ISSN 2005-9299

2021 한국음성학회 가을 학술대회

발표 논문집

주제: 딥러닝 기반의 음성 연구를 위한 음성학/공학/말장애의 학제적 접근

날짜: 2021년 11월 19일(금), 20일(토)

장소: 온라인

주최: 사단법인 한국음성학회

주관: 사단법인 한국음성학회, 서울대학교

후원: 한국연구재단, ㈜LG유플러스, (사)한국언어재활사협회,
㈜리드스피커코리아, ㈜셀바스AI

한국음성학회
The Korean Society of Speech Sciences

이 발표논문집은 2021년도 정부재원(교육부)으로 한국연구재단의 지원을 받아 발간되었음

모시는 글

한국음성학회는 11월 19, 20일 양일에 걸쳐 온라인으로 가을 학술대회를 개최합니다. 가을 학술대회는 우리 학회의 가장 큰 행사로서 그동안의 학술적 성과를 발표하고 의견을 교환하는 소중한 자리입니다. 올해는 학문과 기술의 최신 동향을 살펴볼 수 있는 3편의 특강과 회원들의 학술적 성과를 공유하는 48여 편의 구두 및 포스터 발표가 마련되어 있습니다. 풍성한 학술대회가 될 수 있도록 초청 특강에 응해 주신 강연자들과 적극적으로 논문 발표를 신청해 주신 회원 여러분께 감사의 말씀 드립니다. 학술대회 준비위원회와 조직위원회 관계자 여러분의 도움이 없었다면 학술대회는 불가능했을 것입니다. 가을 학술대회를 준비하느라 노고를 아끼지 않은 이주경 학술위원장을 비롯한 학술대회 준비위원회 여러분에게 감사의 말씀을 드립니다. 서울대 스마트 강의실 운용을 중심으로 학술대회의 조직과 준비에 큰 도움을 주신 정민화 교수님과 서울대학교 인문대학에 감사를 표합니다. 이 자리를 빌려 한국연구재단, ㈜LG유플러스, (사)한국언어재활사협회, ㈜리드스피커코리아, ㈜셀바스AI 등 후원해 주신 여러 기관에 감사의 말씀을 드립니다. 학술대회 준비를 도와주신 회장단 및 학회 관계자 여러분의 노고에도 감사드립니다.

한국음성학회 가을 학술대회는 전통적으로 단풍과 함께 가을을 맺고 첫눈과 더불어 겨울을 맞는 때에 열려 왔습니다. 전염병 창궐의 여파로 학회 활동이 위축되지 않을까 노심초사하며 보낸 한 해였습니다. 최근 일상적 생활로 전환되고 있어 다행으로 생각합니다. 내의 흐름은 쉼이 없고 연못의 맑음은 그림자를 취한다고 합니다. 학술대회와 총회를 마지막으로 학회의 큰 행사는 마무리됩니다만 회원들의 학문 활동은 쉼이 없겠지요. 이제 고요한 성찰을 통해 지난해를 되돌아보고 알찬 한 해를 준비하는 시간이 남아 있습니다. 남은 올해 잘 마무리하시고 내년에도 건강과 행복이 함께 하시기를 기원합니다.

2021년 11월

한국음성학회 회장 박 한 상

2021 한국음성학회 가을 학술대회 조직위원회

학술대회장	회장	박한상(홍익대)
조직위원회	위원장	성철재(충남대)
	위원	정민화(서울대), 홍기형(성신여대), 장태엽(한국외대), 이주경(서울시립대), 최성희(대구가톨릭대), 윤원희(계명대), 박상희(대구사이버대)
학술위원회	위원장	이주경(서울시립대)
	위원	김지연(우송대), 남호성(고려대), 박기영(ETRI), 박형민(서강대), 송윤경(동명대), 안현기(서울대), 오재혁(건국대)

2021 한국음성학회 가을 학술대회 일정표

11월 19(금) 서울대학교 스마트강의실 및 zoom 플랫폼을 이용한 웹퍼런스

<table>
<tr><th>시 간</th><th colspan="3">발 표 및 내 용</th></tr>
<tr><td>10:00~10:20</td><td colspan="3">개회식 (서울대학교 스마트강의실에서 실시간 온라인 송출)</td></tr>
<tr><td>10:20~11:20</td><td colspan="3">특강 I (서울대학교 스마트강의실에서 실시간 온라인 송출)</td></tr>
<tr><td>11:20~11:30</td><td colspan="3">휴식</td></tr>
<tr><td rowspan="2">11:30~12:30</td><td colspan="3">구두 발표 I (zoom을 이용한 실시간 온라인 미팅)</td></tr>
<tr><td>음성학(3편)</td><td>말장애(3편)</td><td>음성공학(3편)</td></tr>
<tr><td>12:30~13:40</td><td colspan="3">점심</td></tr>
<tr><td>13:40~14:50</td><td colspan="3">포스터 (홈페이지에 포스터 게시 및 zoom을 이용한 실시간 온라인 미팅)</td></tr>
<tr><td>14:50~15:00</td><td colspan="3">휴식</td></tr>
<tr><td>15:00~16:00</td><td colspan="3">특강 II (서울대학교 스마트강의실에서 실시간 온라인 송출)</td></tr>
<tr><td>16:00~16:10</td><td colspan="3">휴식</td></tr>
<tr><td rowspan="2">16:10~17:10</td><td colspan="3">구두 발표 II (zoom을 이용한 실시간 온라인 미팅)</td></tr>
<tr><td>음성학(3편)</td><td>말장애(3편)</td><td>음성공학(3편)</td></tr>
<tr><td>17:10~17:20</td><td colspan="3">휴식</td></tr>
<tr><td>17:20~17:50</td><td colspan="3">총회 (서울대 스마트강의실에서 온라인 송출)</td></tr>
</table>

11월 20일(토)

<table>
<tr><th>시 간</th><th colspan="3">발 표 및 내 용</th></tr>
<tr><td>10:00~11:00</td><td colspan="3">특강 III (서울대학교 스마트강의실에서 실시간 온라인 송출)</td></tr>
<tr><td>11:00~11:10</td><td colspan="3">휴식</td></tr>
<tr><td rowspan="2">11:10~12:10</td><td colspan="3">구두 발표 III (zoom을 이용한 실시간 온라인 미팅)</td></tr>
<tr><td>음성학(3편)</td><td>말장애(3편)</td><td>음성공학(3편)</td></tr>
<tr><td>12:10~12:30</td><td colspan="3">스펙트로그램 리딩 (서울대학교 스마트강의실에서 실시간 온라인 송출)
사회: 이주경(서울시립대)</td></tr>
<tr><td>12:30~12:50</td><td colspan="3">시상식 및 폐회식 (서울대학교 스마트강의실에서 실시간 온라인 송출)
사회: 이주경(서울시립대)</td></tr>
</table>

2021 한국음성학회 가을 학술대회 세부 일정표

2021년 11월 19일(금요일)

특강 I (서울대학교 스마트강의실에서 실시간 온라인 송출)

좌장: 박기영 (한국전자통신연구원)

시간	내용
10:20 ~ 11:20	"딥러닝 기반의 음성지능 연구와 실제적 응용" 박전규(한국전자통신연구원)

구두 발표 I (zoom을 이용한 실시간 온라인 미팅)

