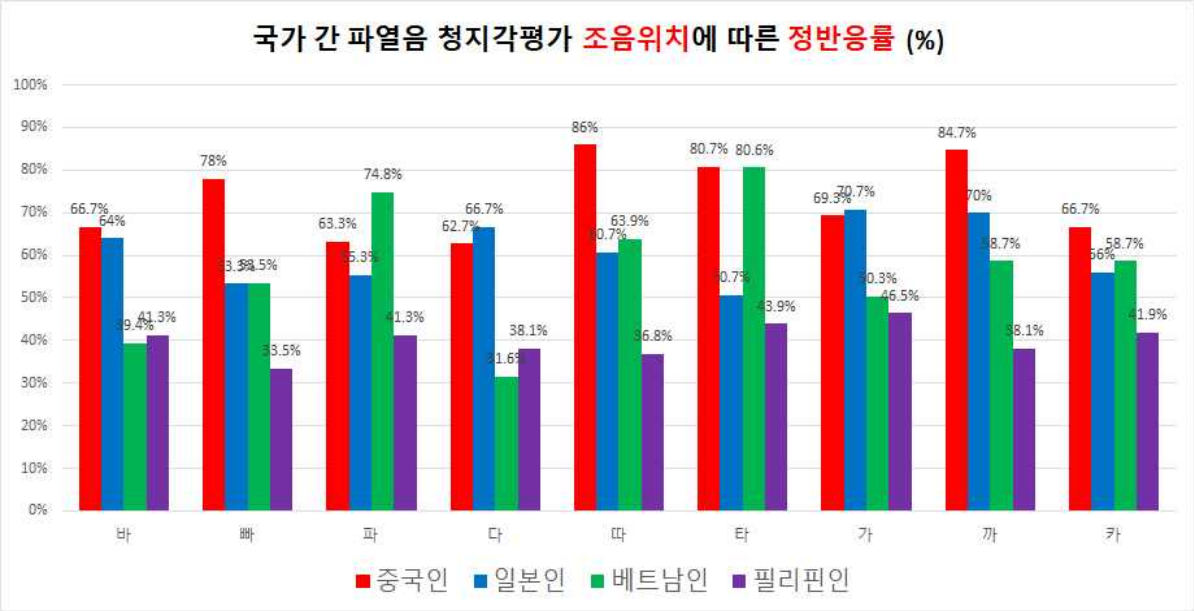


그림 1. 파열음 청지각 평가에서 조음위치에 따른 출신 국가 별 정반응률(%)

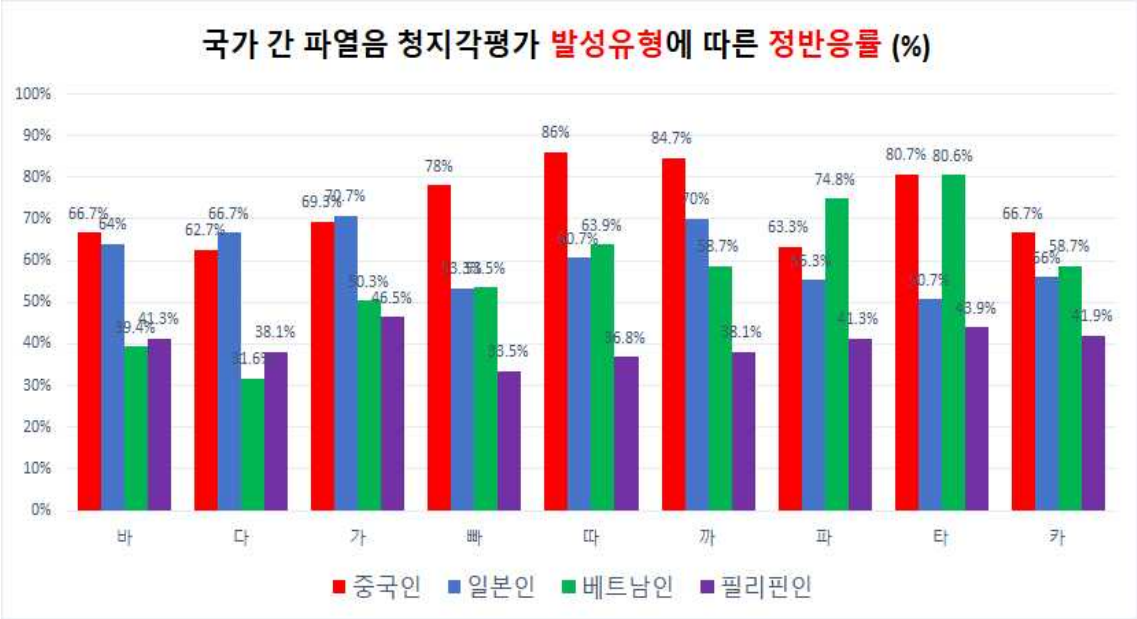


그림 2. 파열음 청지각 평가에서 발성유형에 따른 출신 국가별 정반응률(%)

말소리가 제한된 아동을 위한 말리듬을 이용한 난타 프로그램의 개발과 적용

박 영 혜[*,**] · 최 성 희[*] · 최 철 희[*]
대구가톨릭대학교 일반대학원 언어청각치료학과[*]
아이 앤 맘 언어심리발달센터[**]

Development and Application of Nanta Program using Speech Rhythm for Children with Limited Speech Sound Production

Yeong Hae Park[*,**], Seong Hee Choi[*], Chul-Hee Choi[*]
Dept. of Audiology & Speech-Language Pathology
Daegu Catholic University, Mom Language Psychological Development Center

inm2018@daum.net, shgrace@cu.ac.kr, cchoi@cu.ac.kr

모든 언어에는 고유한 리듬이 있으며, 리듬은 언어 속에 내재된 강세 구조와 운율과 관련이 있다. 난타란, 북과 같은 타악기를 이용한 “두드리기”의 한국 전통 음악인 사물놀이에서 유래된 것으로 리듬과 박자(beat)상황만으로 설정된 비언어극이다(Hoe & Mun, 2013). 특히, 북은 청각적인 감각 자극을 통해 울림이나 진동이 다른 타악기에 비해 강하여 청각 감각을 더욱 잘 자극하며, 감정과 정서를 불러일으킨다(Park & Kim, 2018). 정상 아동의 언어발달은 신생아인 경우 생후 2개월 이전에 말소리를 구분하고, 억양이나 강세와 같은 초분절적인 요소의 변화를 가장 먼저 구별한다(Eimas, 1974).

표현 언어는 대부분의 유아들이 첫 단어를 산출하기 전까지 다양한 발성을 통해 기본적인 욕구나 의사소통의도를 표현하게 된다(Jang & Ha, 2019). 따라서 아동의 언어발달은 분절적인 자모음을 이해하거나 산출하기 이전에 리듬이나 억양과 같은 초분절적 요소가 먼저 발달함을 알 수 있다. 난타는 리듬과 비트로 이루어져 있고, 리듬은 일정한 빠르기나 비트의 강약의 조합으로 이루어져 있어 난타를 이용한 발성은 아동의 언어발달 이전에 선행하는 다양한 초분절적 요소인 리듬이나 억양의 발달과 말발달을 촉진시키는 데 도움을 줄 수 있을 것이다.

본 연구는 무발화 아동 및 말소리 목록이 제한된 언어발달이 늦은 아동에게 적용하는 프로그램으로 고안하였다. 난타 프로그램은 호흡, 음소, 리듬에 의한 3단계의 발음으로 구성되었다. 난타 프로그램은 청각 자극, 다양한 큰 소리, 박자, 리듬을 제공했다. 매회기마다 활동은 도입 5분은 자유롭게 북을 치며 북소리에 관심을 가지고 북을 친근하게 느낄 수 있도록 하고, 그날의 기분이나 상황으로 만남 인사를 만들어 부르고, 프로그램에 대한 소개 시간을 가지도록 하였다. 전개 과정은 30분으로 프로그램 활동을 하며 마무리 활동을 5분 동안 실시하였다. 난타활동 프로그램은 초기 단계(1~4회기), 중기단계(5

~11회기), 종결단계(12~15회기)로 나누어 총 15회기 동안 진행되었다. 초기 단계는 북소리와 난타 리듬을 통하여 북소리를 구음으로 연주하며 흥미와 재미를 유발시키도록 한다.

참여자와 연구자의 친밀감 형성 단계를 거치고 상호교류의 중기단계는 초기단계에서 형성한 연구자와의 안정된 신뢰감을 바탕으로 구음이와 단어놀이를 1음절부터 4음절의 말리듬으로 언어를 입 밖으로 표출할 수 있는 표현의 단계를 거친다. 이를 통해 자기표현을 할 수 있도록 기회를 제공하며 종결단계는 긍정적인 상호교류를 함으로써 자기표현 및 언어표현 능력을 확장시키고자 하였다. 프로그램의 회기별 절차는 <appendix>와 같다. 개발된 프로그램의 유용성을 평가하기 위하여 표현 어휘가 매우 제한적인 6명의 아동이 난타 프로그램 중재에 참여하였다. 참여 아동들에게 소리와 박자를 탐구하고 자유롭게 표현하도록 격려하였다. 아동들이 리듬과 함께 모방하고 모방한 말리듬의 음절 길이를 늘림으로써 음성 소리를 낼 수 있도록 하였다.

1회당 40분씩 총 15회기 (주 2회) 실시되었다. 효과성을 탐색하기 위해 취학 전 수용성-표현성 척도(PRES)와 수용성-표현성 어휘시험(REVT)의 원시 점수를 얻어 치료 전후 비교 하였다. 연구 결과, Wilcoxon Singed Rank test로 중재 전후를 비교한 결과, 중재 후 PRES에서 수용 언어 점수(p=.027)과 표현력 언어 점수(p=.024) 및 한국어 표현 어휘 점수(p=.028) 뿐 아니라 한국어 수용 어휘 점수(p=.028)도 통계적으로 유의하게 크게 향상된 것으로 나타났다. 결론적으로, 개발된 난타 리듬 제어 프로그램은 수용적이고 표현적인 어휘와 언어 발달에 상당히 긍정적인 영향을 끼쳤다. 이러한 발견은 호흡 조절과 발성을 통한 리듬 제어 프로그램이 말소리 목록을 다양하게 하고 표현 어휘 산출 향상에 유용할 수 있음을 시사한다.

핵심어: 말리듬, 난타, 청각적 자극, 비트, 말소리 산출

수위 센서를 이용한 음성치료용 물컵 장치 개발

최 성 희*· 채 혜 림*· 임 금 별**· 김 신 혜** · 윤 종 인**
대구가톨릭대학교 일반대학원 언어청각치료학과, 대구가톨릭대학교 의공학과

Development of a Water Cup Device for Voice Treatment Using Water Level Sensor

Seong Hee Choi, Hye-Rim Chae, ShinHye Kim, Geumbyeol Lim, Jong-In Youn
Dept. of Audiology & Speech-Language Pathology
Dept. of Biomedical Engineering, Daegu Catholic University
shgrace@cu.ac.kr, vpinkponyv@naver.com
sinhye0808@cu.ac.kr, star1208@cu.ac.kr, jyoun@cu.ac.kr

반폐쇄성도기법 중 물저항발성은 과소 및 과다기능성 음성장애 환자에게 유용한 치료법으로 널리 사용되고 있다. 본 연구는 음성장애 환자의 튜브를 이용한 물저항발성치료를 위해 수위에 따라 가변적으로 변하는 저항값을 가지고 있는 다양한 센서를 이용하여 물거품 시 수위의 변화를 정량화할 수 있는 물저항발성 음성치료용 물컵 장치를 개발하였다. 물거품의 높이 변화를 정량적으로 측정하기 위해 물이 담긴 비이커에 아두이노를 활용한 2가지 수위 센서(eTape Liquid Level Sensor, Rain Water Level Sensor)를 사용하여 물저항값을 측정하였으며, 수위 높이에 따른 물저항값과 호기 및 발성지속시간을 표시해 주는 시각적 피드백 곡선을 구현하였다. 수위 센서와 아두이노를 연결하여 코드를 적용시켜 수위센서 작동 여부를 확인하였다. 센서가 물에 흔들리지 않게 하기 위해 Solidworks를 이용하여 설계한 뒤 3D프린팅으로 센서를 고정하기 위한 지지대를 설계하였다. 센서를 비이커에 고정시키고 물을 부어 결과를 확인하였다. 또한, 수위 높이에 따른 시각적 피드백을 제공하기 위하여 LED 모듈 센서를 사용하여 수위에 따른 색깔 변화를 측정하였다. eTape Liquid Level Sensor를 이용한 실험에서는 8inch, 12inch 센서 모두 수위가 높을수록 저항값은 작아졌으며, 수위 센서 높이의 4cm 이상부터 저항값이 변화하였다. 빗물 레벨 센서를 이용한 실험에서는 수위에 따른 저항값에 따라 서로 다른 LED 색깔을 나타내었다. 저항값이 400미만일 경우, 반응이 없었으며, 400 < value ≤ 500일 때 빨간색, 680 < value ≤ 730 일 때 빨간색 LED외에 노란색, 730 초과되면 빨간색, 노란색, 파란색 LED색깔을 보였다. 빨대 지름에 따른 수위 변화는 빨대 지름이 5mm보다 10mm일 때 증폭률이 더 크게 나타났다. 본 연구는 진동센서 생체피드백을 이용한 반폐쇄성도운동 치료 모형을 개발하기 위한 예비연구로서 본 실험을 통해 각 센서의 수위 높이에 따른 물저항값의 변화를 측정하였다. 이러한 결과는 수위센서가 물저항발성 시 물거품의 높이를 객관적으로 정량화할 뿐 아니라 시각적피드백을 줄 수 있음을 시사한다.

핵심어: 수위센서, 음성치료, 물컵 장치

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020S1A5A2a0145868).

포스터 발표 III

음성학

좌장: 손민정(한남대)

영어 원어민의 영어 어휘 강세 발화에 대한 음향적 분석

유지윤, 이석재*
연세대학교 영어영문학과

Acoustic characteristics of English lexical stress by native speakers

Jiyun Yoo, Seok-Chae Rhee*
English Language and Linguistics, Yonsei University
jiyun6843@naver.com, scrhee@yonsei.ac.kr

본 발표는 영어 원어민 화자가 영어 어휘 강세(lexical stress)를 실현하는 데 사용하는 음향 지표(acoustic cues)를 분석해([1]. 강세 음절은 비강세 음절보다 긴 모음의 지속 시간(vowel duration), 높은 기본 주파수(f0), 센 강도(intensity)를 가짐), 영어 어휘 강세를 실현하는 음향 지표의 위계를 분석하는 것을 목표로 하는 연구이다. 영어는 강세 박자 언어(stress-timed language)로 한 음절이 다른 음절에 비해 돋들리는 어휘 강세가 실현되고, 기존 선행 연구들은 어휘 강세를 실현하는 음향 지표의 위계에 대한 서로 다른 결과를 제시해왔다. 본 발표는 어휘의 음절 개수와 강세 음절의 위치에 따라 어휘 강세 실현에서 주요하게 사용하는 음향 지표가 다를 것이라는 가설을 테스트한다. 분석 대상으로는 강세 음절의 위치가 다른 2음절, 3음절, 그리고 4음절 영어 단어를 영어 사전에서 추출한 미국식 영어 원어민 화자의 발화 자료로 하고, 해당 자료에서 어휘의 음절별 음향 지표(모음 지속 시간, 기본 주파수, 강도)를 측정하였다. 분석 결과 ult-stressed, antepenult-stressed words의 경우 모음 지속 시간을 가장 주요하게 사용했고, penult-stressed words의 경우 기본 주파수를 가장 주요하게 사용하였다.

참고문헌

[1] Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *The Journal of the Acoustical Society of America, 27*(4), 765-768.

한국 초등학생의 L2 영어 낭독 발화 평가에서 분절적 · 초분절적 자질이 청자 이해도에 미치는 영향

차 호 윤, 이 석 재
연세대학교 영어영문학과

Effects of Segmental and Suprasegmental Features on the Intelligibility in Korean Elementary English Learners' L2 read-speech

HoYoon Cha, Seok-Chae Rhee
Dept. of English Language and Literature, Yonsei University
hoyoon.heather@gmail.com, scrhee@yonsei.ac.kr

한국 초등학생의 영어 낭독 발화를 평가하는 데 있어 분절적 자질과 초분절적 자질이 청자(평가자) 이해도에 미치는 영향을 발화자의 연령 및 능숙도에 따라 대조적으로 밝히는 데 목적을 두고, 음성 자료의 수집과 평가를 진행했다. 영어 발화자는 총 30명의 한국인 초등학생이고, 각 발화자는 총 24개의 영어 문장을 발화했으며, 이는 연구자와 발화자 일대일 상황에서 녹음되었다. 녹음된 음성파일을 총 3명의 평가자가 각기 세 차례에 걸쳐 '총체적 이해도 평가', '초분절적 자질 평가', '분절적 자질 평가'를 시행했다. 그 결과 한국 초등영어학습자 영어 발음에 대한 청자 이해도에는 분절적 자질과 초분절적 자질 모두 유의미한 영향을 미치지만, 분절적 자질이 더 큰 영향력을 나타냈다. 또한 학습자의 연령과 영어 능숙도와는 무관하게 분절적 자질이 가장 큰 영향력을 나타냈으며, 초분절적 자질 중에서는 리듬 자질이 모든 연령집단과 영어 능숙도가 가장 낮은 집단에서 높은 영향력을 보였다. 억양 및 휴지 자질은 집단에 따라 상이한 영향력을 보였다. 마지막으로, 영어 발화 문장에 대한 청자의 총체적 이해도가 높을수록 청자 이해도에 대한 분절적 자질의 영향력이 높게 나타났으며, 초분절적 자질의 영향력은 총체적 이해도와 무관하게 나타났다.