음성학 및 음운론	**좌장: 안현기(서울대)**			
	시간	구분	제목	저자
	11:30~11:50	PH1	Laryngeal movements in the production of Korean stop consonants	Sunhee Kim(SNU), Tran, Savariaux, Gerber, Vallée(Univ. Grenoble Alpes), Inyoung Kim(NAVER LABS Europe)
	11:50~12:10	PH2	The role of F0 trajectory in the emotion identification	Tae-Jin Yoon (Sungshin Women's University)
	12:10~12:30	PH3	대화 참여자의 성별에 따른 발화의 시간적 실현 양상	유도영, 신지영(고려대학교)
말장애 및 음성의학	**좌장: 송윤경(동명대)**			
	시간	구분	제목	저자
	11:30~11:50	SD1	만 4-6세 아동의 운율구 위치에 따른 치조마찰음 /ㅅ/의 특성	이호, 성철재(충남대학교)
	11:50~12:10	SD2	한국어 'ㅅ' 연장음에 대한 지각 연구	박진(가톨릭관동대학교), 박소현(충남대학교)
	12:10~12:30	SD3	진동센서를 이용한 공명발성, 입술트릴, 튜브 발성시 전기성문파형과 안면진동 특성	안지호, 최성희, 이경재, 최철희 (대구가톨릭대학교)
음성공학	**좌장: 박형민(서강대)**			
	시간	구분	제목	저자
	11:30~11:50	SE1	법과학적 활용을 위한 삼성 스마트폰 음성 녹음 파일 원본과 편집본의 메타데이터 구조 및 속성 분석	안서영, 유세희(성신여자대학교), 김경화(대검찰청), 홍기형(성신여자대학교)
	11:50~12:10	SE2	장애발화 음성 코퍼스 구축 시 필요한 메타데이터 제안	이선우, 김선희, 정민화(서울대학교)
	12:10~12:30	SE3	한국어 감정 인식 딥러닝 알고리즘 개발에 대한 연구	김태용, 이보원(인하대학교)

포스터(홈페이지 내 '학술대회'에 포스터 게시 & zoom을 이용한 실시간 미팅)
좌장: 공은정(항공대), 이수복(우송대), 홍기형(성신여대)

시간	구분	제목	저자
13:30 ~ 15:00	P01	뇌성마비 성인과 일반 성인 및 뇌성마비 아동 간의 모음 /아/ 음향학적 특성 비교	정필연, 심현섭(이화여자대학교)
	P02	청유문과 의문문 읽기 과제를 통한 자폐아동과 일반아동의 운율특성 비교	김현경, 성철재(충남대학교)
	P03	성대결절 아동의 연속발화에 대한 공기역학적 특성	정혜주, 최성희, 이경재, 최철희(대구가톨릭대학교)
	P04	언어재활사의 주관적·객관적 음성피로도	전혜원, 성철재(충남대학교)
	P05	Usefulness of Vocal Fatigue Index for evaluation of laryngeal hypertension	Ji Sung Kim, Dong Wook Lee (ChungBuk National University Hospital)
	P06	구어처리과제의 발달: 문장폭기억과제를 중심으로	오소정(동명대학교)
	P07	1차원 합성곱 신경망 구글넷 기반 스테레오 음성의 도래 방향 추정	이정혁, 김홍국(광주과학기술원)
	P08	상호의존정보를 이용한 임베딩 기반 음성감정인식	박순찬, 김형순(부산대학교)
	P09	도메인 적대적 학습 기반의 억양 음성 인식	나형주, 장민지, 나희정, 박정식(한국외국어대학교)
	P10	L2 화자의 유창성 수준에 따른 짧은 발화 음성 대상 언어 식별	나희정, 나형주, 장민지, 박정식(한국외국어대학교)
	P11	다자간 대화 음성인식 시스템 개발	박전규, 강점자, 동성희, 박기영, 오유리, 이성주, 최우용(한국전자통신연구원)
	P12	Graph Attention Network를 활용한 프레임 단위 화자 특징 결합	허정우, 심혜진(서울시립대학교), 박재한, 이가희(KT), 유하진(서울시립대학교)
	P13	Transformer 및 Parallel WaveGAN을 이용한 다화자 음성합성	최연주, 엄지섭, 김회린(KAIST)
	P14	Multi-Task 기반의 공감형 발화 분류 모델	김종인, 정민화(서울대학교)
	P15	지각 훈련을 통한 한국어 폐쇄음 음향 신호 가중치의 L2 학습	오은진(이화여자대학교)
	P16	The influences of processing levels on the perception-production link of L2 phonotactics	Song Yi Kim, Jeong-Im Han (Konkuk University)

시간	구분	제목	저자
13:40 ~ 14:50	P17	Word recognition in English coronal place assimilation: Eye tracking study	Eunkyung Sung(CUFS), Sehoon Jung(Kyungsung University), Sunhee Lee(CUFS), Seulgi Oh(HUFS)
	P18	프랑스어 강세 음절에 오는 비강 모음의 음향적 특징	박혜숙, 김선희(서울대학교)
	P19	코로나19로 인한 마스크 착용이 아동 말소리 발달에 미치는 영향 연구	박범준, 윤수연(충남대학교)
	P20	Different phonetic reduction patterns of English function words between native English and Korean speakers	Wooji Park, Seok-Chae Rhee (Yonsei University)
	P21	Korean learners' production of English /l/: Ultrasound and acoustic analyses	Joo-Kyeong Lee(University of Seoul)

특강 II (서울대학교 스마트강의실에서 실시간 온라인 송출)

좌장: 박상희(대구사이버대)

시간	내용
15:00 ~ 16:00	"언어병리학 음성장애 영역에서의 AI 연구 최신 경향 및 임상 특성" 김근효(부산대병원)

구두발표 II

음성학 및 음운론	**좌장: 오은진(이화여대)**			
	시간	구분	제목	저자
	16:10~16:30	PH4	An acoustic study of Korean mothers' vowel space	Eon-Suk Ko(Chosun University), Sunghye Cho(University of Pennsylvania)
	16:30~16:50	PH5	서울말 내포문 wh-섬 제약의 지각 및 반응시간 연구	윤원희(계명대학교)
	16:50~17:10	PH6	Research on English Speakers' Production of Word Boundary Stops Based on Voicing Found in Spontaneous Speech	Yungdo Yun(Dongguk University)
말장애 및 음성의학	**좌장: 이옥분(대구사이버대)**			
	시간	구분	제목	저자
	16:10~16:30	SD4	마비말장애 음성인식 성능에 영향을 미치는 언어학적 특징 분석	여은정, 김선희, 정민화(서울대학교)
	16:30~16:50	SD5	연축성 발성장애 환자에서 발화범위 프로파일의 특성	이승진(한림대학교), 김재옥(강남대학교), 임성은(강남세브란스병원), 임재열(연세대학교)
	16:50~17:10	SD6	말소리장애아동에 대한 인공지능과 언어재활사의 음성인식 비교	천시온, 최성희, 이경재, 최철희 (대구가톨릭대학교)
음성공학	**좌장: 유하진(서울시립대)**			
	시간	구분	제목	저자
	16:10~16:30	SE4	Attention 인공신경망을 통한 한국어 방언의 억양 패턴 학습 및 방언 식별	이주영(서울대학교), 김경화(대검찰청), 정민화(서울대학교)
	16:30~16:50	SE5	트랜스포머 기반 실시간 한국어 음성인식	오유리, 박기영(한국전자통신연구원)
	16:50~17:10	SE6	실시간 시청각 발화구간검출 알고리즘	정세영, 박형민(서강대학교)

2021년 11월 20일(토요일)

특강 III (서울대학교 스마트강의실에서 실시간 온라인 송출)

좌장: 정현성(한국교원대)

시간	내용
10:00 ~ 11:00	"AI 시대, 음성학의 방향" 남호성(고려대)

구두발표 III

음성학 및 음운론	**좌장: 성은경(사이버한국외대)**			
	시간	구분	제목	저자
	11:10~11:30	PH7	A longitudinal study of individual and age-related differences in perception-production links in second language speech	Donghyun Kim (Kumoh National Institute of Technology)
	11:30~11:50	PH8	The Effects of Focus-on-Form Pronunciation Instruction on The Evaluation of High School Students' Utterances	Ji Sun Yuk (Sejong Global High School)
	11:50~12:10	PH9	프랑스인 한국어 L2 학습자의 평음, 격음, 경음의 지각 습득 과정과 패턴 - 선택적 주의를 중심으로	이보람(소르본 누벨 대학교)
말장애 및 음성의학	**좌장: 김지연(우송대)**			
	시간	구분	제목	저자
	11:10~11:30	SD7	안면 마스크 착용이 발화의 음향학적 특성에 미치는 영향	김덕애, 최성희, 이경재, 최철희(대구가톨릭대학교)
	11:30~11:50	SD8	변성기 남성의 발화범위 및 음성범위 프로파일	김재옥(강남대학교), 이승진(한림대학교)
	11:50~12:10	SD9	발화 환경에 따른 한국어 모음 /오, 우, 으/의 포먼트 특성	박지연, 성철재(충남대학교)
음성공학	**좌장: 김경화(대검찰청)**			
	시간	구분	제목	저자
	11:10~11:30	SE7	전체 맥락과 깊이별 분리 합성곱을 이용한 Conformer 기반 음성 인식기의 순방향 모듈 개선	정승훈, 김홍국(광주과학기술원)
	11:30~11:50	SE8	영어-한국어 대화체 자동통역을 위한 Cascade 및 End-to-End 접근 방식 비교	방정욱, 이민규, 윤승, 김상훈(한국전자통신연구원)
	11:50~12:10	SE9	딥러닝 기반 단일 클래스 분류를 사용한 한국어 키워드 검출	이승현, 박형민(서강대학교)

차 례

포스터발표 I

특강 II

구두발표 II

[음성학 및 음운론]

[말장애 및 음성의학]

[음성공학]

특강 I

좌장: 박기영(한국전자통신연구원)

딥러닝 기반의 음성지능 연구와 실제적 응용
(박전규, 한국전자통신연구원)

딥러닝 기반의 음성지능 연구와 실제적 응용

박 전 규
한국전자통신연구원 복합지능연구실

Deep Learning-based Speech Intelligence Research and its Practical Applications

Jeon Gue Park
Integrated Intelligence research Section, ETRI
jgp@etri.re.kr

딥러닝 기술 혁신에 따라 시각 및 언어지능 영역과 함께 음성지능도 비약적 성능 개선을 거듭하고 있다. 이러한 기술적 고도화를 배경으로 기존의 음성 검색, AI 비서, 콜센터 녹취 등의 응용영역에 더해서 다양한 회의, 상담, 대담 등을 대상으로 하는 회의록 서비스가 확산 일로에 있다. 그러나 현재의 음성인식 기술은 다수의 참여자가 다양한 거리와 위치에서 발성하는 경우, 과도한 소음 환경 또는 예측하지 못한 소음이 발생하는 경우, 화자간의 발성겹침이 있는 경우, 유아나 노인 등이 구사하는 사투리 등을 포함하는 자연스런 구어체 발성에 대해서는 심각한 정확도 저하가 발생하고 있다. 나아가 전문 영역에 구애받지 않는 자연스런 대화처리, 사용 언어에 제약이 없는 실시간 자동통역, 장시간 대화를 받아적고 사람처럼 요약해 주는 딕테이션, 성우의 연기나 동화 구연 수준의 자연스런 음성합성 등과 같은 음성지능의 궁극적 목표는 여전히 일반의 기대와 다르게 난제로 남아 있다. 본 고에서는 이와 같은 물리적, 환경적, 인적 성능병목 요인을 극복하는 고난도 음성지능 연구개발 현황을 살펴보고 현재 수준의 기술에 기반하는 다양한 실제적 응용 사례를 고찰해 보고자 한다.

[감사의 글] 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2019-0-01376, 다중 화자간 대화 음성인식 기술개발)

구두 발표 I: 음성학 및 음운론

좌장: 안현기(서울대)

Laryngeal movements in the production of Korean stop consonants

Sunhee Kim[1], Thi Thuy Hien Tran[2], Christophe Savariaux[2],
Silvain Gerber[2], Nathalie Vallée[2], Inyoung Kim[3]
[1]Dept. of French Language Education, Seoul National University, Korea
[2]GIPSA-lab/Univ. Grenoble Alpes, France, [3]NAVER LABS Europe, France
sunhkim@snu.ac.kr, {thi-thuy-hien.tran, christophe.savariaux, silvain.gerber,
nathalie.vallee}@gipsa-lab.fr, inyoung.kim@naverlabs.com

In Korean, as in some Asian languages, syllable final stops are often produced without a perceptually salient burst due to a non-audible release of the occlusion (Edmondson et al., 2010) and are called unreleased stops (Ladefoged & Maddieson, 1996). Although this tendency is phonologically well described as the result of a lenition process (Cho et al., 2002), it is still not well understood phonetically. The aim of our study is to describe the production process of Korean non-release stops taking into account laryngeal and supralaryngeal articulations and in particular to test the hypothesis of a laryngeal action that would contribute to a decrease in air pressure behind the occlusion causing a non-audible release (Tran et al., 2021). Using an auxiliary of the EVA-2* connected to an electroglottograph EG2-PCX2, we recorded the vertical movement of the larynx and the signal of the vocal fold oscillations in synchronization to the audio recording of the acoustic signal. Twenty-one Korean native speakers (10 male speakers, 11 female speakers) participated in the experiment. The corpora consist of 3 repetitions of 34 monosyllabic and disyllabic words inserted in a carrier sentence. A systematic lowering of the larynx accompanies productions of /p, t, k/ with differences in amplitude movement between onset and coda positions and between male and female speakers was observed. Results also suggest that Korean unreleased stops are not glottalized.