Durations of Two English Stops across Word Boundaries

Yungdo Yun
Dharma College, Dongguk University
yungdoyun@dongguk.edu

This study discusses durations of English consecutive stops across word boundaries. When two stops are adjacent across a word boundary, one is a word-final stop and the other is a word-initial stop. They share the stop closure. It is known that closure durations of voiceless word-final stops are longer than those of voiced stops in English and the places of articulation of both word-final and word-initial stops do not affect their closure durations. But the VOTs of the word-initial stops are known to be affected by the voicing and their places of articulation. In this study, the shared closure duration of the consecutive stops followed by the VOT of the word-initial stop was measured as the duration of the two consecutive stops. Both English and Korean speakers produced English nonsense word pairs that contain the consecutive stops. The results showed that English speakers' production was affected by both stops; however, they relied more on the voicing and the places of the word-initial stops than on those of word-final stops. But the Korean speakers did not rely on the voicing and the places of articulation of the word-final stops at all due to the influence of the well-known coda neutralization in Korean phonology. They relied only on those of the word-initial stops. The average durations of the consecutive stops produced by the Koreans were longer than those by the English speakers.

응급의료영역 음성대화 데이터 구축

이주영[1], 최서경[2], 지승훈[1], 강지민[2], 김종인[3], 김도희[4], 김보령[5], 조은기[5], 김호정[5], 장정민[4], 김준형[6], 구본혁[6], 박형민[6], 김선희[5], 정민화[1]

서울대학교 언어학과[1], 서울대학교 영어영문학과[2], 서울대학교 인지과학협동과정[3], 서울대학교 외국어교육과[4], 서울대학교 불어교육과[5], 서강대학교 전자공학과[6]

Building Conversational Speech Data in an Emergency Medical Domain

Jooyoung Lee[1], Seo Gyeong Choi[2], Seunghun Ji[1], Jeemin Kang[2], Jongin Kim[3], Dohee Kim[4], Boryong Kim[5], Eungi Cho[5], Hojeong Kim[5], Jeongmin Jang[4], Jun Hyung Kim[6], Bon Hyeok Ku[6], Hyung-Min Park[6], Sunhee Kim[5], Minhwa Chung[1]

Dept. of {Linguistics[1], English Language & Literature[2], Cognitive Science[3], Foreign Language Education[4], French Language Education[5]}, Seoul National University, Dept. of Electronic Engineering, Sogang University[6]

excalibur12@snu.ac.kr, csganna@snu.ac.kr, seunghun.ji@snu.ac.kr, bling1104@snu.ac.kr, prows12@gmail.com, dohee826@snu.ac.kr, jadebr@naver.com, eungi78@snu.ac.kr, hojeong43@snu.ac.kr, jjungmini@snu.ac.kr, imalbert@naver.com, k01032633280@gmail.com, hpark@sogang.ac.kr, sunhkim@snu.ac.kr, mchung@snu.ac.kr

응급실 상황에서의 의료 대화는 환자의 증상을 빠르게 파악하기 위하여 의사의 질문에 대한 환자의 증상을 설명하는 내용으로 이루어진다. 의사는 대화 내용 중 진료에 중요한 정보를 파악하여 기록하지만, 대화 내의 모든 정보를 정확하게 기록하는 것은 어렵다. 또한, 진찰 후 진료 내용 전문을 확인하기도 쉽지 않다. 본 연구는 심뇌혈관 응급환자의 음성·영상 관련 임상정보 자동 추출을 목적으로 서울대학교병원 응급의학과에서 수집 중인 의료 대화 데이터의 수집 방법과 음성인식을 통한 임상정보 자동 추출을 위한 원시 음성 데이터의 가공 방법을 제안한다.

본 연구에서 사용하는 의료 대화 데이터는 양질의 음성 자료 수집, 음성인식 기술을 통한 임상정보 자동 추출 도구 개발, 그리고 최종적으로는 진찰에 필요한 정보를 객관적인 기준에 의해 저장하고 관리하는 정형화 음성 데이터베이스 구축을 목적으로 수집한다. 현재까지의 의료 데이터 분석은 의사의 진찰 메모와 기억에 의존하는 것이 크며, 녹음된 음성 데이터는 시간상의 한계로 활용이 어려웠다. 이를 극복하기 위해 양질의 음성 자료를 수집하여 음성인식 기술을 통해 대화 내용을 텍스트 형태로 변환하고, 최종적으로 환자의 증상과 관련된 임상정보를 객관적인 기준에 의해 자동으로 추출하여 데이터베이스화 하는 것이다.

현재 서울대학교 언어학과에서 보유 중인 음성인식 모델은 음성 학습량으로는 약 7,200여 시간으로 양질의 음성 대화에 대해서는 높은 인식률[3)]을 보인다. 그러나 기계에서 나오는 배경 잡음과 대화 미참여자들의 배경 대화 등이 포함되는 환경에서는 아직 효과적인 어렵기 때문에 의료 대화 인식에 필요한 기초적인 텍스트 데이터 확보가 우선시되어야 한다.

의료 대화 데이터는 현재까지 202건(약 22시간 분량)이 수집되었다. 녹음 환경은 동일 대화에

3) 대화형 음성 데이터 기준 Word Error Rate 약 20% (후처리 이전 기준)

대해 2채널 마이크와 16채널 마이크를 각각 이용하여 수집하였고, 16채널 마이크로 수집된 음성의 경우는 빔포밍(beam-forming) 기법을 통해 환자와 의사 각각의 음성을 효과적으로 녹음할 수 있도록 설계되었다. 또한, 모든 음성은 기본적으로 1채널, 16비트의 bit-depth, 그리고 16,000Hz의 샘플 수로 저장되었다. 의사와 환자/보호자 사이에 거리가 있고, 서로 마주보고 있는 상황이기 때문에 각 대화 참여자의 소리를 효과적으로 녹음하기 위해서는 여러 위치에서 녹음이 가능한 16채널 마이크가 적합하다.

수집 음성 데이터의 텍스트 구축은 Praat[1]의 Annotation 기능을 이용하여 수행하였다. 우선 의사와 환자/보호자의 발화를 구분하여 각각의 Tier로 표시하였고, 표기법대로 전사하는 철자전사(Orthographic Transcription)와 발음 나는 대로 전사하는 청취전사(Auditory Transcription)로 구분하여 전사하였다. 청취전사는 발음열 생성기(KoG2P)[2]를 사용하여 철자전사로부터 자동으로 추출한 다음 결과를 수동으로 검토하였다. 본 데이터는 대화체 발화 특성상 두 명 이상의 화자가 동시에 발화하는 발화 겹침이 발생하는데, 별도의 잡음 Tier를 통해 발화 겹침임을 표시하였다. 또한, 추가적인 정보 파악을 위해 환자 발화에 한하여 방언 발화 구간을 표시한 방언 Tier와 대화 내 각종 잡음을 표시한 잡음 Tier도 함께 기록하였다.

본 연구에서는 심뇌혈관 응급환자의 임상정보 추출을 목적으로 의료 대화 데이터 202건을 2채널 음성과 16채널 마이크로 수집하였으며, 원시 데이터로부터 대화 참여자별 철자전사, 청취전사, 환자의 방어 발화 구간 표시, 발화 겹침 표시, 잡음 표시로 나누어 전사하여 다양한 정보를 포함한 텍스트 데이터를 구축하였다. 구축된 텍스트 데이터는 향후 의료 영역에서의 음성인식 모델의 학습 데이터로 활용될 수 있으며, 환자의 주요 임상정보의 종류와 분포를 구축된 텍스트 데이터를 통해 확인할 수 있다는 점에서 중요하다 할 수 있다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었습니다. (IITP-2020-2018-0-01833)

참고문헌

[1] Paul Boersma, David Weenink (2018): Praat: doing phonetics by computer [Computer program]. Version 6.0.37, retrieved 14 March 2018 from http://www.praat.org/

[2] https://github.com/scarletcho/KoG2P

영-한 음차표기에서 나타나는 음운현상 분석: 음절 말 폐쇄음에서의 모음삽입과 이중모음의 단모음화 현상을 중심으로

전지현, 이석재
연세대학교 영어영문학과

Korean Transliteration of English Words: focusing on vowel epenthesis and monophthongization

Jihyun Jeon, Seok-Chae Rhee
Dept. of English Language and Literature, Yonsei University
wjswlgus6623@gmail.com, scrhee@yonsei.ac.kr

본 연구는 영어 차용어의 한국어 음차표기에서 나타나는 이중모음의 단모음화 현상과 음절 말 폐쇄음에서의 모음삽입 현상에 관한 연구이다. 연구의 목적은 모음삽입과 단모음화 현상을 중심으로 영어 단어의 알파벳 표기를 한글 표기로 전환할 때에 나타나는 실제 영어의 한글 음차표기 실현 양상을 분석하는 데에 있다. 영어 차용어는 한국어로 음차표기 될 때에 음운론적 변화를 겪는데, 영어의 이중모음은 단모음 또는 두 모음의 연접으로 표기되고 음절 말의 폐쇄음은 받침 또는 모음삽입을 동반한 다음 음절의 어두음으로 나타나며 받침과 어두음이 동시에 나타나기도 한다. 이러한 한국어와 영어 간 음운 체계의 차이에서 발생하는 음운론적 변화는 음차표기의 혼용이라는 문제를 일으킨다. 기존의 연구들은 영어차용어의 음운론을 이론적으로 규명하였으나 실제적인 음차표기의 양상을 보여주지는 않는다. 따라서 본 연구에서는 기존의 영어차용어 음운론이 실제로도 적용되는지를 확인하고자 한다. 다섯 가지 단서(유무성성, 음절 수, 단어 내 위치, 이중모음 여부, 조음 위치)를 기준으로 영단어를 선별하여 음절 말 폐쇄음의 모음삽입 현상을 살펴볼 것이며, 세 가지 단서(이중모음의 종류, 모음의 철자 수, 후행하는 말음의 유무)를 기준으로 선별한 영단어로 이중모음의 단모음화 현상을 살펴볼 것이다. 구글 폼의 형태로 설문지의 링크를 배포하여 다양한 성별, 연령대, 출신 지역, 영어 숙련도를 가진 백여 명의 응답자를 대상으로 한 약 3만 개의 데이터를 분석 대상으로 한다. 이를 통해 한국인 영어 학습자가 음절 말 영어 폐쇄음과 이중모음을 음차표기 할 때에 각각의 단서에 따른 영어차용어 음운규칙의 실현 양상을 알아보고 단서 간의 위계 차이를 검증할 것이다. 더불어 앞선 응답자의 속성 정보에 따라 두 음운론적 변화의 실현 양상에 차이가 존재하는지를 규명하려 한다.

한국인 학습자 발화 프랑스어 구강 및 비강모음에 대한 음향적 특징 연구

김 도 희*, 윤 상 아**, 김 선 희***
서울대학교 외국어교육과*, 한국외국어대학교 프랑스어과**, 서울대학교 불어교육과***

Acoustic characteristics of French oral and nasal vowels produced by Korean learners

Dohee Kim*, Sanga Yoon**, Sunhee Kim***
Dept. of Foreign language Eduacation, Seoul National University*
Dept. of French, Hankuk University of Foreign Studies**
Dept. of French Education, Seoul National University***
dohee826@snu.ac.kr, tiddlysa@naver.com, sunhkim@snu.ac.kr

프랑스어 모음의 대표적인 특징 가운데 하나는 비모음이 존재하는 것이다. 현대 프랑스어에서의 비모음은 /ɛ̃, ɔ̃, ɑ̃/의 3개가 있으며 이에 대응되는 구강모음으로는 /ɛ, o, ɔ, a/가 존재한다. 프랑스어 비모음의 음향적 특징은 주로 구강 및 비강모음 대립쌍을 비교 분석하는 연구들로서, 원어민의 비모음은 공통적으로 구강모음보다 후설모음으로 실현되는 것으로 보고되었다(Carignan, 2011, Carignan, 2014; Delvaux et al., 2002). 즉, 음향 실험에서 비모음은 공통적으로 F2 값이 감소하는 경향을 보인다는 것이다. 또한, 전설모음 대립쌍인 /ɛ/-/ɛ̃/에서는 비강모음이 구강모음에 비해 F1은 증가하고 F2는 감소함을 보여 혀높이가 낮아지고 후설로 발음되는 반면에, 후설 원순모음 /ɑ̃ /, /ɔ̃/, /ɔ̃/의 경우는 상응하는 구강모음과 비교할 때 일정한 경향보다는 화자별로 차이를 보이는 것으로 보고하였다(Carignan, 2014; Montagu, 2002). Delvaux et al.(2002)는 후설 원순모음인 /ɔ/-/ɔ~/은 성별에 따라 결과가 다르게 나타나는데 남성의 경우 혀의 위치가 거의 비슷하지만 여성은 비모음이 구강모음보다 후설에 위치하며 때로는 혀높이가 높아지는 것으로 보고하고 있다.

본 연구는 한국어 음운 체계에 존재하지 않는 비모음을 한국인 학습자들이 어떻게 발음하고 있는지 그 음향적인 특징은 규명하는 것을 목적으로 한다. 이를 위하여 한국인 학습자들이 발화한 구강모음과 비모음을 원어민의 발화와 비교하였다. 구강모음과 비모음의 비교를 위하여 /ɛ/-/ɛ̃/, /a/-/ɑ̃ /, /o/-/ɔ̃/ 3개의 대립쌍이 포함된 단어 48개를 유도구문을 이용하여 녹음하였다. 음절 유형은 /CV/ 형태로 다양한 자음 환경을 고려하여 구성하였고 유도구문 « Il a dit CV comme ça. » 에 위치시켜 각 2회씩 녹음하였다. 음성 녹음은 서울대학교 응용언어학 실험실 내 방음실에서 Sennhiser e815S 마이크와 Focusrite Scarlett 6i6 오디오 인터페이스를 컴퓨터에 연결하여 Goldwave 프로그램으로 녹음하였고, 녹음은 44,100Hz, 16bit, mono로 진행하였다.

실험에는 프랑스 원어민 7명과 한국이 학습자 9명이 참여하였는데, 본 연구는 예비연구로서 이 가운데 원어민 2명과 한국인 학습자 2명의 발화를 분석하였다. 분석을 위하여 음향 분석 프로그램 praat(version 6.1.040)을 이용하여 목표로 하는 개별 모음을 분할하고 음성 레이블링을 실시한 후 모음 구간에서 F1, F2, F3의 중간값을 추출한 후 평균값을 비교하였다.

실험 결과 원어민의 경우, 전설 비모음 /ɛ̃/에서 F1은 증가하고 F2는 감소하였으며 F3는 증가하였다. 후설 원순 비모음 /ɑ̃/에서는 상응하는 전설 구강모음 /a/에 비하여 F1과 F2는 감소하고 F3는 증가하였다. 후설 원순 비모음 /ɔ̃/에서는 F1, F2, F3 모두 증가하였다. 대립쌍 /ɛ/-/ɛ̃/에서 F1(t=-12.013, p<.001)과

F2(t=-28.985, p=<.001)에서 차이가 유의미했고 F3는 차이가 무의미했다(t=-1.583, p=.120>.05). 대립쌍 /a/-/ɑ̃/에서는 F1(t=12.141, p<.001), F2(t=8.174, p=<.001), F3(t=-6.143, p=<.001)에서 모두 유의미한 차이가 있었다. 또한, 대립쌍 /o/-/ɔ̃/에서 F1(t=-7.279, p<.001)과 F2(t=-12.058, p<.001)에서 통계적으로 차이가 있었지만 F3(t=-1.986, p=.053>.05)는 그 차이가 무의미했다.

한국인 학습자의 음성의 경우 전설 구강모음 /ɛ/에 비하여 비모음 /ɛ̃/에서 F1은 증가하고 F2 및 F3는 감소하였다. 비모음 /ɑ̃/은 상응하는 구강모음 /a/에 비하여 F1, F2, F3가 모두 감소하였다. 후설 원순 비모음 /ɔ̃/는 모음 /o/에 비하여 F1과 F2는 증가하고 F3는 감소하였다. 통계 분석 결과 /ɛ/-/ɛ̃/에서 F1(t=-12.670, p=<.001), F2(t=19.910, p=<.001), F3(t=2.579, p<.05)에서 유의미한 차이가 있었고 /a/-/ɑ̃/에서는 F1(t=8.290, p=<.001)과 F2(t=5.223, p<.001)에서 통계적으로 차이가 있었지만 F3(t=0.220, p=.826>.05)에서는 무의미함을 알 수 있다. /o/-/ɔ̃/에서 F1(t=-7.694, p<.001), F2(t=3.179, p<.001), F3(t=3.179, p=<.01)에서 모두 유의미한 차이를 보였다.