Évaluation Vocale Assistée (EVA): Assisted Voice Assessment. “The EVA2™ Assisted Voice Assessment system is designed to study most parameters of speech production: sound, pitch, voice intensity, air flow rates, pressures...”(http://www.sqlab.fr/evaRootFR.htm)

The role of F0 trajectory in the emotiom identification

Tae-Jin Yoon
Dept. of English Language and Literature, Sungshin Women's University
tyoon@sungshin.ac.kr

Modulation of pitch, loudness, duration and voice quality across the syllables in an utterance conveys both linguistic such as prominence, prosodic phrasing and non-linguistic information such as speaker's emotional status. Speakers' emotional states are psychological states signally by neurophysiological changes and associated with thoughts, feelings and behavioral responses. When uttered, the utterance emits emotional status of the speaker as well as linguistic information. This paper reports an acoustical analysis of RAVDESS(The Ryerson Audio-Visual Database of Emotional Speech and Song) Emotion Speech Corpus (Livingstone & Russo, 2018). The corpus contains seven types of emotion (calm, happy, sad, angry, fearful, disgust emotions in addition to a neutral emotion). Some previous works have addressed the roles of intonation, especially which of pitch contour patterns in manifesting emotional speech. However, the intonation parameters that have been used before were rather simplistic, in that simplified F0 parameters such as max F0, min F0, F0 range are used. This simplistic approach may be partially responsible for the low classification rates of emotional types. The aim of the paper is to illustrate that dynamic f0 contours, as modeled using generalized additive modeling method, can reliably serve as a behavioral response for speaker's emotional states. The baseline model of F0 change over time for each emotion with speaker as a random effect could account for 57.9% of the deviance. The baseline model is augmented with additional fixed effects such as Intensity, Utterance Duration per emotion and random effects such as speaker by emotion, the model could account for up to 76.3% of the total deviance. The findings of the study confirm previous observations about the informative role of pitch in expressing emotion.

대화 참여자의 성별에 따른 발화의 시간적 실현 양상

유 도 영, 신 지 영
고려대학교 국어국문학과

Temporal Structure of Utteracne according to Gender of Interlocutors

Doyoung Yoo, Jiyoung Shin
Dept. of Korean Language and Literature, Korea University
heyd723@gmail.com, shinjy@gmail.com

이 연구는 화자의 성별과 대화 상대방의 성별에 따라, 발화 내 조음 구간과 휴지 구간의 시간적 실현 양상에 차이가 있는지 살피고자 한다. 이를 통하여 사회언어학적 차이가 발화의 시간적 구조에 어떠한 영향을 주는지 고찰하는 것을 연구의 목표로 한다.

앞선 연구들에 따르면, 발화의 시간적 실현은 대화 참여자의 성별에 따라 다른 실현 양상을 보인다. 독백 발화를 관찰한 유도영·신지영(2019, 2020)에 따르면, 남성 화자는 여성 화자보다 더 길고 빈번한 휴지를 산출하며, 더 짧은 조음 구간의 길이를 보인다고 하였다. 발화자의 성별 이외에도 대화에 참여하는 상대의 성별에 의하여 발화의 운율 패턴이 달라질 수 있는데, Gravano et al. (2011)은 대화 상대방의 성별이 무엇인지에 따라, 발화 내용을 계획할 때에 사용되는 화자의 운율 실현이 달라질 수 있음을 관찰하였다. 이러한 연구 결과는 화자와 대화 상대방의 성별이 전체적인 발화의 시간적 패턴을 다르게 만드는 주요한 변수일 수 있음을 시사한다. 따라서 발화의 전반적인 시간 패턴을 연구하기 위해서는 화자의 성별뿐 아니라 대화 상대방의 성별 또한 고려되어야 한다. 따라서 이 연구는 발화자와 대화 상대방의 성별을 모두 고려하여, 성별에 따라 발화의 시간적 실현이 어떠한 차이를 보이는지 살피고자 하였다.

이를 살피기 위하여, 친밀도와 나이가 통제되고 성별을 변수로 삼은 20대의 대화 자료 18개를 분석하고, 발화 길이, 조음 구간의 길이, 발화 내 휴지 길이, 조음 속도를 살폈다. 분석 결과 남성 화자의 발화 길이가 여성 화자의 발화 길이보다 짧았고, 남성 화자의 휴지 길이가 여성 화자의 휴지 길이보다 길었다. 또 대화 상대방의 성별에 따라서도 실현 양상에 차이를 보였는데, 남성 화자는 여성과 대화할 때보다 남성과 대화할 때 조음 구간의 길이와 휴지 길이 및 발화 길이가 길어졌다. 여성은 여성 상대와 대화할 때 조음 구간의 길이가 더 길어지고 발화 속도가 더 빨라지는 결과를 보였다. 이와 같은 결과를 통하여 화자의 성별뿐 아니라 상대의 성별 또한 발화의 운율적 실현을 달리하는 주요한 변수일 수 있음을 확인하였다.

참고 문헌

유도영·신지영. 2019. 과제, 성별, 세대에 따른 휴지의 실현 양상 연구. 말소리와 음성과학, 11(2), 33-44.
유도영·신지영. 2020. 휴지 단위와 호흡 단위의 실현 양상 연구. 말소리와 음성과학, 12(1). 19-31.
Gravano, A., Levitan, R., Willson, L., Beňuš, Š., Hirschberg, J., & Nenkova, A. (2011) Acoustic and prosodic correlates of social behavior, INTERSPEECH-2011, 97-100.

구두 발표 I: 말장애 및 음성의학

좌장: 송윤경(동명대)

만 4~6세 아동의 운율구 위치에 따른 치조마찰음 /ㅅ/의 특성

이 호, 성 철 재
충남대학교 언어병리학과

Characteristics of alveolar fricative ‘s’ according to position in prosodic phrase in children aged 4~6 years

Ho Lee, Cheol Jae Seong
Dept. of Speech-Language Pathology, Chungnam National University
soul_hoya@naver.com, cjseong49@gmail.com

우리말의 마찰음 중 치조마찰음 /ㅅ/는 2~3세에 출현하여 6, 7세이 완전히 습득된다(김영태, 1996). 치조마찰음 /ㅅ/는 가장 늦게 발달이 이루어지고 조음하기 어려운 음소이기에 조음할 때 생략, 대치, 왜곡 등의 오류가 나타난다(전희정·이승환, 1999). 이러한 이유로 치조마찰음 /ㅅ/에 대한 연구는 지속적으로 이루어져 왔다. 치조마찰음 /ㅅ/에 대한 연구는 시간적 분석, 주파수 정보, 스펙트럼 모먼트 분석 등 음향적 평가에 대한 연구와 정상 발달과 특정 장애 아동과의 차이, 연령별 특성, 오류패턴과 음절 구조에 대한 청지각적 연구로 활발히 진행되고 있다. 현재까지의 연구에서 나아가 연결발화에서 운율구 위치에 따라 나타나는 치조마찰음 /ㅅ/의 특성을 살펴보고자 하였다.

연구대상은 정상발달 범주에 속한 4세 17명, 5세 18명, 6세 19명으로 총 54명의 아동이 연구에 참여하였다. 녹음과제는 치조마찰음 /ㅅ/와 모음(/ㅏ/, /ㅖ/, /ㅣ/, /ㅗ/, /ㅜ/)이 결합된 단어로 2개의 강세구를 구성한 5개의 문장으로 아동이 그림을 보고 발화할 수 있도록 5장의 그림을 제작하였다. 아동이 산출한 문장에서 첫 번째 강세구와 두 번째 강세구의 음성을 추출하였고, Praat의 ExperimentMFC를 활용하여 평가를 위한 실험용 플랫폼을 구성하였다.

1차 청지각 평가는 조음정확도 평가로 제시된 음성파일을 듣고 얼마나 정확하게 들리는지 1~5점 척도로 평가하게 하였다. 2차 청지각 평가는 청자의 반응 평가로 음성파일을 듣고 가장 가깝게 들리는 음소를 선택하여 반응을 평가하였다. 1차 청지각 평가에는 운율구 내 위치, 연령, 모음에 따른 결과를 알아보기 위해 삼원분산분석을 실시하였고, 2차 청지각 평가에서는 /ㅅ/로 정조음된 자료를 대상으로 왜도와 무게중심을 관찰변수로 놓고 삼원분산분석을 실시하였다. 통계처리는 SPSS 26을 이용하였다.

<표 1> 운율구 내 위치와 연령, 모음환경에 따른 조음정확도 기술통계

모음	첫 번째 강세구			두 번째 강세구		
	4세	5세	6세	4세	5세	6세
a	3.13(±1.567)	3.83(±1.508)	4.11(±1.195)	2.90(±1.591)	3.62(±1.604)	3.98(±1.261)
e	3.08(±1.650)	3.60(±1.596)	3.56(±1.539)	2.91(±1.631)	3.55(±1.526)	3.65(±1.457)
i	3.12(±1.606)	3.99(±1.274)	3.66(±1.381)	2.19(±1.445)	2.71(±1.480)	3.02(±1.474)
o	3.16(±1.672)	4.02(±1.305)	4.24(±1.081)	2.65(±1.482)	3.12(±1.434)	3.42(±1.477)
u	2.70(±1.587)	2.98(±1.487)	3.36(±1.458)	2.79(±1.489)	3.30(±1.493)	3.74(±1.284)

운율구 내 위치와 연령, 모음환경에 대한 평균과 표준편차는 <표 1>과 같다. 결과를 살펴보면, 조음정확도에서 운율구 내 위치, 연령, 모음환경에 대한 주효과는 유의미하게 나타났다($p<.001$). 또한 위치와 모음

환경에 대한 상호작용 효과에서도 유의미한 결과가 나타났다(p<.001). 위치와 모음환경(p=.204), 연령과 모음환경(p=.133), 위치와 연령과 모음환경(p=.438)에 대한 상호작용 효과는 나타나지 않았다.

<표 2> 운율구 내 위치에 따른 연령별 청자의 반응 빈도

	강세구	첫 번째 강세구			두 번째 강세구			총빈도
	연령	4세	5세	6세	4세	5세	6세	
청자 반응	[s, ɕ]	52	71	83	45	66	82	395
	기타	28	12	8	31	18	8	109
빈도		80	83	91	76	84	90	504

운율구 내 위치에 따른 연령별 청자의 반응 빈도는 <표 2>에 제시하였다. [s]와 /시/의 [ɕ]를 제외한 다른 음소를 기타음소로 봤을 때, 연령이 증가할수록 기타 음소에 반응하는 빈도가 낮아졌다. 그리고 연령이 증가할수록 운율구 위치와 상관없이 /ㅅ/로 반응하는 것으로 나타났다.

<표 3> 운율구 내 위치와 연령, 모음환경에 따른 왜도와 무게중심 기술통계

	모음	첫 번째 강세구			두 번째 강세구		
		4세	5세	6세	4세	5세	6세
왜도	a	5.704(±5.524)	0.769(±1.518)	4.056(±5.107)	3.395(±2.054)	1.971(±3.527)	2.631(±3.450)
	e	2.592(±3.685)	1.928(±3.408)	2.828(±6.011)	4.016(±3.759)	1.988(±3.396)	2.107(±2.619)
	i	1.177(±1.645)	0.329(±0.729)	1.452(±1.635)	3.342(±3.630)	0.694(±1.305)	1.653(±3.752)
	o	4.059(±5.594)	2.166(±2.795)	1.233(±2.184)	2.774(±3.324)	2.212(±2.128)	3.004(±3.876)
	u	5.229(±7.646)	2.587(±3.771)	1.623(±3.040)	4.863(±6.491)	1.025(±1.871)	1.265(±2.275)
무게중심	a	1975.142 (±2871.752)	5146.135 (±3237.529)	3293.226 (±3703.158)	2046.469 (±1931.976)	4341.372 (±3110.747)	3390.025 (±2761.72)
	e	3687.234 (±3247.186)	5236.091 (±4006.377)	4797.678 (±3588.606)	2846.348 (±3679.436)	4762.315 (±3376.399)	3724.005 (±2666.338)
	i	4388.058 (±2376.683)	5919.24 (±2162.785)	3798.589 (±2255.493)	3630.48 (±3768.299)	5237.708 (±2557.653)	4477.985 (±1999.174)
	o	2603.828 (±2780.126)	4141.966 (±3684.299)	4835.348 (±3554.18)	3332.918 (±3857.235)	3160.489 (±2541.035)	3742.098 (±3803.956)
	u	4263.297 (±3659.619)	3657.364 (±3162.725)	4739.255 (±2938.68)	3798.469 (±3978.618)	4934.509 (±3716.653)	4603.568 (±3055.252)

정조음으로 평가된 음성자료에 대한 왜도와 무게중심의 기술통계 값은 <표 3>과 같다. 통계 결과, 왜도는 연령에서 주효과가 관찰되었으나(p<.001), 다른 변수에서는 유의하지 않았다. 무게중심에서도 연령 주효과가 유의하게 나타났고(p<.005), 모음환경과의 상호작용 효과는 관찰되지 않았다.

본 연구는 치조마찰음 /ㅅ/의 특성을 운율구 내 위치에 따른 청지각적 특성과 음향학적 관점으로 살펴보았다. 연령이 높을수록 모든 모음 환경에서 /ㅅ/의 조음정확도와 정조음으로 반응한 빈도가 높게 나타난 것은 김영태(1996), 전희정·이승환(1999)의 연구결과와 같이 조음이 완성되어 가는 것을 알 수 있다. 왜도와 무게중심에서도 연령별 유의한 차이가 나타난 것은 연령이 어릴수록 조음의 정확도와 정교함이 떨어지는 것을 나타낸다. 향후 조음 치료시 문장 수준의 훈련 방법의 적용과 운율 특성을 고려한 연구가 지속되길 기대한다.

1. 2. 참고문헌

3. [1] 김영태(1996), 그림자음검사를 이용한 취학 전 아동의 자음정확도 연구.『말-언어장애연구』, 7-34
4. [2] 전희정·이승환(1999), 2-7세 아동의 /ㅅ/와 /ㅆ/ 말소리 발달 연구. 이화여자대학교 대학원 석사학위논문
5. [3] 김지연·성철재(2018), 아동이 산출한 치조마찰음 /ㅅ/에 대한 청지각적·음향학적 연구.『말소리와 음성과학』, 제 10권 3호(p41-48).

한국어 'ㅅ' 연장음에 대한 지각 연구

박 진*, 박 소 현**
가톨릭관동대학교 언어재활상담학과*
충남대학교 언어병리학 학과간협동과정**

Korean Listeners' Identification and Discrimination of Prolonged /ㅅ/ as Prolongations

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

목적: 본 연구는 기본적으로 한국어 화자들을 대상으로 특정 말소리 연장에 대한 지각 양상을 알아보고자 한다. 구체적으로 한국어 평마찰음 'ㅅ'의 연장음 인식에 있어 범주적인(categorical) 또는 연속적인(continuous) 특성을 보이는지, 그리고 이러한 지각 양상에 남녀 간 차이가 나타나는지를 알아보고자 한다. **연구방법:** 한국어를 모국어로 하는 대학생 50명(남, 여 각 25명)을 대상으로 '가을'문단(김향희, 1996) 내 특정 문장에 포함된 평마찰음 'ㅅ'의 길이를 20 ms에서 정수배로 1000 ms까지(즉, 20 ms, 40 ms, 60 ms ..... 960 ms, 980 ms, 1000 ms) 연장한 총 51개의 문장자극을 생성해 무작위로 들려준 후, 이를 비정상 지각 정도와 관련해 1에서 100점까지 척도(100으로 갈수록 더 비정상적)로 측정하도록 하였다. 곡선추정회귀분석을 통해 마찰음의 연장 길이가 청각적으로 인식하는 부자연스러운 정도의 점수(1에서 100까지)에 미치는 영향과의 관계를 회귀식으로 나타냈으며, 이를 위해 모형설명력 값(수정된 R제곱)이 가장 높은 회귀식을 선택하였다. 그리고 비정상 인식 관련 남녀 차이를 알아보기 위해 one way ANOVA를 실시하였다. **연구결과 및 논의:** 분석결과, 마찰음의 연장 길이와 관련된 청각 인식 점수(1에서 100까지)간의 관계는 삼차모델(cubic model)에 가까웠으며, 관련해 남녀 차이는 나타나지 않았다. 이러한 연구결과는 말더듬 연장 인식 및 평가와 관련해 심도있는 논의와 통찰을 제공할 것으로 사료된다.