실험 결과를 요약하면, 두 그룹 모두 세 개의 비모음에서 대응되는 구강모음보다 F1 및 F2에서 차이가 유의미하게 나타났고, F3의 경우 두 그룹 간에 다소 차이가 있었다. 원어민의 경우 전설 원순 비모음 /ɑ̃/과 구강모음 /a/에서 원순성이 뚜렷하게 차이가 있었고 비모음보다 구강모음에서 비교적 원순성이 더 강하게 실현되었다. 반면, 한국인 학습자의 경우 전설 비모음 /ɛ̃/과 후설 원순 비모음 /ɔ̃/에서 F3의 차이가 유의미하게 나타났으며 구강모음보다 비모음에서 원순성이 더 강하게 실현되는 것을 확인하였다.

원어민 음성의 결과는 선행 연구의 결과와 비교하여 /ɔ̃/을 제외하고 /ɛ̃/과 /ɑ̃/에서 비슷한 양상을 보였다. 즉, 비모음 /ɛ̃/과 /ɑ̃/에서 모두 F2 값이 감소하여 후설로 이동하여 실현된 것으로 나타난다. /ɔ̃/의 경우에는 기존 연구에서 대체로 F1과 F2가 감소하였지만(Carignan, C., 2014), 본 연구에서는 F1, F2가 증가하여 다른 결과가 나타났다. 그러나 선행 연구에서도 비모음 /ɔ̃/은 성별 및 화자에 따라 결과의 차이가 다소 존재하기 때문에 본 실험에서도 참여자 수를 더 늘려 결과 양상을 살펴볼 필요가 있을 것이다. 한국인 학습자 음성의 경우는 원어민과 비교하여 비모음에서 세 개의 모음에서 F1과 F2의 변화 양상은 비슷한 경향을 보였으나 원순성을 반영하는 F3에서는 차이를 보였다.

본 연구는 예비 연구로 전체 실험자 총 16명 중 4명의 화자에 대한 결과만 제시한 것으로 이후 좀 더 다양한 학습 수준의 실험자들을 비교 분석한다면 다른 실현 양상이 도출될 것으로 보인다. 이러한 연구는 프랑스어 발음 교육에서 발음 평가와 발음 교정 피드백 자료로 이용할 수 있을 것으로 기대한다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2020-2018-0-01833).

참고문헌

Carignan, C. (2011). Oral Articulation of Nasal Vowels in French. *Proceedings of the 17th International Congress of Phonetic Sciences*, 408-411.

Carignan, C. (2014). An acoustic and articulatory examination of the “oral” in “nasal”: The oral articulations of French nasal vowels are not arbitrary. *Journal of phonetics,* 46, 23-33.

Delvaux, V., Metens, T., & Soquet, A. (2002). Propriétés acoustiques et articulatoires des voyelles nasales du français. *Actes des XXIVèmes Journées d’étude sur la parole*, 348-352.

Montagu, J. (2002). L’articulation labiale des voyelles nasales postérieures du français: comparaison entre locuteurs français et anglo-américains. *Actes des XXIVèmes Journées d’étude sur la parole*, 253-256.

구두 발표 I

음성학 I

좌장: 윤태진(성신여대)

경북 방언 wh 의문문의 작용역에 따른 운율 구조의 변화

윤 원 희
계명대학교 영어영문학전공

Changes of Prosodic Structures of Wh-scopes in Gyeongsang Korean

Weonhee Yun
Dept. of English Language and Literature, Keimyung University
whyun@kmu.ac.kr

경북 방언 내포절의 wh-어구는 모문의 의문문 종결 어미에 따라 모문 작용역을 갖거나 내포문 작용역을 갖는 의문문으로 이해된다. 모문 작용역일 경우 고 평탄조 또는 저 평탄조의 운율 특징을 갖으며, 내포문 작용역 의문문과 다른 억양 패턴을 보인다. 이 연구는 모문 작용역일 경우 의문사 억양의 종류와 관계없이 내포문 작용역인 문장과 구분되는 새로운 단서를 찾고자 하였다. 그 결과 고 평탄조나 저 평탄조와 관계없이 내포문 동사의 F0 정점과 내포문 보문소의 F0 값의 차이는 큰 변화가 없는 반면, 내포문 작용역일 경우 큰 차이를 나타낸다. 또한 모문 동사의 F0 정점과 모문 종결 어미의 F0의 차이도 의문사 억양의 종류와 관계없이 차이가 크지 않으나, 내포문 작용역일 경우 크게 나타난다. 결과적으로 작용역에 따른 운율 특징의 차이는 내포 동사와 모문 동사에서 F0 정점과, 내포 동사와 함께하는 보문소, 그리고 모문 동사와 연결된 종결 어미의 F0 값의 차이로 일관되게 설명할 수 있다.

발화 상대자의 성별·연령이 발화자의 발화량에 미치는 영향

정 희 재, 신 지 영
고려대학교 국어국문학과

The Influence of Interlocutors' Gender and Age on Amount of Talkers' Utterance

Heejae Jeong, Jiyoung Shin
Department of Korean Language and Literature, Korea University
mugen702@korea.ac.kr, shinjy@korea.ac.kr

이 연구의 목적은 한국어 일상 대화에서 성별, 연령의 사회 변수가 발화량의 실현 양상에 어떤 영향을 미치는지 살피는 데에 있다. 구체적으로 발화자와 발화 상대자의 사회 변수에 따라 발화 비율이 어떻게 달리 실현되는지 밝히어 이를 확인하고자 한다.

이를 위해 1) 친밀도와 연령이 통제되고 성별만을 달리하는 40쌍의 대화 자료와 2) 친밀도가 통제되고 성별, 연령을 모두 달리하는 62쌍의 대화 자료를 각기 분석하였다. 각 대화마다 음절 수에 기초한 대화의 참여자별 발화 비율을 산출하였다. 이를 토대로 성별, 연령 변수에 따라 어떠한 사회적 집단이 발화 비율을 더 높게 실현시켰는지 관찰하였다.

그 결과 성별만을 달리하는 집단에서는 발화 비율의 차이가 유의하지 않았다. 마찬가지로 연령만을 달리하는 집단에서도 발화 비율의 차이가 유의하지 않았다. 그러나 성별과 연령 변수를 모두 달리하는 집단에서는 유의한 차이가 나타났다. 이때 유효한 지표는 연장자의 성별이었다. 연장자는 남성일 때에 여성일 때보다 유의하게 발화 비율을 높게 실현하였다. 즉, 남성 연장자 집단은 다른 성별, 연령 집단보다 발화 비율을 높게 실현하였고, 대화 상대자와의 발화량의 차이를 더 컸다.

친밀한 관계의 일상 대화에서 성별과 연령 변수 어느 하나만으로는 발화 비율에 미치는 영향을 충분히 포착할 수 없었다. 발화 비율의 유의한 차이가 '연장자의 성별'에서만 나타났다는 결과는 성별과 연령 변수의 영향을 두루 고려해야만 발화 비율과 사회 변수와의 상호작용을 충분히 포착할 수 있다는 것을 보여 준다.

감사의 글

이 논문 또는 저서는 2019년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2019S1A5A2A03044146)

Acoustics of noise-adapted and clear speech in individuals with elevated depressive symptoms

Hoyoung Yi[1] and Rajka Smiljanic[2]
[1]Speech Language and Hearing Sciences, Texas Tech University Health Sciences Center
[2]Linguistics, The University of Texas at Austin
[1]hoyoung.yi@ttuhsc.edu, [2]rajka@austin.utexas.edu

Introduction

Talkers adapt their speech in response to different challenging conditions including the presence of environmental noise (noise-adapted speech, NAS) or talking to listeners who have difficulty understanding them (clear speech, CS). They spontaneously modify their speech production from hypo- to hyper-articulated forms to facilitate speech communication (H&H theory, Lindblom, 1990). Not all talkers however are equally skilled in producing listener-oriented intelligibility-enhancing speaking style modifications (Yi et al., 2019; Smiljanic & Gilbert, 2017).

Depression is a commonly occurring disorder that affects a wide variety of chronic physical and social disabilities (Kessler & Bromet, 2013). Patients with Parkinson' s disease or stroke, who require speech therapy, often experience depression symptoms (Cummings, 1992; Dickson et al., 2008). Previous research has revealed that listeners benefited less from clear speech modifications produced by talkers with high depressive (HD) symptoms compared to talkers with low depressive (LD) symptoms (Yi et al., 2019).

The current study compared NAS and CS modifications, two common types of responses to communicative challenges, in talkers with elevated depressive symptoms. The findings will provide a better understanding of the nature of communicative deficits in individuals with high depressive symptoms. The findings have the potential to aid speech therapy planning for maximizing intelligibility in talkers with speech sound disorders accompanied with depressive symptoms.

Methods

Five talkers classified as having high depressive (HD) symptoms and five talkers classified as having low depressive (LD) symptoms participated in the experiment. The Center for Epidemiological Studies Depression Scale (CES-D; Radloff, 1977) is a short self-report scale designed to measure depressive symptoms for use with general and clinical populations in order to identify elevated depressive symptoms with high internal consistency. When the participants scored 16 or greater on CES-D, they were classified as having a higher likelihood of a major depressive disorder, though they were not medically diagnosed as clinically depressed at the time of the recording. Participants were classified as having LD symptoms if they scored 15 or lower on CES-D. All 10 talkers read 80 meaningful sentences from the Basic English Lexicon (Calandruccio & Smiljanic, 2012) in each of the three following conditions: 1) conversational speech (CO), 2) noise-adapted speech (NAS), and 3) clear speech (CS). Talkers produced all of the sentences in CO first and then in CS. To elicit clear speaking style, they were instructed to read the sentences as if they were talking to someone who is hard of hearing or a non-native speaker of their language. Next, they were instructed to read the sentences while hearing 6-talker babble noise through the headphones. All talkers were English monolingual speakers. Acoustic analyses included speech rate (syllables per second), mean F0 (Hz), F0 range (Hz), and energy in the 13 kHz range (dB) of all CO, NAS, and CS productions.

Results

Each of the four acoustic outcomes was submitted to a linear mixed effects regression model using R. Fixed effects included depressive symptoms (HD vs. LD), speaking style (CO, NAS, & CS), and an interaction

between depressive symptoms and speech style. Talkers and sentences were included as random intercepts. Wald test was applied to examine the overall effect of main factors and the interaction. The results revealed significant main effects of speaking style and interaction of depressive symptoms with speaking style in all acoustic measures (all p values were smaller than .001). The effect of depressive symptoms was not statistically significant for all acoustics measures. The significant interactions were further investigated using pairwise contrasts evaluated with Bonferroni' s tests using the 'emmeans' function in R. **Speech rate** was significantly slower in NAS and CS compared to CO, and slower in CS compared to NAS for both LD and HD talkers (all p values < .001). Even though both talker groups significantly slowed down in NAS and CS sentences, the effect was greater for LD talkers than for HD talkers. **Energy in 1-3 kHz** was significantly increased in NAS and CS compared to CO for LD talkers (all p values < .001). HD talkers also significantly increased energy in 1-3KHz in NAS compared with CO and CS (all p values < .001) but not in CS relative to CO (p = .062). **F0 mean** was significantly higher in NAS compared to CS and CO for both LD and HD talkers (all p values < .001). LD talkers also showed significantly higher F0 mean in CS compared with CO but HD talkers did not (p = 1.0). **F0 range** was significantly wider in NAS and CS compared to CO for both LD and HD talkers (all p values < .001 except NAS vs. CO for HD: p = .048) but, there were no differences in F0 range between NAS and CS. Both talker groups produced wider F0 range in NAS and CS sentences but, the effect of F0 range modification in NAS and CS was smaller for HD talkers compared to LD talkers.

Discussion

Similar to LD talkers, talkers with HD symptoms implemented NAS modifications in all acoustic measures and CS modifications in speech rate and f0 range. However, HD talkers did not make significant increases in energy in the 1-3 kHz range and in F0 mean when producing CS compared with CO. It appears that HD talkers, similar to LD talkers, could modify their spoken output in response to noise along a number of acoustic-articulatory dimensions. CS results suggest that HD takers may have more difficulty in making conversational-to-clear speech adjustments the absence of a real communication partner. These findings have important implications for speech therapy planning aimed at maximizing intelligibility in individuals with speech sound disorders accompanied with depressive symptoms. While not meant for formal diagnosis, this work demonstrates reduced communication behaviors as observable signals in individuals with depressive symptoms. It remains to be determined how NAS and CS modifications produced by individuals with HD symptoms affect intelligibility outcomes and what other acoustic-articulatory modifications are implemented by HD and LD talkers in CS and NAS. Because of frequent comorbidity of depression in individuals with dysarthria, future work should examine CS and NAS modifications and intelligibility outcomes in the clinical population.

References

Calandruccio, L., & Smiljanic, R. (2012). New sentence recognition materials developed using a basic non-native English lexicon. *Journal of Speech, Language, and Hearing Research, 55*(5), 1342-1355.

Cummings, J. L. (1992). Depression and Parkinson's disease: a review. *The American Journal of Psychiatry, 149*(4), 443.

Dickson, S., Barbour, R. S., Brady, M., Clark, A. M., & Paton, G. (2008). Patients' experiences of disruptions associated with post-stroke dysarthria. *International Journal of Language & Communication Disorders, 43*(2), 135-153.

Kessler, R. C., & Bromet, E. J. (2013). The epidemiology of depression across cultures.*Annual review of public health,34*, 119-138.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In*Speech production and speech modelling*(pp. 403-439). Springer, Dordrecht.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied psychological measurement, 1*(3), 385-401.

Smiljanic, R., and Gilbert, R. (2017). Intelligibility of noise-adapted and clear speech in child, young adult, and older adult talkers. *Journal of Speech, Language, and Hearing Research*. doi:10.1044/2017_JSLHR-S-16-0165

Yi, H., Smiljanic, R., & Chandrasekaran, B. (2019). The effect of talker and listener depressive symptoms on speech intelligibility.*Journal of Speech, Language & Hearing Research,62*(12).

한국인 영어학습자의 단어 및 어구 강세-비강세 음절의 모음 길이 구현과 발음 평가 등급과의 관계

박 기 훈, 박 혜 숙, 이 석 재
연세대학교 교육대학원 조기영어교육전공

The relationship between English pronunciation assessment and the stressed/unstressed vowel duration in words and phrases spoken by Korean elementary English learners

Kihoon Park, Hyesook Park, Seok-Chae Rhee
Division of Graduate School of Education, Yonsei University
manim0317@naver.com, pss7078@daum.net, scrhee@yonsei.ac.kr

본 발표는 한국인 초등 영어학습자의 발음 평가 등급에 따라 한 단어나 어구(phrase)에서 구현되는 강세 음절과 비강세 음절 사이의 모음 길이 실현 양상을 대조 관찰하고, 측정된 등급별 모음길이 비율 차이와 L2 영어 유창성 평가 등급과의 관련성을 고찰하고자 함이다.

연구에 사용된 기반 음성코퍼스는 K-SEC(2004)이며 여기에 초등학생의 문장발화만을 추출하고 유창성 평가과정을 거쳐 5단계의 등급 구분(1등급((Novice)~5등급(Advanced))을 마친 “Rated K-SEC(2018)”의 2음절어 단어와 기능어와 내용어로 구성된 어구의 모음들이 길이 측정의 대상이 되었다.

연구는 측정된 강세-비강세 음절 모음길이 비율값(강세음절 모음길이/비강세음절 모음길이)을 바탕으로 웰치 로버스트 일원배치 분산분석(Welch Robust One-way ANOVA)과 Games-Howell 테스트(Games-Howell Post-hoc Multiple Comparison)를 통해 단어와 어구 내 각각 비강세음절 모음길이 대비 강세음절 모음길이 비율이 유창성 등급 구분별로 아래 결론에서처럼 세 등급으로 유의한 차이를 보인다는 점을 밝히었다.

결론 1: 단어 내 강세-비강세 모음길이 비율에 따른 등급 구분은 [1등급<2등급=3등급<4등급=5등급]으로 통계적 유의함을 보인다.
결론 2: 어구 내 강세-비강세 모음길이 비율에 따른 등급 구분은 [1등급=2등급<3등급=4등급<5등급]으로 통계적 유의함을 보인다.

<단어 내 강세-비강세 모음길이 비율 Games-Howell 사후 검정 결과>

평가 등급(I)	평가 등급(J)	Ratio 평균 차이(I-J)	표준오차	유의확률
1	2	-0.4335	0.1189	0.003**
	3	-0.6059	0.1031	<.0001***
	4	-1.2073	0.1320	<.0001***
	5	-1.5000	0.2175	<.0001***
2	3	-0.1724	0.1026	0.448
	4	-0.7738	0.1317	<.0001***
	5	-1.0665	0.2173	<.0001***
3	4	-0.6014	0.1176	<.0001***
	5	-0.8941	0.2090	0.001***
4	5	-0.2927	0.2247	0.691

<어구 내 강세-비강세 모음길이 비율 Games-Howell 사후 검정 결과>

평가 등급(I)	평가 등급(J)	Ratio 평균 차이(I-J)	표준오차	유의확률
1	2	0.0155	0.1235	1.000
	3	-0.7890	0.1383	<.0001***
	4	-1.1779	0.1759	<.0001***
	5	-3.2069	0.5313	0.001**
2	3	-0.8045	0.0875	<.0001***
	4	-1.1934	0.1395	<.0001***
	5	-3.2224	0.5204	0.002**
3	4	-0.3889	0.1528	0.090
	5	-2.4179	0.5241	0.010**
4	5	-2.0290	0.5352	0.027*

프랑스인 한국어 학습자의 비음성 연구 - 초성 비음 /ㄴ/을 중심으로

이보람
소르본 누벨 대학교 음운·음성학 연구소

Study on the nasality of /n/ in the initial position by French learners of Korean

Boram LEE
Laboratoire de Phonétique et Phonologie, UMR7018, CNRS/ Université Sorbonne-Nouvelle (France)
lee.boram@sorbonne-nouvelle.fr

1. 서론

본 연구는 프랑스인 한국어 학습자를 대상으로 어두 초성 비음/ㄴ/과 어중 초성 비음/ㄴ/의 발화 실험을 통해 프랑스인 학습자와 한국인 모어 화자의 발화 양상을 비교·대조하였다.

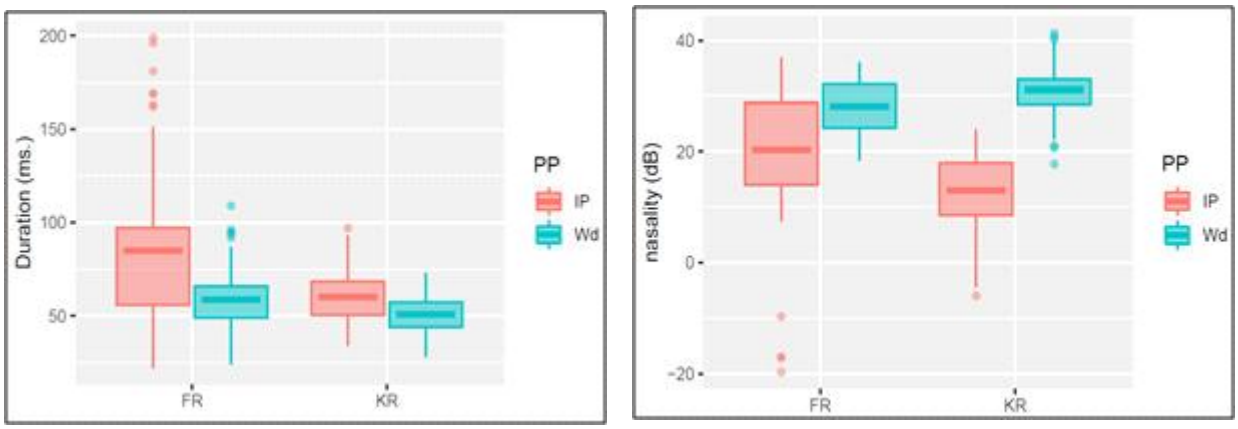
한국어교육에서 비음에 관한 연구는 종성 비음과 비음화에 관한 연구가 주를 이루고 있다. 그런데 최근에 한국어의 denasalization(비음 소실) 현상에 관한 몇몇 연구가 발표되었다(Yoshida, 2008; Kim, 2011; Jang et al., 2018). 이들 연구에 따르면, 어두에 위치한 비음은 비음성이 매우 약하며, 길이도 짧고, 음향학적으로 파열음과 유사성을 보인다. 이러한 한국어의 초성 비음 소실의 특성으로 인해서, 많은 한국어 학습자가 '네'를 '데'로 인식하고, '물'을 '불'로 인식하는 경우를 흔하게 볼 수 있다.

특히, 비음은 자연언어에서 매우 보편적인 자음이기 때문에, 한국어 학습자가 자신의 모국어 비음과 한국어 비음을 어떻게 인식하여 습득하는지 Flegd(1995, 2003)의 SLM(음성학습모델) 관점에서 살펴볼 수 있을 것이다. 프랑스어에도 한국어처럼 세 가지 비음 /n/, /m/, /ŋ/이 존재하며, 한국어와 다르게 비모음이 존재한다. SLM 가설에 따르면, 프랑스인 학습자들은 어중 초성 비음은 한국인과 비슷하게 발화하지만, 어두 초성 비음은 한국인보다 비음성이 크게 실현될 것으로 예측할 수 있다.

2. 연구 방법 및 결과

본 연구를 위해 프랑스인 한국어 학습자 15명과 한국어 모어 화자 10명이 발화 실험(대화문 형식)에 참가하였다. 200개의 실험 문장(어두 초성 /ㄴ/ 2문장 x 어중 초성 /ㄴ/ 2문장 x 2번 반복 x 25명 실험자)을 Pratt로 비음의 길이와 비음성(amplitude P0)을 분석하고, SPSS 18.0으로 독립표본 t-검정을 실시하였다. 결과는 아래 <그림1>과 같다.

<그림1> 언어 화자별 비음 /ㄴ/의 위치에 따른 길이(ms.) 평균값과 비음성(dB) 평균값 비교



두 그림은 각 그룹의 수치를 그래프화 한 것으로, 왼쪽은 /ㄴ/의 길이(ms.)를 나타내고 오른쪽은 /ㄴ/의 비음성(dB)을 나타낸다. FR은 프랑스인 화자를, KR은 한국인 화자의 약자이며, PP는 비음의 위치로 IP는 어두 초성 비음을 Wd는 어중 초성 비음을 의미한다.

실험 결과, 언어 화자별 어중 초성 /ㄴ/의 비음성 평균값은 유의미하지 않았지만, 어두 초성 /ㄴ/의 길이 평균값(t=4.12, df=80, p=. 000)과 비음성 평균값(t=3.62, df=97, p=. 000)은 통계적으로 유의미하게 나타났다. 다시 말해, 프랑스인 학습자는 한국인보다 어두 초성 비음을 더 길게 발화하고 (85ms vs. 60ms), 비음성을 더 강하게 발화한다(20dB vs. 13dB)는 사실을 의미한다.

3. 결론 및 제언

프랑스인 한국어 학습자의 비음의 양상을 살펴보기 위해서 한국어 모어 화자와 어두 초성과 어중 초성 비음 /ㄴ/을 비교하여 살펴보았다. 발화 실험 결과를 통해 프랑스인 한국어 학습자의 어중 초성 비음 /ㄴ/은 한국인과 유사한 양식을 띄지만, 어두 초성 비음 /ㄴ/은 한국인과 비음의 길이와 비음성에서 다르게 나타남을 알 수 있었다. 이를 통해, 프랑스인 한국어 학습자가 어두 초성 비음의 음성 미세 정보(fine detail phonetic) 습득하는 데에 어려움을 겪음을 알 수 있다. 또한 이처럼 비음성이 강한 프랑스인 화자의 발화 양상이 한국어 어두 비음 인지에 영향을 미칠 가능성을 제기한다. 반면, 이처럼 비음성이 강한 어두 초성 비음이 한국인 화자에게 어떻게 인식되는지 추가적인 연구가 필요할 것으로 사료된다. 또한, 본 연구의 결과를 바탕으로 SLM의 관점에서 학습자 한국어 수준별 초성 비음의 인식과 발화에 관한 후속 연구가 필요할 것이다.

[참고문헌]

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, *92*, 233-277.

Flege, J. E. (2003). Assessing constraints on second-language segmental production and perception. *Phonetics and phonology in language comprehension and production: Differences and similarities*, *6*, 319-355.

Jang, J., Kim, S., & Cho, T. (2018). Focus and boundary effects on coarticulatory vowel nasalization in Korean with implications for cross-linguistic similarities and differences. *The Journal of the Acoustical Society of America*, *144*(1), EL33-EL39.

Kim, Y. S. (2011). *An acoustic, aerodynamic and perceptual investigation of word-initial denasalization in Korean* (Doctoral dissertation, UCL (University College London).

Yoshida, K. (2008). Phonetic implementation of Korean denasalization and its variation related to prosody. *IULC Working Papers*, *8*(1).

구두 발표 II
말장애 및 음성의학

좌장: 김재옥(강남대)

심한 말소리장애아동에게 적용한 통합치료접근법(ITP-SSSD)이 말소리 비일관성과 음운산출능력에 미치는 효과

고 유 경[1], 김 수 진[2]
[1]바른소리언어치료센터, [2]나사렛대학교

Effects of The Integrated Treatment Program for Chidren with Severe Speech Sound Disorders(ITP-SSSD) on Speech Inconsistency and Phonological Production Ability

Yoo-Kyeong Ko[1], Soo-Jin Kim[2]
[1]Barunsori Speech-Language Clinic, [2]Korean Nazarene University
bobgog@hanmail.net, sjkim@kornu.ac.kr

초록

본 연구는 말소리 비일관성, 음운오류패턴, 조음문제를 동시에 보이며, 언어장애를 동반한 심한 말소리장애아동을 위하여, 통합치료접근법을 개발하여 근거 기반에 의한 치료효과를 검증하고자 하였다. 대상은 만 3-5세의 심한 말소리장애아동으로 대상자간 중다간헐기초선 실험설계를 적용하였다. 연구결과, 말소리 비일관성이 감소되었고, 음소정확도가 증가되는 치료효과를 보였다. 평균음운길이와, 단어단위 정확률도 증가되는 양상을 나타냈다. 본 연구는 말소리 비일관성, 음운오류패턴, 조음문제를 복합적으로 보이며, 언어장애를 동반한 심한 말소리장애아동이 언어문제의 제한 없이, 복잡한 말소리문제들을 해결할 수 있는 통합치료접근법을 개발하여, 근거기반에 의한 치료 효과를 확인하였다는 점에서 의의가 있다.

● 이 논문은 제 1저자의 박사학위논문(2020)을 요약한 것임.

20-30대 베트남 결혼 이주여성들의 한국어 운율 특성[4)]

김난숙, 성철재*
충남대학교 언어치료센터, 충남대학교 언어학과

Prosodic Characteristics of Korean observed in Vietnamese Married Immigrant Women in their 20s and 30s

Nansook Kim, Cheoljae Seong
Speech & Language Clinic, Chungnam National University
*Linguistics, Chungnam National University
kidsmirae@gmail.com, cjseong49@gmail.com

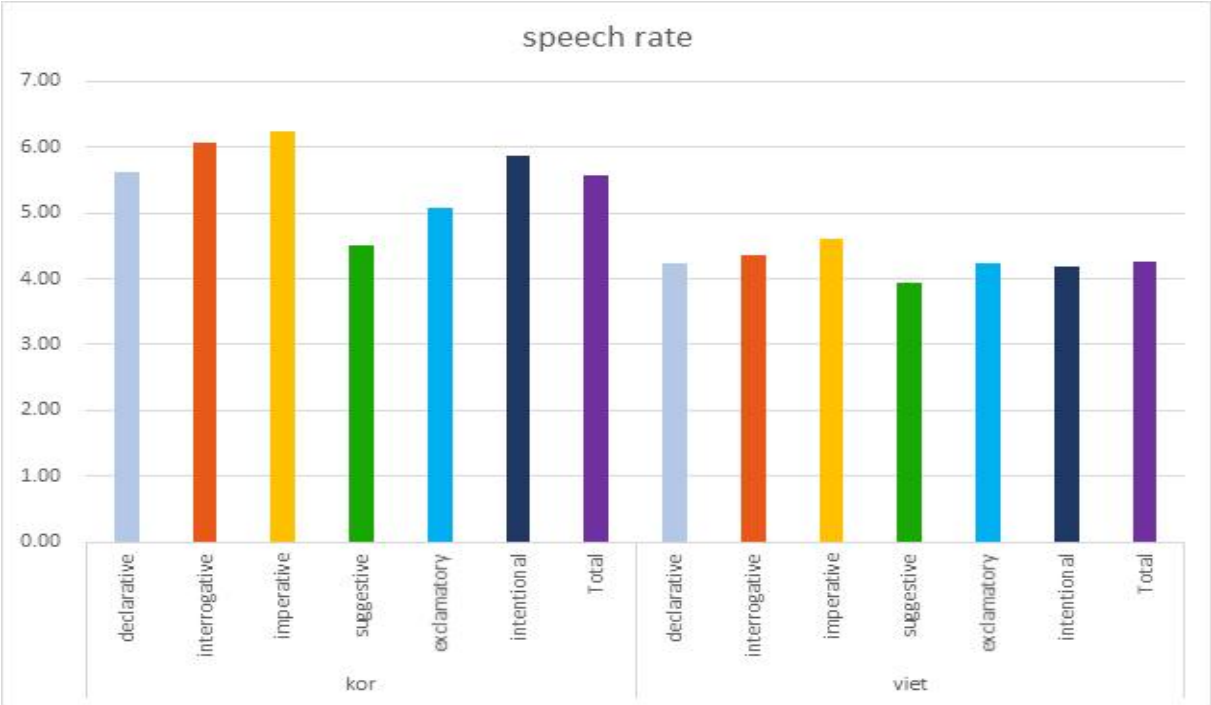
결혼으로 인한 이주여성들 중에서 가장 높은 비율을 차지하고 있는 베트남인 이주여성들의(통계청, 2018) 한국어 운율 특성을 알아보고, 이들의 한국어 운율을 평가할 수 있는 도구 개발의 기초를 마련하고자 본 연구를 수행하였다. 베트남인 결혼 이주 여성 20-30대 51명과 한국인 20-30대 여성 103명을 대상으로 하였으며, 청지각 평가를 통하여 선정된 한국인 상위 51명과 베트남인 51명의 발화 자료를 최종 분석에 사용하였다.

발화 자료를 위한 6가지 실험과제는 직접화법형으로 읽도록 하였고, 분석에 사용된 총 52문장 중 17문장을 선택하여 두 집단의 발화 자료에 대한 청지각 평가를 하였다. 복합명사가 포함된 7문장을 제외한 45문장들은 음향분석을 통한 운율 분석을 하였으며, 운율 분석에 사용된 변수는 총 30개로 음도, 강도, 말속도, 음도 기울기 관련 변수들이다.

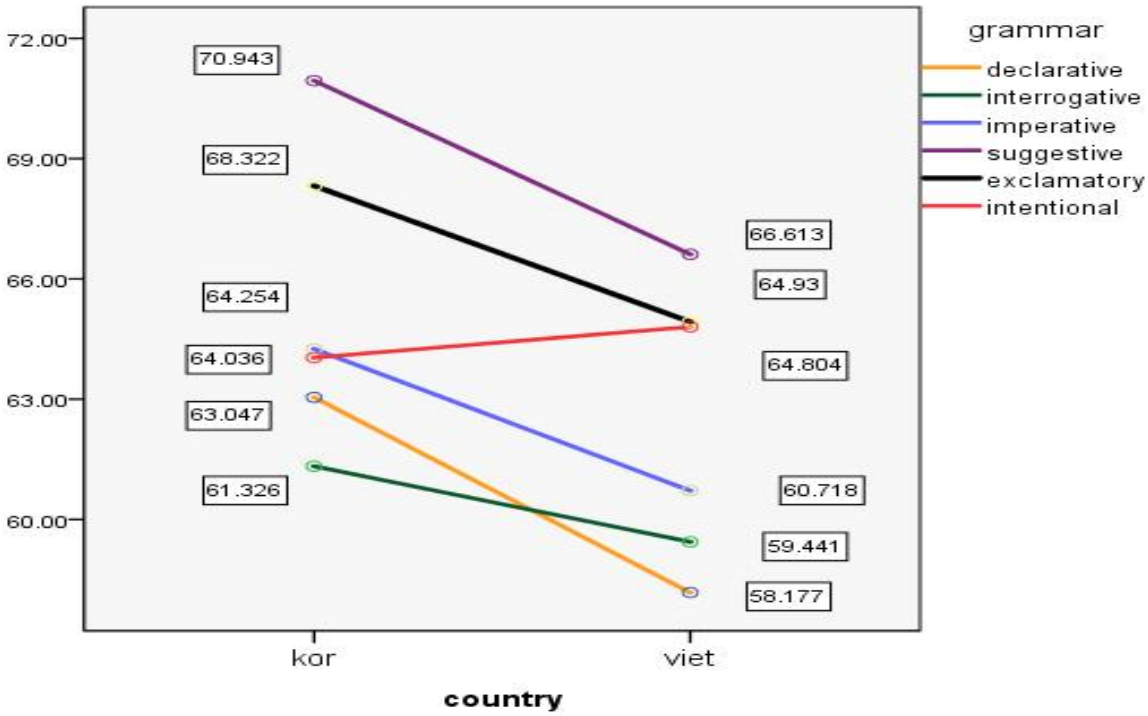
운율 분석 결과는 다음과 같다. 문두 2음절에서 집단(국가), 첫 음절의 발성유형, 음절 구조에 따른 분석을 한 결과, 베트남인들은 강음 CVC 구조일 때 두 번째 음절이 낮아진데 비하여 한국인은 첫 번째 음절이 강음일 때 두 번째 음절의 음도가 더 높아졌다. 문형에 따른 음도 기울기는 한국인의 경우 평서문의 음도 회귀 기울기와 음도 변화율 기울기가 다른 문형들에 비하여 유의하게 작았다. 한국인의 문형별 말속도는 명령문 > 의문문 > 의도문 > 평서문 > 감탄문 > 청유문 순으로 템포가 빠른 반면, 베트남인의 문형별 말속도는 명령문 > 의문문 > 의도문 > 감탄문 >평서문 > 청유문 순으로 빨랐다.