진동센서를 이용한 공명발성, 입술트릴, 튜브 발성시 전기성문파형과 안면진동 특성

안지호, 최성희, 이경재, 최철희
대구가톨릭대학교 언어청각치료학과

Characteristics of eletroglottography and facial vibrations using vibrate sensor during resonance voice, lip-trill, tube phonation

Sion Cheon, Seong Hee Choi, Kyoungjae Lee, Chul-Hee Choi
Dept. of Audiology & Speech Language Pathology, Daegu Catholic University,
a_jiho@naver.com, shgrace@cu.ac.kr, kjlee0119@cu.ac.kr, cchoi@cu.ac.kr

음성치료시 성대 건강과 좋은 음질을 위해 공명 음성 산출을 목표로 한다. 본 연구는 /ㅏ/ 모음 연장발성, 공명발성(허밍발성, 카주발성), 입술 트릴, 튜브발성(공기 중 혹은 물안에서) 시 성대접촉과 안면진동특성을 살펴 보기 위하여 진동센서와 전기성문파형검사를 사용하였다. 대부분의 과기능 음성장애 환자들은 목주변의 긴장과 강한 성대 접촉의 특성을 보이므로, 음성 치료 시 소리 산출의 중심 부위를 목수준에서 위쪽으로 이동하도록 하여 후두의 긴장을 줄이고 올바른 발성을 유도하도록 한다. 공명 발성은 발성하기 쉽고, 안면 조직에서 진동을 통해 생성되는 음성을 의미하는데, 발성 시 억압된 음성도 아니고, 기식화된 음성도 아닌 최적의 성대 내전을 목표로 한다. 본 연구에서는 최근 3개월 이내 감기를 앓지 않으며, 후두질환이 없고, 음성 문제를 보고하지 않으며, 1급 언어재활사에 의한 GRBAS청지각적 평가에서 G0으로 평정된 20대 성인 7명이 참여하였다. 진동센서 장치에 사용된 도구는 Power aupply(OPE-3z05QI), Probe, Vibration Sensor(VS-BV203), DAQ(PSL-DAQ), Breadboard & Cable, PC, DAQ-device와 잡음이 많은 실제 신호를 정확하고 효율적으로 측정할 수 있는 형태로 변환시켜 주기 위해 PSL-DAQ를 함께 사용하였고 안면센서는 목, 코, 뺨 세 가지 부위에서 부착하여 측정하였다. 이와 동시에 공명발성, 입술트릴, 튜브발성 시 전기성문파형을 측정하기 위해 전기성문파형검사를 실시하였다. 후두의 양측 갑상연골(thyroid cartilage)에 전극(electrodes)을 부착시켜 성문접촉률(CQ), 성문개방율(OQ), 성문접촉속도율(SQ)을 측정하였다. 연구 결과, /ㅏ/ 발성은 목의 진동 크기가 3.36으로 가장 높았으나, 신체 부위간 유의한 차이는 없었다. 카주 발성도 목의 진동 크기가 0.3927로 가장 높았으나, 신체 부위간 유의한 차이는 없었다. 한편, 입술 트릴의 경우, 뺨의 진동 크기가 4.833으로 목이나 코에 비해 유의하게 높게 나타났다. 빨대 발성은 뺨의 진동이 1,937로 가장 높았으나, 통계적으로 유의한 차이는 없었다. 반면, 물저항발성은 뺨이 11.095로 목이나 코에 비해 유의하게 높게 나타났다. 한편, CQ, OQ, SQ는 모든 발성 유형 간 유의한 차이를 보이지 않았다. 본 연구를 통해 안면진동센서는 환자들에게 진동 감각에 대한 바이오피드백을 제공함으로써, 공명발성 훈련이나 반폐쇄성도훈련에 대한 효과를 정량화하고, 공명발성을 적절히 수행하는지에 대해 모니터링하는 데 도움을 줄 수 있을 것이다.

Acknowledgment : 본 연구는 한국연구재단 지원을 받아 이루어졌음(NRF-2020S1A5A2a0145868)

구두 발표 I: 음성공학

좌장: 박형민(서강대)

법과학적 활용을 위한 삼성 스마트폰 음성 녹음 파일 원본과 편집본의 메타데이터 구조 및 속성 분석

안 서 영[1], 유 세 희[1], 김 경 화[2], 홍 기 형[3]
[1]성신여자대학교 미래융합기술공학과, [2]대검찰청 법과학 분석과,
[3]성신여자대학교 서비스디자인공학과

Analysis of metadata structure and attributes of original and edited Samsung smartphone voice recording files for forensic use

SeoYeong Ahn[1], SeHui Ryu[1], KyungWha Kim[2], KiHyung Hong[3]
[1]Dept. of Future Convergence Technology Engineering, Sungshin W. University, [2]Forensic Science Division, Supreme Prosecutor's Office, Korea,
[3]Dept. of Service Design Engineering, Sungshin W. University
98_0731@honglab.org, rsh725@honglab.org, savoix@spo.go.kr, khhong@sungshin.ac.kr

스마트폰의 대중화로 인하여 근래 범죄의 증거자료로 제출되는 녹취 파일은 대부분 스마트폰을 통하여 생산되고 있으며, 스마트폰을 기반으로 한 녹음 파일의 무결성(위변조) 여부가 수사와 재판 과정에서 주요 쟁점으로 떠오르고 있다. 가장 높은 국내 시장 점유율을 가진 삼성 스마트폰은 통화 및 음성 녹음, 그리고 편집이 가능한 어플리케이션이 탑재되어 유통되고 있으며, 자체 어플리케이션을 통한 편집은 외부 어플리케이션을 통한 편집과 다르게 원본 파일과의 유사성이 높기에, 무결성을 입증하기 위해 더 정밀한 분석 기법 개발이 필요하다. 본 연구에서는 삼성 스마트폰 26개 기종에서 생성된 녹음 원본 및 편집 파일의 메타데이터 구조와 속성을 분석하여, 기기의 녹음 파일과 자체 제공 어플리케이션을 통한 편집 파일의 메타데이터 구조와 속성을 분석하였다. 연구 결과, 원본과 편집본 사이의 음성 파일 메타데이터 구조 및 속성값에서 유의미한 차이가 있음을 확인하였다.