문형에 따른 문미 억양구의 상대적 발화시간 비율은 한국인의 경우 청유문 > 감탄문 > 명령문 > 의도문 > 평서문 > 의문문 순으로 길게 나타났으며, 베트남인의 경우에는 청유문 > 감탄문 > 의도문 > 명령문 >의문문 > 평서문 순으로 길게 나타났다. 한국인은 선택의문문의 말속도와 음도비가 유의하게 크게 나타났지만 베트남인들은 의문문 유형간 말속도와 음도비에 유의한 차이가 없었다. 음도비와 강도비는 모두 한국인이 유의하게 작았는데, 이는 베트남인의 문미 강세구의 음도와 강도가 높다는 것을 의미한다.

4) 이 논문은 제 1 저자 박사학위 논문의 일부를 수정하고 보완한 것임.



<그림 1> 국가별 문형에 따른 말속도



<그림 2> 국가별 문형에 따른 상대적 발화시간 비율

뇌성마비로 인한 마비말장애 성인의 자음정확도와 말명료도 비교연구

여은정, 정민화
서울대학교 언어학과

Comparative Study of Percentage of Correct Consonants and Speech Intelligibility in Adults with Dysarthric Adults with Cerebral Palsy

Eun Jung Yeo, Minhwa Chung
Dept. of Linguistics, Seoul National University
ej.yeo@snu.ac.kr, mchung@snu.ac.kr

1. 서론

뇌성마비는 중추신경계통 손상으로 나타나는 운동장애로, 말소리 산출기관 조절에 문제가 있을 경우 마비말장애를 동반하여 말명료도가 저하되기도 한다. 말명료도란 화자가 의도한 메시지가 청자에게 성공적으로 전달되는 정도로, 말 문제에 있어 중요한 척도로 간주된다[1].

국내에서 마비말장애 장애정도를 판정할 때는 조음 능력, 그중에서도 객관적 지표인 자음정확도에 중점을 두고 있다[2]. 선행연구는 마비말장애 화자의 자음정확도와 말명료도 간 높은 상관관계를 확인하였다[3]. 그러나 자음정확도만으로 말명료도를 얼마나 설명할 수 있는지에 대한 연구는 미비한 실정이다.

본 연구는 자음정확도의 말명료도에 대한 설명력을 알아보고자 한다. 이를 위해 마비말장애인 70명을 대상으로 자음정확도 기반의 조음중증도와 말명료도를 비교분석하고, 조음중증도 집단 별 자음정확도와 말명료도 간 상관분석 및 단순회귀분석을 실시한다.

2. 분석 대상

2.1 음성 데이터와 조음중증도 점수

본 연구에서는 마비말장애인 발화의 말명료도를 분석하기 위하여 QoLT 마비말장애 DB[4]를 사용하였다. 마비말장애 화자 70명의 음성과 화자별 APAC 조음 평가 결과를 4점 척도로 환산한 조음중증도 점수[5)]가 사용되었다. 각 조음중증도 집단 별 화자는 51명, 13명, 4명, 2명이었다. 음성 자료는 문장 읽기 과제로, 화자 당 5개의 문장을 2회 반복한 자료이다.

2.2 말명료도 점수

말명료도 평가는 2년 이상의 임상 경험이 있는 언어재활사 5명이 실시하였다. 한 화자의 문장을 모두 듣고 5점 척도의 청지각적 평가(0-정상, 1-경도, 2-경도-중등도, 3-중등도-중도, 4-중도)를 하였다. 평가는 일주일 간격으로 두 차례 진행되었고 화자의 순서는 무선배치되었다.

각 화자에 대한 항목 평가 점수는 가장 빈도수가 높은 점수가 선택되었다. 각 말명료도 집단 별 화자는 4명, 21명, 26명, 12명, 7명이었다.

3. 분석 결과

조음중증도와 말명료도의 비교를 위해 말명료도 평가에서 정상으로 평가된 화자 4명은 분석에서 제외되

5) 1-경도; 85~100%, 2-경도-중등도; 65~84.9%, 3-중등도-중도; 50~64.9%, 4-중도; 50% 미만

었다. 조음중증도 점수와 말명료도 점수를 비교분석한 결과, 일치도는 .44로 나타났다.

표 1은 조음중증도 별 말명료도를 제시한 것으로, 중도 조음중증도 집단을 제외한 모든 집단의 화자 50% 이상이 말명료도에서 더 낮게 평가된 것을 보여준다. 이는 높은 자음정확도로 인해 말명료도가 낮음에도 장애 정도가 과소평가될 수 있음을 시사한다. 반면 자음정확도가 50% 이하일 경우에는 자음정확도만으로 말명료도를 설명할 수 있음을 암시한다.

표 1. 조음중증도 별 말명료도 분포

조음중증도	말명료도	빈도	비율(%)	조음중증도	말명료도	빈도	비율(%)
경도(1)	**1**	**21**	**41**	중등도-중도(3)	1	0	0
	2	20	39		2	0	0
	3	6	12		**3**	**2**	**50**
	4	0	0		4	2	50
경도-중등도(2)	1	0	0	중도(4)	1	0	0
	2	**6**	**46**		2	0	**0**
	3	4	31		3	0	0
	4	3	23		**4**	**2**	100

또한, 자음정확도가 말명료도에 미치는 영향을 알아보기 위해 상관분석 및 단순회귀분석을 실시하였다. 이때 말명료도 점수는 각 화자에 대한 평가 10개의 평균값이 사용되었다. 전체 화자 66명의 자음정확도와 말명료도를 상관분석한 결과, 두 점수 간에 유의미한 상관관계가 있었다, r= -.75, p < .05. 그러나 조음중증도 집단 별 자음정확도와 말명료도의 상관관계를 살펴보았을 때, 경도 조음중증도 집단만 유의했으며 약한 상관관계를 보였다, r= -.54, p < .05. 단, 화자 2명만이 속한 중도 집단은 통계적 의미를 계산할 수 없었다.

전체 화자 66명에 대한 단순회귀분석을 실시한 결과, 자음정확도는 결정 계수 R^2= .57의 설명력으로 말명료도를 유의미하게 예측하였다, $F(1, 64) = 84.55$, p < .05. 경도 조음중증도 집단의 경우, 자음정확도는 결정 계수 R^2= .28의 설명력으로 말명료도를 유의미하게 예측하였다, $F(1, 45) = 18.48$, p < .05. 이는 자음정확도가 높은 집단의 경우, 자음정확도의 말명료도에 대한 예측력이 떨어짐을 보여준다.

4. 결론

본 연구는 뇌성마비 마비말장애인의 자음정확도 별 말명료도가 어떤 분포로 나타나는지, 말명료도에 대한 자음정확도의 설명력은 어떠한지 살펴보았다. 그 결과, 자음정확도가 50% 미만으로 현저히 떨어질 경우에는 자음정확도만으로 말명료도를 판단할 수 있지만, 그 이상일 경우에는 정보가 충분하지 않다는 것으로 나타났다. 본 연구의 결과는 소수의 데이터를 분석한 것으로 뇌성마비 마비말장애인 집단을 대표하기에 제한적이지만, 말명료도에 대한 자음정확도의 전반적인 설명력을 알아보았다는 것에 그 의의가 있다.

감사의 글

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 연구개발지원사업으로 수행되었음(과제번호: R2019080018)

참고문헌

[1] 이옥분, 박상희, 남현욱. (2012). 뇌성마비 화자의 말명료도 매개변수 간의 상관성 연구. 언어치료연구, 21(3), 115-126.

[2] 홍새미, 정필연, & 심현섭. (2018). 마비말장애 발화의 청지각적 평가방법 비교: 세부평가와 범주평가. *Communication Sciences & Disorders, 23*(1), 242-253.

[3] 김수진 (2002). 언어장애인의 명료도에 영향을 미치는 말요인: 문헌연구. 말소리, 43, 25-44

[4] Choi, D. L., Kim, B. W., Lee, Y. J., Um, Y., & Chung, M. (2011). Design and creation of dysarthric speech database for development of QoLT software technology. *Oriental COCOSDA*, 47-50.

공명튜브발성 시 물의 저항에 따른 성대진동특성

채 혜 림, 최 성 희
대구가톨릭대학교 일반대학원 언어청각치료학과

Charateristics of Vocal fold Vibration depending on the Water Resistance in Resonance Tube Phonation

Hey-Rim Chae, Seong Hee Choi
Dept. Audiology and Speech-Language Pathology, Daegu Catholic University
mona@cu.ac.kr, shgrace@cu.ac.kr

튜브나 빨대를 물에 넣어 물의 저항을 이용하는 음성치료 기법인 물저항 치료(water resistance therapy, WRT)는 반폐쇄성도훈련(semi occluded vocal tract exercises, SOVTE) 중 하나이다. 튜브나 빨대를 물에 넣는 깊이에 따라 저항이 달라지는데 얕은 깊이에서는 저항이 낮고 깊이가 깊을수록 저항이 증가하며 성대의 내전 또한 증가한다고 알려져 있다(Guzman et al., 2015b; Simberg & Laine, 2007). 그러나 전기성문파형검사(electroglottography, EGG)를 활용한 선행연구들을 살펴보았을 때 물의 깊이에 따른 성대접촉 변화는 연구자마다 상이한 결과가 보고되었다(Andrade et al., 2014; Guzman et al., 2015a; Tyrmi et al., 2017). 따라서, 본 연구에서는 EGG를 이용하여 물저항 발성 전과 4가지 물의 깊이(2cm, 4cm, 7cm, 10cm)에 따른 성대접촉의 변화를 살펴보고자 하였다. 성대의 구조적 이상 및 기질적인 병변이 없는 정상성인 9명(남자 5명, 여자 4명, 평균연령 24.89±세3.62)을 대상으로 물저항 발성 전 모음 /우/와 4가지 깊이에서 물저항 발성을 실시하는 동안 EGG를 이용하여 성문폐쇄율(closed quotient, CQ), 성문개방율(opened quotient. OQ), 성문접촉 속도율(speed quotient, SQ)을 측정하였다. 물의 깊이에 따른 차이를 알아보기 위해 반복측정 분산분석(repeated measures ANOVA)을 실시하였다. 연구 결과, 물의 깊이가 깊어질수록 CQ와 SQ는 유의하게 증가하였고, OQ는 유의하게 감소하였다. CQ와 OQ는 물저항 발성 전과 깊이 2, 4, 7, 10 cm에서 유의한 차이가 있었으며, 깊이 10 cm와 깊이 2, 4, 7 cm와 유의한 차이가 있었다(p < .05). SQ는 물저항 발성 전과 깊이 2, 4, 7, 10 cm에서 유의한 차이가 있었으며, 깊이 10 cm와 깊이 2, 4 cm에서 유의한 차이가 있었다(p < .05). SOVTE 시 성도를 좁힙으로써 음향 임피던스가 증가하게 된다. 그 결과 음향-공기역학적 상호작용으로 성문 기류 저항과 성대의 음압에 영향을 주게 되어 glottal flow가 빨라지게 되는데, 결국 물의 깊이에 따른 저항의 증가로 인해 성대의 저항이 더욱 증가하게 되어 이러한 결과가 나타난 것으로 사료된다.

핵심어: 튜브 깊이, 물 저항, 성대 진동, 전기성문파형검사

Python을 이용한 아동용 단모음 조음 훈련 프로그램

김 하 정, 성 철 재
충남대학교 언어병리학과, 충남대학교 언어학과

The Monophthong Vowels Articulation Training Program for Children Using Python

Hajung Kim, Cheoljae Seong
Speech-Language Pathology, Linguistics, Chungnam National University
sandou@hanmail.net, cjseong49@gmail.com

조음음운장애 중 음운 발달에 대한 연구는 주로 자음에 집중되어 왔다. 이는 모음이 다양한 음향적 특성을 지닌 자음과 달리 서로 비슷한 진동주기를 가져 지각하기 쉬우며, 정확한 위치에서 조음되지 않더라도 오조음으로 인식하기보다는 넓은 관용으로 받아들여질 수 있기 때문이다. 하지만 실제 언어치료 현장에서는 모음 오류가 흔히 관찰되지만 적절한 검사도구나 치료방법의 부재로 어려움을 겪기도 한다(박성지, 2008).

이에 본 연구에서는 아동을 대상으로 모음 포먼트 값을 활용한 단모음 조음 훈련 프로그램을 제안한다. 모음 오류가 있는 아동의 모음 특성을 포먼트를 이용하여 분석하고, 정상 아동의 모음별 평균 포먼트 값과 비교하여 객관적이고 구체적인 개선 방안을 제시하는 것이 연구의 목적이다. 이를 위해 아동의 모음에 대한 음향음성학적 연구를 분석하고 모음 연구에서 포먼트가 차지하는 비중을 다루었다.

모음 포먼트 중 제 1포먼트와 제 2포먼트는 모음의 특성을 대표하며, 이러한 포먼트 값의 비교를 통해 아동의 모음을 분석할 수 있다. F1 주파수는 협착이 성도의 전반부에서 이루어질수록 낮아지고, 후반부에서 이루어질수록 높아진다. F2 주파수는 성문 쪽과 경구개 쪽 가까이에서 협착이 일어날수록 높아지고, 반대로 입술이나 연구개 쪽 가까이에서 협착이 일어날수록 낮아진다. 즉, F1은 모음의 개구도와 관련이 있으며, 저모음일수록 높아지고 고모음일수록 낮아진다. F2는 모음의 전후설과 관련이 있다. 전설모음일수록 F2가 높으며, 후설모음일수록 F2가 낮아진다.(조성문, 2003). 아동이 모음을 정확하게 산출하기 위해서는 모음의 F1과 F2 값이 정상 범위 내에 포함되어야 한다.

단모음 조음 훈련 프로그램은 이러한 모음의 특성을 반영하여 범용 프로그래밍 언어인 Python을 활용하고 Praat의 기능을 구현한 Parselmouth 라이브러리를 기반으로 만들어졌으며, 음성을 직접 녹음하여 분석하는 사용자용 프로그램과 음성파일을 불러들여 분석하는 연구자용 프로그램으로 구분하여 제작되었다. 사용자용 프로그램은 실제 언어치료 현장에서 사용할 수 있으며, 한 번의 시도에 한 명의 아동을 대상으로 정보를 시각화하여 표현한다. 연구자용 프로그램의 경우, 한 번에 많은 양의 음성파일을 불러들여 각각의 정보를 간단하게 텍스트로 저장할 수 있다는 장점이 있다.

프로그램의 환경설정에서는 사용 목적에 따라 Slow Version과 Fast Version을 선택할 수 있

다. 두 가지 버전은 속도에 따른 차이를 보이는데, Slow Version은 속도는 느리지만 이론적으로 더욱 정확한 포먼트 값을 측정할 수 있으며, Fast Version은 매우 빠른 속도로 포먼트 값을 추출할 수 있으나 Slow Version에 비해 정밀도가 약간 낮은 편이다. 프로그램 초기 설계 단계에서는 연구자를 위해 정밀도가 높은 Slow Version을 먼저 제작하였다. 한편, Slow Version이 정밀도는 높으나 속도가 느리다는 점을 참고하여 언어치료 현장에서 사용할 경우, 임상가의 편의를 위해 Fast Version을 추가 제작하였다. Fast Version은 분석시간의 단축과 아울러 일정 수준 이상의 신뢰성 있는 결과를 도출하도록 구현하였다.

구현한 프로그램이 모음 포먼트 값을 정확하게 판단하고 분석 가능한지 판단하기 위해 모음 424개를 대상으로 Praat에서 추출한 값과 비교하였고, 그 결과, Praat에서 추출한 값과 본 연구에서 구현한 프로그램이 추출한 값이 일정한 수준 내에서 동일한 값을 나타냈다. 포먼트 산점도를 이용하여 진행한 종합적인 평가에서도 각 모음이 위치 별로 뚜렷하게 분포하는 것을 알 수 있다. 연구에서 적용한 알고리듬은 박지연 & 성철재(2018)를 참고하였다[3].

두 번째로, 아동용 단모음 조음 훈련 프로그램이 사용자를 고려하여 편의성, 기술성, 전문성, 디자인, 활용성 등 5가지 항목에서 적절하게 설계되었는지 평가하기 위해 언어병리 전문가 9명을 대상으로 설문조사를 실시하였다. 그 결과, 편의성, 기술성, 전문성, 활용성 등 4가지 항목이 6(만족), 디자인의 한 가지 항목이 5(약간 만족)에 가까운 값을 나타냈다. 언어병리 전문가에 의해 진행된 본 연구에 대한 사용자 평가는 대체로 '만족' 수준이었다.

음성 평가는 다차원적이어서 주관적인 평가와 객관적인 평가를 모두 포함하여 이루어져야 한다. 음향음성분석 프로그램을 이용한 방법은 사람의 귀를 이용한 청지각적 분석의 주관성을 보완하는 장점이 있다. 아동용 단모음 조음 훈련 프로그램은 임상에서의 객관성 확보로 인해 언어재활사가 정확한 피드백을 제공하여 아동의 조음 훈련에 도움을 줄 것으로 예상된다. 언어재활사와 모음 재활을 목표로 하는 대상자가 훈련을 보다 쉽게 진행할 수 있을 것이다.

참고문헌

[1] 조성문(2003). 현대 국어의 모음 체계에 대한 음향음성학적인 연구. *한국언어문화. 24*. 427-441.

[2] 박성지(2008). 아동 모음 오류의 포먼트 분석의 타당성에 관한 연구. *한국언어치료학회. 17*(3). 117-131.

[3] 박지연 & 성철재(2018). Praat을 이용한 아동 포먼트 자동 세팅 스크립트 구현. *말소리와 음성과학 10*(4). 1-10.

구두 발표 III

음성학 II

좌장: 고언숙(조선대)

후설모음 /ㅡ, ㅜ, ㅓ, ㅗ/에 대한 음향음성학적 연구: 선행 인접 자음과의 상관성을 중심으로

조 남 민, 황 미 경
한국기술교육대학교 교양학부, 이화여자대학교 국어국문학과

Acoustic Phonetic Study of Back vowel /ㅡ, ㅜ, ㅓ, ㅗ/: Focused on the Correlation with Adjacent Preceding Consonants

Nam Min Cho, Mikyeong Hwang
School of Liberal Arts, Koreatech, Dep. of Korean Linguistics and Literature, Ewha Univ.
nmcho93@koreatech.ac.kr, hmkbia@gmail.com

본고는 한국어 표준어 화자의 후설모음 /ㅡ, ㅜ, ㅓ, ㅗ/ 발화에 대한 음향음성학적 특징에 있어 인접한 선행 자음이 후설모음의 실현에 미치는 영향과 그 원인을 규명하고자 하였다. 후설모음 /ㅡ, ㅜ, ㅓ, ㅗ/의 음가가 점진적으로 변화함에 따라 각 모음의 음향공간의 겹침 현상에도 변화가 나타나는데 이러한 음역 겹침이 어떤 조건하에서 일어나는지에 대한 면밀한 분석의 필요성이 제기된다.

이에 본고는 후설모음이 나타나는 음운 환경을 인접 선행 자음의 음성적 특징에 따라 분류하여 실험을 진행하였고, 그 과정에서 선행 자음과 후설모음으로 구성된 음절의 단어 내 위치 등을 고려하여 음향음성학적 분석을 시행하였다. 이러한 분석을 바탕으로 선행 자음에 따른 후설모음의 음향공간의 변화 양상을 제시하였고, 선행 자음의 특성과 후설모음의 포먼트의 관계에 대해 논의하였다.

실험 결과 마찰음이 선행한 경우 높은 F1 값과 F2 값을 나타냈으며, 파찰음이 선행한 경우 F2 값에서 큰 차이를 보였다. 유음과 비음이 선행한 경우 F1 값은 낮게 형성되지만, F2 값은 상대적으로 더 높게 나타났다. 그리고 [+grave] 자질을 갖는 폐쇄음이 선행할 경우 F1 값은 비어두 /ㅡ/에서 유의미한 차이가 나타났으며, F2 값은 어두 /ㅡ/와 /ㅜ/를 제외한 모든 집단에서 선행자음의 [grave] 자질의 유무에 따른 유의미한 차이가 나타났다.

선행 자음의 특성과 후설모음의 포먼트의 관계에 대한 논의 결과는 다음과 같다. 첫째, 선행 자음이 치조나 경구개, 구개 등에서 조음될 경우 F2 값의 상승을 유발하는데 구개에 비하여 치조에서 더 높은 F2 값이 나타난다. 둘째, 선행 자음의 소음 구간이 F2 값의 상승을 유발하는 요인으로 추정되는데, 조찰성(strident)이 강한 마찰음이 강한 소음 에너지를 보유함으로써 더 높은 F2 값을 만들어내기 때문이라고 보았다. 셋째, 상술한 선행 자음의 특성은 전설모음보다는 후설모음에서, 저모음보다는 고모음에서 효과적으로 작용하여 더 높은 F2 값이 나타난다. 넷째, 유음과 비음 모두 저주파수 에너지를 지니는 특성을 가진 공명음(sonorant)으로 유음과 비음은 후행 모음이 낮은 F1 값을 갖도록 영향을 미치는 반면, 유음과 비음의 높은 F2 값은 후행 모음이 높은 F2 값을 갖도록 영향을 미친다. 다섯째, [grave] 자질을 가진 폐쇄음 /ㄱ/는 음의 개방 단계에서 소음성 특징을 나타낸다. 상술한 바와 같이 선행 자음의 소음성은 후행모음의 F2 값이 상승하는 역할을 하게 되는데, /ㄱ/가 어두에 올 경우 소음성을 갖게 됨으로 인해 후행하는 모음의 F2 값 상승에 영향을 미치는 반면, 비어두에서는 소음 구간이 약화되거나 사라지기 때문에 F2 값 상승에 영향이 미치지 못하는 것으로 보았다.

한국어 단모음의 RNR 평균 비교

윤 지 현[1], Jiayin Gao[2], Takayuki Arai[3], 성 철 재[1]
1 충남대학교 언어학과
2 Linguistics and English Language, University of Edinburgh
3 Dept. of Information and Communication Sciences, Sophia University

Comparison of Rhamonics to Noise Ratio of Korean Vowels

Jihyeon Yun[1], Jiayin Gao[2], Takayuki Arai[3], Cheoljae Seong[1*]
1 Dept. of Linguistics, Chungnam National University
2 Linguistics and English Language, University of Edinburgh
3 Dept. of Information and Communication Sciences, Sophia University
* cjseong49@gmail.com

한국어 모음에서 측정된 음질 특성은 음성장애 분야를 중심으로 많은 연구가 이루어지고 있다. 이러한 병리 관련 연구에서는 분석 대상 모음이 주로 /ㅏ/ 또는 /ㅏ, ㅣ, ㅜ/로 제한된다. 반면, 서윤정.신지영(2018)에서는 '한국인 표준 음성 데이터베이스'에 수록된 수도권 출신 정상 화자 309명의 한국어 8개 단모음 연장 발화 음성에 대해 발성유형 관련 음향 매개변수 f0, jitter, shimmer 및 소음 대 조화음 비율(Noise to harmonics ratio [NHR])을 측정하고, 화자의 성별과 연령대 및 모음에 따라 나누어 결과를 보고한 바 있다. 한편 김난숙.성철재(2017)에서는 과소비성을 유발할 수 있는 아데노이드 편도 비대 또는 알레르기성 비염 아동 집단의 음성을 분석하였는데, 두 집단을 구분하는 음향 변수로서 켑스트럼의 조화음(라모닉)과 소음의 비율에 해당하는 Rhamonics to noise ratio [RNR]가 스펙트럼 변수인 조화음 대 소음 비율(Harmonics to noise ratio [HNR]) 및 NHR보다 유용한 것으로 나타났다.

본 연구에서는 한국어 단모음 범주별로 RNR 측정값 평균을 구하여 모음에 따른 차이가 있는지 알아보고, 이 결과를 다른 발성유형 관련 음향 매개변수와 비교하였다. 분석 대상으로는 일본 도쿄 소재 대학에 유학 중인 19-29세 수도권 출신 한국어 모어 화자 9명(여5, 남4)이 단음절을 /CV-다/ 및 /아-CV-다/ 조건으로 발화한 음성자료 중 한국어 7개 단모음 /ㅣ, ㅐ, ㅏ, ㅓ, ㅗ, ㅜ, ㅡ/를 사용하였다. 음절 초성 자음은 /ㅁ, ㅂ, ㅃ, ㅍ, ㄴ, ㄷ, ㄸ, ㅌ, ㄹ, ㅅ, ㅆ/ 중 하나에 해당한다. 분절한 각 모음 표본에 대해 Praat 6.1.13에서 일괄적으로 중간 26 ms 전후에 묵음구간 각 15 ms를 연결하여 추출하고 PowerCepstrum 객체를 생성해 RNR을 측정하였다. 화자(9)/음절위치(2)/자음(11)/모음(7) 조건별 교차표 각 셀에서 RNR 평균값 1386개를 구하고, 남녀별 기술통계량을 계산한 결과는 <표 1>과 같다.

	n	mean	sd	median	mad	min	max
F	770	0.739	0.450	0.649	0.428	0.060	3.112
M	616	0.506	0.292	0.420	0.224	0.060	1.778

표 1 RNR 조건별 평균치 데이터셋의 기술통계량

R 패키지 lmerTest의 lmer() 함수에 식 [rnr ~ vowel * onset * position + sex + (1 | speaker)]을 입력하여 4676개 관측치를 바탕으로 선형혼합모형의 효과량을 추정하였다. 응답변수는 RNR 측정치이고, 고정효과는 순서대로 모음, 자음, 음절위치와 삼원교호작용 및 성별이다. 무선효과로는 화자별 절편을 추가하였다. car::Anova() 함수를 이용한 고정효과의 제2종 Wald 카이자승 검정 결과는 <표 2>와 같다. 이

에 따르면 모음, 자음 변인의 주효과 및 자음:음절위치 교호작용이 유의수준 0.0001에서 통계적으로 유의하다.

	Chisq	df	Pr(>Chisq)
vowel	266.633	6	< 2.2e−16***
onset	233.500	10	< 2.2e−16***
position	0.080	1	0.7776
sex	2.183	1	0.1396
vowel:onset	54.219	60	0.6859
vowel:position	6.673	6	0.3521
onset:position	93.387	10	1.138e−15***
vowel:onset:position	63.570	60	0.3519

표 2 RNR에 대한 선형혼합모형의 고정효과 검정 결과 (***p < 0.001)

R emmeans 패키지의 emmeans() 함수를 이용하여 위 모형으로부터 추정한 RNR 주변평균을 음절위치 및 선행자음에 따라 세분한 각 21개 모음 쌍의 평균차를 Bonferroni 교정으로 검정하였다. 총 462쌍 중 유의수준 0.05에서 31쌍, 유의수준 0.01에서 16쌍의 RNR 평균차가 유의했다. 한편 성별, 음절위치, 자음 변인에서 각각 효과량의 평균값을 취한다는 가정하에 계산한 모음별 추정 주변평균은 <표 3>과 같다. 21개 모음 쌍의 추정 주변평균 차이에 대해 Bonferroni 교정하여 검정한 결과(Kenward-Roger df = 4514), ‘i-e’ (p = 0.0474)와 ‘i-ʌ’, ‘i-ɨ’, ‘e-a’, ‘e-ɨ’, ‘ʌ-ɨ’, ‘o-u’ (ps > 0.1)의 7개 쌍을 제외한 나머지 14쌍의 모음 간 비교에서 RNR 평균차가 유의했다(p < 0.001).

vowel	emmean	SE	df	lower.CL	upper.CL
a	0.793	0.082	7.62	0.602	0.984
e	0.726	0.082	7.64	0.535	0.917
ɨ	0.669	0.082	7.65	0.478	0.860
i	0.646	0.082	7.64	0.456	0.837
ʌ	0.607	0.082	7.64	0.416	0.797
u	0.492	0.082	7.65	0.301	0.683
o	0.449	0.082	7.65	0.258	0.640

표 3 RNR에 대한 선형혼합모형의 모음별 추정 주변평균 (내림차순 정렬)

동일한 음성자료에 대해 Praat 6.1.13으로 측정한 HNR, NHR 및 Cepstral peak prominence [CPP] 값을 각각 응답변수로 하여 위와 동일한 예측변인을 투입한 통계모형 식을 적용하고, 모형별로 각 462쌍의 모음 간 추정 주변평균 차이를 검정하였을 때 HNR > CPP > RNR > NHR 순으로 차이가 유의한 모음 쌍이 많았다. 어절위치(2)와 자음 조건(11)별 모음 비교 각 21쌍에서, Bonferroni 교정한 유의수준 0.05를 기준으로 차이가 유의한 모음 쌍은 HNR에서 총 175쌍, CPP에서 77쌍, RNR에서 31쌍, NHR에서 10쌍이었다.

참고문헌

김난숙, 성철재. (2017). 과소비성 (Hyponasality) 집단의 음향 특성과 분류 변인. 언어학, (78), 31-61.
서윤정, 신지영. (2018). 모음 연장 발성이 보이는 연령대별 음향음성학적 특성 연구. 말소리와 음성과학, 10(4), 67-76.

L2 영어 말하기 유형별 (즉흥자유체 vs. 낭독체) 발화 평가 채점 세부 요인이 말하기 평가 총체적 점수에 미치는 영향에 대한 비교

이 석 재
연세대학교 영어영문학과

Comparison of evaluation factors on the scores of L2 English productions in spontaneous and read speech

Seok-Chae Rhee
Dept. of English Language and Literature, Yonsei University
scrhee@yonsei.ac.kr

본 발표는 한국인 영어학습자들의 “즉흥 자유발화(대화형 & 독백형)”와 “낭독발화”에 대한 각 평가 채점 요인이 총체적 말하기 평가 점수에 미치는 영향성 정도를 대조 비교하는 데 있다.

유창성 평가의 대상이 된 음성코퍼스는 YELC 2011 Speech Corpus이고, 이에 대해 4인의 평가자가 평가루브릭에 따라 2018년 말하기 평가를 수행하였다. 즉흥 자유발화에 있어 평가 채점 각 요인은 Segment, Prosody, Word Choice, Grammar이고, 낭독발화에서는 Word Choice와 Grammar가 제외되고 (보고 읽는 낭독발화이기 때문) 발화 운율 유창성 부분을 세부 영역으로 나누어 Segment, Stress-Rhythm, Speed-Pause, Intonation으로 구분되었다.

평가 결과에 대한 다중회귀분석 결과는 아래와 같다.

Ia). 즉흥자유체 발화 (대화형)

모형		비표준화계수 B	비표준화계수 표준화오차	표준화계수 베타	t	유의확률	공선성통계량 공차	공선성통계량 VIF
1	(상수)	−.317	.041		−7.795	.000		
	Segment	.268	.021	.273	12.864	.000	.230	4.352
	Prosody	.215	.022	.212	9.901	.000	.226	4.426
	Grammar	.161	.015	.163	10.584	.000	.437	2.289
	Word	.374	.014	.424	26.542	.000	.405	2.470
R=.96, R^2=.921, Adjusted−R^2=.920, F=2228.590, p<.000								

Ib). 즉흥자유체 발화 (독백형)

모형		비표준화계수 B	비표준화계수 표준화오차	표준화계수 베타	t	유의확률	공선성통계량 공차	공선성통계량 VIF
1	(상수)	−.386	.028		−14.035	.000		
	Segment	.192	.014	.175	13.980	.000	.296	3.377
	Prosody	.344	.016	.336	21.397	.000	.188	5.307
	Grammar	.149	.016	.141	9.169	.000	.196	5.106
	Word	.366	.015	.378	23.806	.000	.185	5.411
R=.962, R^2=.926, Adjusted−R^2=.926, F=4975.107, p<.000								

II). 낭독체 발화

모형		비표준화계수 B	비표준화계수 표준화오차	표준화계수 베타	t	유의확률	공선성통계량 공차	공선성통계량 VIF
1	(상수)	−.161	.041		−3.928	.000		
	Stress−Rhythm	.140	.031	.146	4.490	.000	.096	10.440
	Segment	.443	.019	.452	23.888	.000	.281	3.556
	Rate−Pause	.264	.021	.252	12.402	.000	.243	4.114
	Intonation	.172	.028	.176	6.070	.000	.120	8.322
R=.959, R^2=.920, Adjusted−R^2=.919, F=2292.473, p<.000								

결론: 텍스트 없이 자신의 생각을 직접 영어로 말해야하는 즉흥자유체 발화에서는 말소리 자체 측면보다는 대화와 독백에서 공히 ‘영어 단어의 선택’이 적정했는지 여부가 말하기 평가에 가장 큰 영향을 끼쳤

으며, ‘문법 준수’ 여부는 영향이 가장 적었고, 대화형 발화에서는(Ia) 분절음들을 옳게 발화했는지 여부가 운율적인 특징을 제대로 발화했는지 여부보다 총체적 말하기 평가 결과에 미치는 영향이 더 컸으며, 독백형에서는(Ib) 이와 반대로 운율적 특징을 제대로 준수하며 발화했는지 여부가 분절음들을 제대로 발화했는지보다 더 큰 영향력을 보이는 것으로 나타났다. 한편 스크립트를 보고 읽는 낭독체 발화에서는 분절음들의 정확한 발음 여부가 다른 운율적 각 요인들보다 총체적인 말하기 유창성 평가 점수에 더 큰 영향을 끼치는 것으로 나타났다.