장애발화 음성 코퍼스 구축 시 필요한 메타데이터 제안

이 선 우[1], 김 선 희[2], 정 민 화[1]
서울대학교 언어학과[1], 서울대학교 불어교육학과[2]

Metadata recommendations for disordered speech corpora

Seonwoo Lee[1], Sunhee Kim[2], Minhwa Chung[1]
Dept. of linguistics, Seoul National University[1],
Dept. of french education, Seoul National University[2]
lsw5220@snu.ac.kr, sunhkim@snu.ac.kr, mchung@snu.ac.kr

본 연구는 장애발화 음성 코퍼스의 상호운용성을 높이기 위해 코퍼스 구축 시 메타데이터로서 고려해야 할 환자 정보를 제안하고자 한다. TalkBank 코퍼스와 기구축된 장애발화 음성 코퍼스를 메타분석하여 다양한 장애군에서의 메타데이터를 일반적 정보와 의료적 정보로 나누어 분석한 뒤, 언어재활사 설문을 통하여 장애군별 우선순위를 확인하였다. 그 결과, 성별과 연령은 거의 모든 장애군에서 필수적인 일반적 정보였고, 의료적 정보에는 진단과 밀접하게 관련된 요인들이 포함되었다.

한국어 감정 인식 딥러닝 알고리즘 개발에 대한 연구

김 태 용, 이 보 원
인하대학교 전기컴퓨터공학과 DSP연구실

Deep Learning Experiments on Korean Speech Emotion Recognition

Taeyong Kim, Bowon Lee
Dept. of Electrical Computer Engineering, Inha University
taeyong.kim@dsp.inha.ac.kr, bowon.lee@dsp.inha.ac.kr

화자가 전달하고자 하는 의미를 정확하게 이해하기 위해서는 말을 이루고 있는 텍스트 정보와 그 안에 내포되어 있는 감정을 인식하는 것이 중요하다. 본 논문에서는 한국어 음성을 이용하여 인간의 감정을 인식하는 성능을 향상시키기 위해 연구를 수행하였다. 연구의 객관적인 타당성을 보여주기 위하여 한국지능정보사회진흥원에서 제공한 한국어 감정 데이터셋을 사용하여 연구하였다. 상기 데이터셋은 총 7가지 감정(Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise)으로 구성되어 있으며 각 감정에 대하여 30대 여성 성우의 감정별 3,000개 발화를 포함한 음성 녹음파일로 구성되어 있다. 총 21,000개의 음성파일에 대하여 입력 음성파일에 대한 감정을 인식하는 감정인식 알고리즘을 개발하였다. 상기 모델은 시계열로 구성된 음성데이터를 효율적으로 분석하기 위하여 LSTM으로 구성된 모델을 기반으로 하였으며, 모델의 과적합 문제를 해소하기 위하여 dropout기법도 적용하였다. 실험 결과에서 상기 모델을 통한 감정인식을 수행한 결과 7가지 감정분류 테스트에서 약 98% 정확도를 나타내었다.

본 논문은 2021년 대한민국 교육부와 한국연구재단의 지원 (NRF-2021S1A3A2A01087325) 및 산업통상자원부의 산업기술혁신 사업 (10073154, 인간 내면상태의 인식 및 이를 이용한 인간친화형 인간-로봇 상호작용 기술 개발) 의 지원을 받아 수행된 연구임

포스터 발표

좌장: 공은정(항공대)
이수복(우송대)
홍기형(성신여대)

뇌성마비 성인과 일반 성인 및 뇌성마비 아동 간의 모음 /아/ 음향학적 특성 비교

정 필 연, 심 현 섭
이화여자대학교 아동발달센터, 이화여자대학교 언어병리학과

Comparison of the acoustic characteristics of the vowel /a/ in adults and children with cerebral palsy and healthy adults

Pilyeon Jeong, Hyunsub Sim
Ewha Womans University Center for Child Development and Disability
Dept. of Communication disorders, Ewha Womans University
jpy@ewha.ac.kr, simhs@ewha.ac.kr

본 연구의 목적은 뇌성마비 성인의 음향학적 특성을 일반 성인 및 뇌성마비 아동과 비교해 보는 것이다. 피험자로는 뇌성마비 성인 4명(평균 29.25세), 일반성인 4명(평균 26.25세), 그리고 뇌성마비아동 4명(평균 6.07세)이 참여하였다. 뇌성마비 성인과 일반 성인 간에는 성별과 연령을 동일하게 통제하였고, 뇌성마비 성인과 뇌성마비 아동은 경직형과 혼합형(경직형과 불수의운동형)으로 유형을 일치시켜 배치하였다. 실험 과제는 모음 /아/ 연장발성이었고, 측정변수는 F0, Jitter, Shimmer, NHR, F1, F2, 음성강도(intensity)였다. 모음연장발성 과제는 총 3회 실시하였고, 가장 긴 음성샘플을 선택하여 분석하였다. 음향학적 분석을 위해 음성샘플 가운데 안정화 구간을 선택하여 Praat으로 분석하였다. 세 집단 간 통계적 유의성을 검정하기 위해 Kruskal-Wallis 분석을 실시하였고, 사후분석은 Mann-Whitney *U* 검정으로 분석하였다. 연구결과 세 집단 간에 MPT, Jitter, Shimmer, NHR, F2, 음성강도에서 유의한 차이가 있었다. 어떤 집단에서 차이가 있는지 사후분석을 실시한 결과, 뇌성마비 성인과 일반 성인 간에는 MPT, Jitter, Shimmer, NHR, F2에서 통계적으로 유의한 차이가 있는 것으로 나타났다(all p<.05). 뇌성마비 성인과 뇌성마비 아동 간에는 Jitter, Shimmer, NHR, 음성강도에서 유의한 차이를 보였다(all p<.05). 이러한 결과는 뇌성마비 성인은 발성 및 조음 시 일반 성인에 비해 음질과 호흡능력에서 어려움을 보이고, 혀의 전후 위치에서도 차이가 있으며 뇌성마비 아동에 비해서는 음질이 더 저하되어 있음을 시사한다.

청유문과 의문문 읽기 과제를 통한 자폐아동과 일반아동의 운율특성 비교

김 현 경, 성 철 재
충남대학교 언어병리학과, 충남대학교 언어학과

Prosodic Feature Comparison of Austistic Children and Ordinary Children Through Reading Task of Suggestive Sentences

Hyunkyung Kim, Cheoljae Seong
Dept. of Speech & Language Pathology, Linguistics, Chungnam National University
kim2209094@naver.com, chseong49@gmail.com

자폐아동은 사회적 이해력에 결함이 있기(Gray, 1995) 때문에 비전형적인 운율이 나타난다(McCann & Pappe, 2003). 이러한 부적절한 운율 능력은 사회적 단서를 잘못 이해하게 하고, 의사소통 의도를 이해하는데 어려움을 갖게 한다(이수정, 2013). 자폐아동의 부적절한 운율 능력에 관한 연구는 평서문과 의문문에서 다수 보고되었고, 자폐아동이 의문문의 문미에서 음도, 강도가 높고 음도 변화가 크며, 일반아동보다 과장된 억양을 사용한다는 것을 알 수 있다(이진형 & 성철재, 2019; 신희백, 2016; 정금수 & 성철재, 2007). 여러 문장 형태에 따른 운율특성을 비교분석할 필요성이 있으나, 자폐아동의 청유문 운율을 분석한 연구는 한정적이다. 때문에, 청유문과 의문문의 운율에서 어떤 음향학적 특성 차이가 있는지 알아보고자 하였다.