한국어 방언 자동 식별을 위한 방언 특징 분석

이주영[1], 김경화[2], 양승희[3], 정민화[1]
서울대학교 언어학과[1], 대검찰청[2], 서울대학교 인문정보연구소[3]

An Analysis on Korean Dialectal Features for Dialect Identification

Jooyoung Lee[1], Kyungwha Kim[2], Seunghee Yang[3], Minhwa Chung[1]

Dept. of Linguistics, Seoul National University[1]
Supreme Prosecutor's Office, Korea[2]
Center for Humanities and Information, Seoul National University[3]

excalibur12@snu.ac.kr[1], savoix@spo.go.kr[2]
sy2358@snu.ac.kr[3], mchung@snu.ac.kr[1]

방언은 한 언어의 음운, 문법, 어휘적 틀 안에서 표준어 혹은 다른 방언과의 차이를 드러낸다. 방언학은 지역 방언의 특징을 이러한 층위에서 분석하여 기술하는 분야이며, 방언을 구분한다는 것은 방언학에서 다룬 내용을 토대로 방언 간 차이를 분석한다는 의미가 된다. 방언 자동 식별은 인간이 방언학적 지식을 통해 방언 구분을 하는 것을 기계가 대량의 데이터와 적절한 특징을 사용하여 대신하려는 시도라고 할 수 있다. 본 연구에서는 한국어 방언 자동 식별에 적합한 방언 표지 선정을 위해 방언학의 기존 연구들을 종합적으로 분석하여 음운, 문법, 어휘, 억양 특징 중 계산 모델링에 적용할 수 있는 방언 표지 후보를 제시하고자 한다. 음운, 문법, 어휘를 분석 범위로 선정한 것은 김공칠 (1969) [1]을 근거로 하였다.

방언학의 기존 연구에서 분절음은 음운, 문법, 어휘적 측면에서, 초분절음은 억양 측면에서 다루어졌다. 음운 관련 연구는 음운변동과 모음 위주의 기술이 주였는데, 단모음화, 고음화, 전설모음화 등의 모음 관련 음운 현상과 경음화, 구개음화 등의 자음 관련 음운 현상을 들 수 있다. 모음은 곽충구 (2003) [2]을 참고해 볼 때, 이중모음 /ㅟ/와 /ㅚ/의 단모음화 여부, /ㅐ/와 /ㅔ/의 합류 여부를 비롯하여 경상 방언의 /ㅡ/와 /ㅓ/ 합류, 제주 방언의 /ㆍ/나 /ㆍㅣ/와 같은 독자적인 모음의 여부에 따라 방언별 모음 체계를 정리하여 연구하였다.

문법 관련으로는 종결어미의 방언형을 중심으로 기술되었다. 장승익 (2019) [3]에서는 전북 방언 특유의 조사로 문말조사 '-요', '-라우'와 어미조사 '-잖여', '-느만', '-드만'을 언급하였다. 어휘는 방언 지도를 제시하여 근접 지역 내의 유사한 어휘 분포와 먼 지역 간의 상이한 어휘 분포를 동시에 보였다. 또한, 억양은 성조 개념으로 두드러지게 언급되었는데, 김차균 (1970) [4]에서는 경남 창원 방언의 성조를 어절과 문장(월) 단위에서 H(높은 소리)와 L(낮은 소리)로 표현하였다. 이외에 장승익 (2012) [5]에서는 전북 방언의 핵억양 패턴을 정리하였다.

이와 같이 한국어 방언학은 분절음에서는 모음 및 자음의 음운변동, 개별모음의 음성적 실현, 종결어미, 방언 지도상의 어휘 분포, H와 L로 표현한 억양으로 정리할 수 있다. 각각의 정리된 연구는 방언학 전문가들의 판단으로 기술하였다는 점에서 수치적으로 표현이 가능한지를 확인할 필요가 있다. 결국, 방언 자동 식별은 방언학에서 기술한 방언 표지를 객관적으로 표현할 수 있을 때 비로소 가능해지기 때문이다. 언어 현상을 객관화하는 연구는 실험음성학에서 살펴볼 수 있다. 음운과 관련하여 장혜진 (2010) [6], 한경임 (2014) [7]에서는 각각 서울과 대구, 제주와 대구 방언을 VOT(Voice Onset Time)과 F0를 활용하여 방언 간의 어두 폐쇄음 차이를 알아보았다.

또한, 이중모음 /ㅢ/와 관련하여 박선우 (2019) [8]는 이중모음 /ㅢ/ 발화 내의 전반부와 후반부로 나

누어 구간의 포먼트 F1과 F2의 변화를 살펴보는 실험을 하였다. 억양으로는 오재혁 (2014) [9]에서 한국어 억양 곡선의 정규화 방안을 제시하여 억양 표시가 그동안 객관적이지 못한 점을 지적하여 억양 특징을 수치적으로 접근하려는 시도를 보였다. 실험음성학은 음운과 억양 관련 연구는 활발하나 문법과 어휘는 음향적 특징과 거리가 있어 연구가 눈에 띄지는 않았다.

위 실험음성학적 성과를 참고할 때 방언 자동 식별은 음운과 억양을 중심으로 이루어질 것으로 보인다. 실험음성학에서 F0, 포먼트, VOT 등을 표지로 사용한 점을 고려하면, 방언 자동 식별에 적용할 객관적인 수치는 프레임(Frame) 단위에서 시작해야 한다. 25ms의 프레임 구간이 10ms씩 발화 구간 내에서 이동하면서 각 프레임의 음향 특징을 추출하는 방식으로 접근할 때, 방언학에서 다룬 방언의 모음과 자음은 각각 F0, 포먼트 F1, F2와 VOT를 계산하여 추출할 수 있다.

음운변동, 문법, 어휘 관련 특징은 음성인식의 결과를 해석하는 방법으로 각 방언의 고유 특징을 알아볼 수 있다. 특히, 음운변동은 음운과 관련 있으나 변동 과정을 객관적으로 나타낼 수 없기 때문에, 변동의 결과를 방언의 어휘로 간주하여 어휘 자체를 인식하는 형태로 접근해야 한다. 이때 음성인식 모델은 어휘 사전에 방언별 어휘를 학습 과정에 추가하는 것이 중요하다.

초분절음적 요소인 억양은 오재혁 (2014) [9]의 방법론을 적용할 수도 있으나, 최근 다양한 분야에서 높은 식별 성능을 보이는 딥러닝 기법과 음성공학적 특징을 활용하여, 모델 스스로 방언별 특징을 학습하는 방법을 생각해 볼 수 있다. 전통적으로 음향음성학에서 활용되어 온 F0나 포먼트, 스펙트로그램 뿐만 아니라 더 나아가 음성인식에서 사용하는 MFCC (Mel Frequency Cepstral Coefficients) 등의 low level 특징도 딥러닝 모델의 입력 표지로 사용할 수 있다.

감사의 글

본 연구는 대검찰청 용역연구개발연구사업 과제 “AI를 이용한 음성 자동 분석 기법 개발”의 연구지원비 지원에 의해 수행되었습니다.

참고문헌

[1] 김공칠. (1969). 방언의 연구. 한글, (143), 142-164.
[2] 곽충구. (2003). 현대국어의 모음체계와 그 변화의 방향. 국어학 (國語學), 41, 59-92.
[3] 장승익. (2019).《 전라북도 방언사전》의 표기에 대한 연구. 국어문학, 71, 97-120.
[4] 김차균. (1970). 경남 방언의 성조 연구. 한글, (145), 111-152.
[5] 장승익. (2012). 전북방언의 핵억양 특징 연구. 한국언어문학, 83, 91-116.
[6] 장혜진, 신지영. (2010). 어두 폐쇄음의 발성 유형 지각에서 나타나는 방언 간 차이: 서울 방언과 대구 방언의 비교를 바탕으로. 한국어학, 49, 369-388.
[7] 한경임. (2014). 한국어 폐쇄음 VOT: 제주방언과 대구방언 비교. 코기토, (75), 157-176.
[8] 박선우. (2019). 한국어 이중모음/의/의 음향적 특성과 유형에 대하여. 현대문법연구, 102, 165-183.
[9] 오재혁. (2014). 한국어 억양 곡선의 정규화 방안에 대한 연구. 한국어학, 62, 395-420.

‘동의’의 정도에 따른 운율 실현에 관한 연구

양봉석, 신지영
고려대학교 국어국문학과

A study on the prosodic realization according to the degree of agreement.

Bongseok Yang, Jiyoung Shin
Department of Korean Language and Literature, Korea University
wdg3938@korea.ac.kr, shinjy@korea.ac.kr

본고는 화자가 상대방의 말에 대해 동의하는 정도에 따른 운율적 실현 양상을 살핌으로써 화자의 동의 정도와 운율 간의 상관관계를 밝히는 것을 목적으로 한다. 구어에서는 화자가 자신의 태도에 따라 운율을 달리 실현하는 경우가 많다. 화자는 상대방의 발화에 대해 얼마나 동의하는가도 운율을 통하여 표현할 수 있다. 대화상에서 상대방의 주장이나 제안 등에 대해 동의를 드러내는 것이 더 공손하며 선호되는 반응이므로, 동의를 잘 드러낼 수 있는 운율을 실현하는 것은 중요하다. 따라서 이에 주목함으로써 원만한 대화를 위한 운율 전략의 한 방법에 대해 탐구해 볼 수 있을 것이다.

이를 위하여 산출실험을 진행하고자 한다. 산출실험에서는 피험자가 상대방의 발화에 대해 얼마나 동의하는가를 드러내되, 오로지 “어 좋아.”라는 표현만을 사용하여 응대해야 하는 과제를 수행하도록 하였다. 이를 통해 화자가 선행발화에 대해 얼마나 동의하는가를 오로지 운율만을 활용하여 드러내도록 한 것이다. 이후 발화의 음높이 범위(pitch span), 강도(intensity), 말속도(speech rate) 및 선행발화와 응대발화 간 간격이 동의의 정도와 상관관계를 갖는지를 분석하고자 한다.

선행연구에 따르면 발화의 음높이 범위가 넓을수록 선행화자의 발화에 대한 관여(involvement)가 높다는 것을 드러낸다[1]. 또한 동의 표현의 경우 말속도가 빠를수록, 강도가 강할수록 더 강한 동의로 지각된다[2]. 그리고 선행발화와 응대발화 간의 간격(gap)은 동의를 표현할 때 대체로 짧게 나타나며, 비동의를 표현할 때는 대체로 길게 나타난다[3]. 이에 따라 본고의 분석 결과는 화자가 선행발화에 대해 동의할수록 발화의 음높이 범위를 넓게, 강도를 강하게, 말속도를 빠르게, 선행발화와 응대발화 간 간격을 짧게 실현할 것으로 예상하였다.

[1] Cruttenden, A. (1997). *Intonation*. Cambridge University Press.
[2] Freeman, V. (2015). The phonetics of stance-taking (Doctoral dissertation).
[3] Pomerantz, A. (1984). Agreeing and disagreeing with assessments: some features of preferred/dipreferred turn shapes. In: Atkinson, J. Maxwell, Heritage, John (Eds.), Structures of Social Action. Studies in conversational analysis. Cambridge University Press, Cambridge, pp. 57-101.

구두 발표 IV
음성공학

좌장: 박정식(한국외대)

GAN 기반 신경망 보코더의 비교평가 및 음질 개선방안 연구

서영주, 최연주, 엄지섭, 정성희, 김회린
한국과학기술원 전기및전자공학부

A study on comparative evaluation of GAN-based neural vocoders and their speech quality improvement

Youngjoo Suh, Yeunju Choi, Jisub Um, Sunghee Jung, and Hoirin Kim
School of Electrical Engineering, KAIST
yjsuh@kaist.ac.kr, wkadldppdy@kaist.ac.kr, twiz0311@kaist.ac.kr, hash2430@gmail.com, hoirkim@kaist.ac.kr

딥러닝 기술의 도입은 음성합성 분야에도 많은 기술적 발전을 가져왔다. 개발된 음성합성 기술은 합성음의 음질 측면에서 인간이 발성한 실제 음성 수준에 근접할 정도로 발전하였다. 초기에는 적용된 기술적 방식으로 인하여 실시간 합성에는 이르지 못했지만 최근에는 기술적인 발전으로 인하여 이마저도 극복하였다. 음성합성 기술에서 이 두 가지 성능을 만족시키기 위해서 음성합성기 전체를 구성하고 있는 두 모듈 즉, 텍스트 정보를 스펙트로그램 (spectrogram) 정보로 변환시키는 음성합성기 본 모듈과 스펙트로그램 정보를 음성파형 정보로 변환시키는 보코더 (vocoder) 모듈에서 동시에 기술적 발전이 이루어져 왔다. 이 중에서 보코더는 최종적으로 인간이 듣고 인지하는 대상이 되는 음성파형을 합성한다는 의미에서 음성합성기 본 모듈에 못지않은 기술적 중요성을 가진다. 고음질과 실시간 합성의 두 가지 특성을 지닌, 최근에 개발되는 이러한 신경망 보코더들은 많은 경우에서 generative adversarial network (GAN) 알고리즘에 기반을 두고 있다. 이러한 배경에서, 본 논문에서는 GAN 기반의 대표적인 신경망 보코더들에 대한 비교평가와 이들의 음질을 개선시키기 위한 기본적인 방안에 대한 연구결과를 다루었다. 실험에서는 MelGAN, Multi-band MelGAN, Parallel WaveGAN, 세 가지 신경망 보코더들을 사용하였다. MelGAN은 다른 두 보코더에 비해 음질에서 가장 만족스럽지 못하고 Multi-band MelGAN은 가장 고속이면서 음질적으로는 중간 정도이며 Parallel WaveGAN은 음질은 양호하지만 돌발적인 잡음을 발생시키는 특징이 있다. 알고리즘의 구조적 개선 없이, MelGAN에서는 generator의 residual stack을 구성하는 layers의 수와 Parallel WaveGAN에서는 generator의 전체 layers의 수를 각각 증가시키는 방안이 음질 상에서 의미 있는 개선을 가져옴을 확인하였다.

<감사의 글>
본 연구는 산업통상자원부의 산업기술혁신사업으로부터 지원을 받아 수행된 연구입니다 (No. 10080667, 음원 다양화를 통하여 로봇의 감정 및 개성을 표현할 수 있는 대화음성합성 원천기술 개발).

합성곱 신경망을 이용한 화자 검증에서 시계열 정보의 활용도 분석

허정우, 심혜진, 정지원, 김주호, 유하진
서울시립대학교 컴퓨터과학부

An analysis of the Utilization of Time Series Information in Speaker Verification Using Convolutional Neural Network

Jungwoo Heo, Hye-jin Shim, Jee-weon Jung, Ju-ho Kim, Ha-Jin Yu
School of Computer Science, University of Seoul
jungwoo4021@gmail.com, shimhz6.6@gmail.com, jeewon.leo.jung@gmail.com, wngh1187@naver.com, hjyu@uos.ac.kr

최근 화자 검증 연구에서는 화자 정보 추출을 위해 주로 심층 신경망을 사용한다. 음성은 시계열 데이터지만 화자 검증에서는 시계열 데이터 처리에 좋은 성능을 보이는 것으로 알려진 순환 신경망 대신 합성곱 신경망을 사용하는 경우가 많다. 본 연구에서는 합성곱 신경망을 사용하여 화자 정보를 추출할 때 시계열 정보가 사용되는 정도를 확인하는 연구를 진행하였다. 본 연구는 신경망 내부에서 시계열 정보를 활용하고 있다면 시계열 정보를 삭제했을 때 성능이 저하될 것이라 가정하였다. 따라서 시계열 정보의 활용도 확인을 위해 특징 지도의 시계열 정보를 시간 축을 기준으로 임의로 섞는 혼합 계층을 사용하여 제거한 뒤, 이에 따른 성능 변화를 확인하였다. 혼합 계층은 합성곱 신경망의 여러 위치에 삽입한 뒤, 각 경우의 성능을 비교하였다. 이는 합성곱 연산을 수행한 정도에 따라 시계열 정보가 얼마나 활용되는지를 확인하기 위함이다. 신경망의 성능은 등록 발성과 입력 발성에서 추출한 화자 모델 간의 코사인 유사도를 바탕으로 동일 오류율을 계산하여 확인하였다. 연구에 사용된 베이스라인은 ResNet18[1]이며, VoxCeleb1 데이터셋을 사용해 학습과 평가를 진행하였다. 입력 특징은 발성 단위 평균과 분산 정규화한 Mel-Filterbank 에너지 특징을 사용하였다.

실험 결과는 표 1과 같다. 신경망에 입력하기 전 시계열 정보를 제거한 경우와 첫 합성곱 이후 시계열 정보를 제거한 경우에 동일 오류율이 각각 11.21%, 9.36%로 합성곱 연산을 거친 이후 시계열 정보의 활용도가 낮아짐을 확인하였다. 시계열 정보는 합성곱 연산을 수행한 정도에 비례하여 활용도가 낮아져, 잔차 블록2 이후 제거한 경우 베이스라인과 성능이 유사하였다. 이처럼 합성곱 신경망을 이용해 화자 정보를 추출하는 과정에서 시계열 정보의 활용도가 낮아지는 것은 기존 연구에서 순환 신경망 대신 합성곱 신경망을 사용한 화자 검증에서 좋은 성능을 보여줄 수 있었던 원인으로 분석된다.

혼합 계층 위치	동일 오류율(%)
베이스라인	7.48
입력층 이전	11.21
첫 합성곱 이후	9.36
잔차 블록1 이후	7.97
잔차 블록2 이후	7.41
잔차 블록3 이후	7.29

<표 1. 혼합 계층의 위치와 동일 오류율>

감사의 글

이 논문은 2020년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초 연구사업임(2020R1A2C1007081)

참고문헌

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Including Linguistic Knowledge in an Auxiliary Classifier CycleGAN for Corrective Feedback Generation

Seung Hee Yang*, Minhwa Chung**
Senior Researcher, Institute of Humanities, Seoul National University*
Department of Linguistics, Seoul National University**
sy2358@snu.ac.kr, mchung@snu.ac.kr

Learners who acquire a second language (L2) usually speak with a non-native accent because of the influence of their mother tongues. Accent conversion technique, which transforms non-native to native accented speech, enables L2 learners to listen to native-accented speech as a corrective feedback in a computer-assisted pronunciation training setting. However, since it is difficult for the L2 learners to evaluate their own pronunciations, it would be beneficial to inform what type of variations occurred, such as coda deletions, three-way distinctions, and prosodic variations.

Motivated by the analysis results in a related work [1], which found that both segmental and suprasegmental criteria are positively correlated with human proficiency ratings in L2 Korean, this paper proposes an augmented Cycle-consistent Generative Adversarial Network (CycleGAN) for accent conversion-based feedback generation. In this network, the generator learns to convert non-native speech to the native counterpart by adversarial training [2],[3]. An additional auxiliary classifier, a two-layer convolutional neural network, is trained on the non-native speech mel-spectrograms and performs three-class classification: “segmental correction,” “suprasegmental correction,” and “no correction.”

For implementation, the proposed model is trained on L2KSC (L2 as Korean Speech Corpus) [4], consisting of 300 short utterances produced by 217 non-native speakers with 27 mother tongue backgrounds, and 107 native speakers. For the auxiliary classifier training, non-native speech samples were manually annotated with the three classes. Average F1 score achieves 0.55, with the highest accuracy in segmental error classification, and confusion patterns observed between “no error” and “suprasegmental correction” classes.

Compared to the conventional accent conversion approaches, the proposed method allows integration of linguistic knowledge in a neural network, so that the feedback also informs the learners of the linguistic variation types and the users are expected to benefit from the knowledgeable feedback. Future experiments can be conducted with more fine-grained linguistic classes.

Keywords: Auxiliary Classifier CycleGAN, Corrective Feedback Generation using Generative Adversarial Network, Using Linguistic Knowledge in a Cycle-consistent Adversarial Training

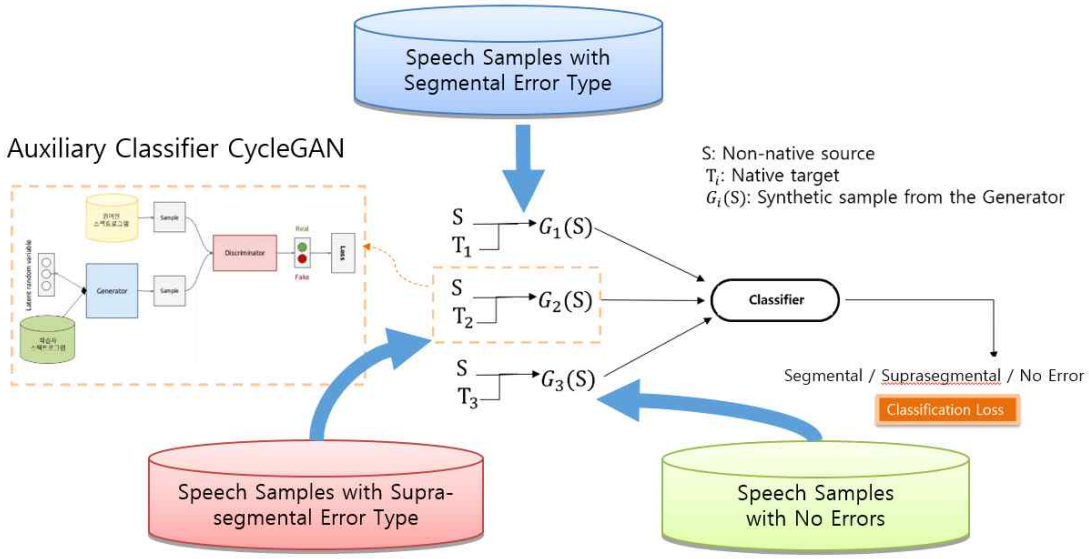


Figure 1. Proposed Auxiliary Classifier CycleGAN consists of three CycleGANs, each corresponding to a linguistic class, and a domain classifier. For each linguistic class, there is a CycleGAN with two discriminators and two generators, which is consistent with the existing CycleGAN approach. The synthetic sample Gi(S) is generated from the source (S). Cycle-consistency loss is calculated according to the difference between real samples (i.e. native speech) and their corresponding reconstructed samples. The domain classifier learns to ensure the discriminability between the generated samples.

Acknowledgment

This research is supported by Ministry of Culture, Sports and Tourism(MCST) and Korea Creative Content Agency(KOCCA) in the Culture Technology(CT) Research & Development Program 2020.

참고문헌

[1] S.H. Yang and M. Chung, Linguistic Factors Affecting Evaluation of L2 Korean Speech Proficiency. In Proceedings of Speech and Language Technology in Education (SLaTE) 2017, Stockholm, Sweden, pp. 53-58, 2017.

[2] P. Isola, J.Y. Zhu, T. Zhou, A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of CVPR, 2017.

[3] S.H. Yang and M. Chung. Self-imitating Feedback Generation Using GAN for Computer-Assisted Pronunciation Training. In Proceedings of Interspeech 2019, Graz, Austria, pp. 1881-1885. 2019.

[4] S. Lee, and J. Chang. Design and Construction of Speech Corpus for Korean as a Foreign Language (L2KSC). In The Journal of Chinese Language and Literature, vol. 33, pp. 35-53. 2005.

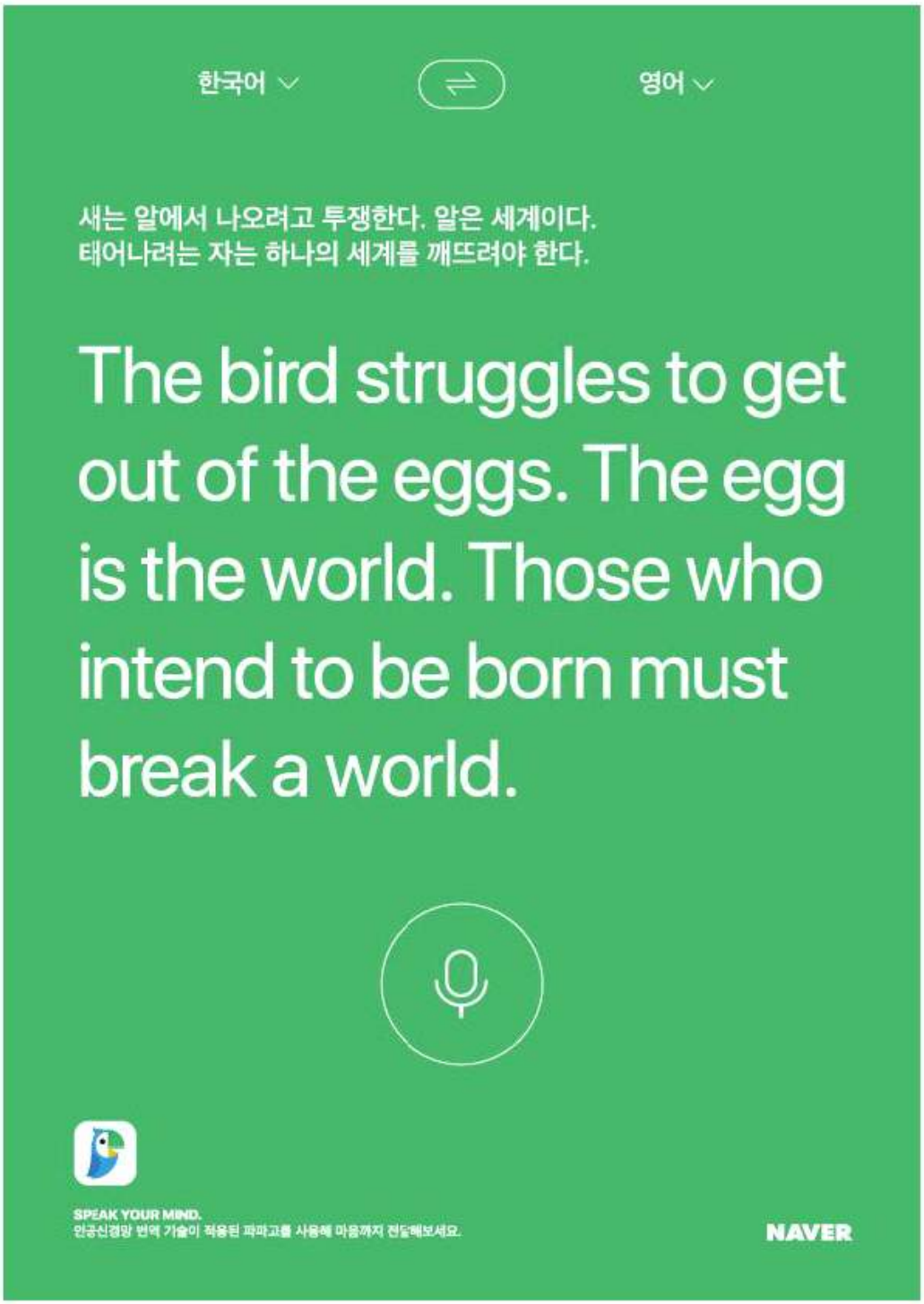
연속음성에서의 Jitter 측정방법

조 철 우
창원대학교 전기전자제어공학부

Jitter Measurement from Connected Speech

Cheolwoo Jo
School of Electrical, Electronics and Control Eng., Changwon National University
cwjo@changwon.ac.kr

본 연구에서는 연속음성에서의 Jitter성분 측정 방법에 대해 고찰하고자 한다. 기존의 Jitter측정방법으로는 지속 발성한 모음을 대상으로 변동성을 측정하는 방법을 주로 사용하여왔다. 문장음성 등 연속음성의 경우는 문장에 따른 운율정보의 영향으로 기존의 측정법으로는 왜곡이 발생하게 된다. 이에 연속 발성에 대해 운율정보의 피치 변동을 상쇄시키는 방법을 제안하고자 한다. 피치 변동을 제거하는 방법으로는 분석구간내에서의 피치 변동을 다항식 보간법에 의해 변동 경향을 대표하는 곡선을 구하고 그 곡선을 기준으로 변이를 제거한다. 이후 변이가 제거된 피치의 궤적으로부터 Jitter를 측정하는 방법을 적용하여 피치 주파수의 변동성을 측정하였다. 제안한 방법의 효용성 측정을 위해 Kay Pentax MEEI DB의 정상음성 표본을 사용하였다.


한국어
영어
새는 알에서 나오려고 투쟁한다. 알은 세계이다.
태어나려는 자는 하나의 세계를 깨뜨려야 한다.
The bird struggles to get out of the eggs. The egg is the world. Those who intend to be born must break a world.
SPEAK YOUR MIND.
인공신경망 번역 기술이 적용된 파파고를 사용해 마음까지 전달해보세요.
NAVER

포뮬러셀레이션구상 이벤라사

bio.qlex.www\\:qttn 9212-061 (20 로뮬

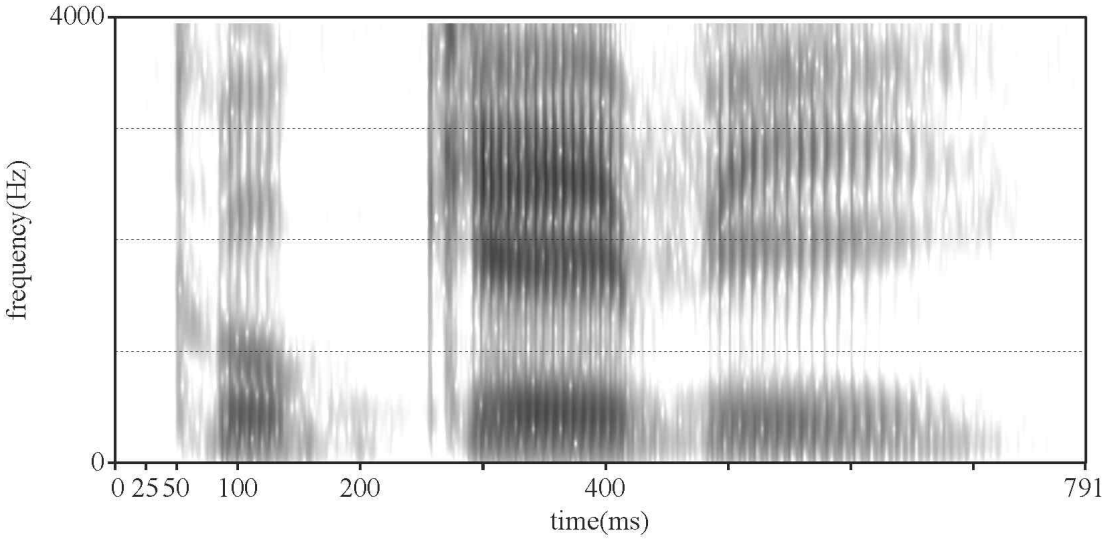
.라니뮬크하초 를표테 로ㅣ표대술샹 을러 표샹위음구상 뒤0202ㄱ

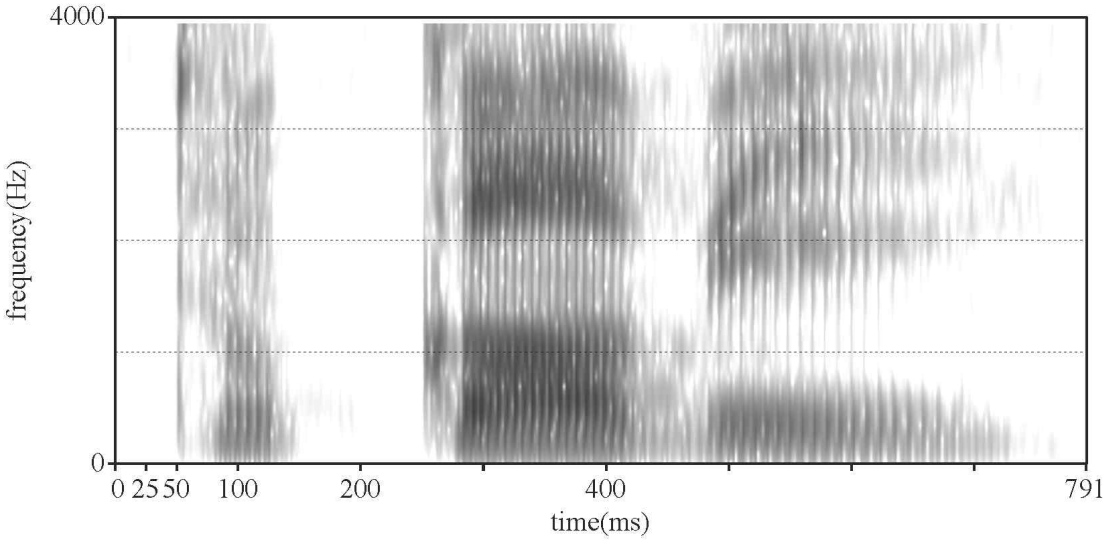
샹위음구상' 기오 배쁜스 배하 을뮈쌓 로샹위음 구상 스표벊라셀레이션구상 (ㅋ) 션구쩡 쁨됨이 로위음 .라니뮬크하초 트이밉기 를표테 로표대술샹 을러 로'표 음 오 배끌러 을뮈쌓 배훙 를구쩡 오씹 뜨뜡꾠 션용응 쁨뻐닢 로커베어샹쨩 위이 로샹크쌩이잉 션샹위음 로구상 배훙 을쁜쨩 션휴도 쁨옵자 크옵트 이샹 러표샹위 .라니뮬샤 트이조로쿳 을위하


음성기술
인공지능 전문 기업 셀바스AI가
이끌어가겠습니다
셀바스AI는 음성인식 및 음성합성 등 핵심 음성기술 분야의
독보적 원천 기술을 보유한 국내 대표 음성기술 기업입니다
셀바스 AI의 인공지능 대표 브랜드인 Selvy를 음성기술에 접목하여
꾸준한 연구개발과 음성인식 성능 향상에 힘쓰고 있습니다
selvas AI
회사 홈페이지 www.selvasai.com
음성기술 홈페이지 speech.selvasai.com
셀비 홈페이지 selvy.ai
Tel. 02 852 7788
E-mail. support@selvasai.com

Proceedings

of the 2020 Fall Conference
of the Korean Society of Speech Sciences





November 28, 2020
online ZOOM Webinar
The Korean Society of Speech Sciences