연구대상은 자폐아동 32명과 일반아동 32명으로 REVT 수용, 표현 결과와 성별을 일치시켜 구성하였다. 모든 대상자에게 2어절의 의문문과 청유문을 읽게 하여 녹음한 후, 문미 어절을 Praat으로 음향분석하였다. 그리고 문미 어절 및 음절의 음도, 강도, 상대적 길이(duration), 음도기울기(slope), 어절의 2개의 피치 정점(pitch peak) 간 음도 기울기(diff_s13_slope) 등 에 따라 차이가 있는지 SPSS for Window 22.0을 이용하여 이원분산분석을 실시하였다. 청지각 평가는 모든 아동이 발화한 문장 전체에 무작위로 선정한 문장을 추가하여 7명의 언어전문가(박사 및 박사과정중인 언어치료사)에게 들려주었다. 그 후, 청지각 평가를 통해 의문문 판별 여부와 의문문, 청유문의 정반응 반응분포, 집단 간 차이를 알아보고자 카이제곱 통계를 실시하였다.

통계 분석 결과 첫째, 의문문은 문미 어절 음도 범위가 자폐아동 집단이 일반아동 집단보다 작았다(일반집단>자폐집단). 청유문은 정규화된 문미 어절 음도 중앙값에서 자폐아동 집단이 일반아동 집단보다 크게 나타났다(일반집단<자폐집단). 일반아동 집단의 경우 문미 어절 음도범위에서 의문문이 청유문보다 넓었다(의문문>청유문). 자폐아동 집단의 정규화된 문미 어절 음도 중앙값에서 의문문이 청유문보다 작았고(청유문>의문문) 문미 어절 음도 범위는 의문문이 청유문보다 컸다(의문문>청유문). 둘째, 청유문에서 자폐아동 집단은 일반아동 집단에 비해 정규화된 문미강도 최고값, 정규화된 문미강도 평균값이 높으며, 문미강도 범위가 넓었다(일반집단<자폐집단). 자폐아동 집단은 의문문보다 청유문에서 문미강도 범위가 넓었다(의문문<청유문). 일반아동 집단은 청유문보다 의문문에서 정규화된 문미강도 최고값, 정규화된 문미강도 평균값이 높았다(의문문> 청유문). 셋째, 문미 음절 상대적 길이가 청유문에서 더 길었고, 자폐아동 집단이 일반아동 집단보다 짧게 발화하였다. 넷째, 자폐아동 집단은 일반아동 집단과 마찬가지로 문미 음절 음도 회귀기울기와 문미 어절 음도 회귀기울기에서 의문문이 청유문보다 높았다. 하지만 의문문의 문미 음절 음도

회귀기울기와 문미 어절 음도 회귀기울기정도는 자폐아동 집단이 일반아동 집단에 비해 더 낮았다. 청유문의 두 최고 음도 변화량기울기(diff_s13_slope)는 집단 간에 유의한 차이를 보였으나(p=.026), 문장 간에는 유의한 차이가 없었다(p=.595). 아래 청유문의 두 최고 음도 변화량기울기는 자폐아동 집단이 일반아동 집단보다 더 컸다(일반집단< 자폐집단). 다섯째, 청지각 평가에서 청자는 자폐아동 집단보다 일반아동 집단의 발화를 더 청유문과 의문문으로 인식하였다. 또한 청자는 자폐아동의 의문문보다 청유문을 더 바르게 인식하지 못하였다.

위의 결과, 의문문에서 자폐아동 집단은 문미의 운율이 약해서 음도 기울기와 음도가 일반아동 보다 낮았고, 청자가 자폐아동의 의문문을 구별하는데 어려움이 있었다. 그러므로 보다 확실히 의문문임을 알 수 있도록 문미 음절의 음도 기울기와 음도를 높여서 질문하는 연습이 필요하다. 청유문에서 자폐아동은 문미 어절과 음절을 짧게 발화하고, 문미의 음도와 강도가 높았는데, 의문문과 청유문의 구별하기 어려울 만큼 과장된 억양을 사용하는 경향이 있었다. 또한 청유문의 전형적인 운율특징인 문미 어절 두 최고음도 사이의 변화량 기울기는 자폐아동 집단이 일반아동 집단보다 높았다. 때문에 청유하는 말의 경우 문장의 끝을 더 길게 연장하여 발화하고 문미 어절의 첫 번째 최고 음도를 높이거나 두 번째 최고 음도를 낮게 발화하는 연습이 필요함을 알 수 있다. 이러한 운율 특성 때문에 청자가 자폐아동이 발화한 청유문을 청유문으로 바르게 지각하지 못하는 문제가 나타나며, 이를 통해 자폐아동의 운율 능력에 결함이 있음을 확인할 수 있었다. 자폐아동의 운율 능력 향상을 위해서 문미의 음도, 강도, 발화 시간 비율, 음도 기울기 등의 중재방법과 치료방안들을 모색해야 할 필요성이 있다는 것을 알 수 있다.

참고문헌

[1] 신희백, 최지은, 이윤경(2016). 문장유형에 따른 고기능 자폐스펙트럼 장애 아동의 운율 특성. 한국음성학회지, 8(2), 65-71.

[2] 이진형 & 성철재(2019). 따라말하기 과제를 통한 자폐아동과 일반아동의 평서문과 의문문의 음향학적 특성 비교. 한국음성학회, 12(2), 39-49.

[3] 정금수 & 성철재(2007). 자폐 범주성 장애아동과 정상아동의 문장 읽기에서의 운율특성 비교. 언어청각장애 연구, 12(4), 625-642.

[4] Gray, (1995). Teaching children with autism to 'read' social situations. In K. A. Quill(ED.), Teaching children with autism: Strategies to enbance communication and socilizaion, 219-241.

[5] McCann & Pappe (2003). Prosody in autism spectrum disorder: a critical review. Jaunal of Language Communication Disorder, 38(4), 325-350

성대결절 아동의 연속발화에 대한 공기역학적 특성

정 혜 주, 최 성 희, 이 경 재, 최 철 희
대구가톨릭대학교 언어청각치료학과

Aerodynamic characteristics of connected speech in children with vocal nodules

Hyeju Jung, Seong Hee Choi, Kyoungjae Lee, Chul-Hee Choi
Dept. of Audiology & Speech Language Pathology, Daegu Catholic University
hyeju1510@gmail.com, shgrace@cu.ac.kr, kjlee0119@cu.ac.kr, cchoi@cu.ac.kr

일반적으로 아동의 생명을 위협하는 것은 아니지만 그들의 삶에 부정적인 영향을 미치기 때문에 적절한 검사를 통하여 최적의 중재를 제공하는 것이 중요하다. 그러나 성대결절이 있는 아동을 대상으로 한 공기역학적인 검사는 매우 제한적이며 특히, 연속발화를 사용한 연구는 수행되지 않았다.

따라서 본 연구는 성대결절 아동과 일반 아동의 공기역학적 특성을 비교하고자 하였다. 또한, 말하기 과제가 공기역학적 특성에 미치는 영향을 알아보기 위해 모음 /ㅏ/ 연장발성 뿐만 아니라 말하기와 노래하기 과제를 포함한 연속발화를 이용하여 공기역학적 측정치를 비교하였다.

본 연구에서는 4세에서 10세 사이의 총 20명(10명 : 성대결절 아동, 10명 : 일반 아동)의 남자 아동이 참여하였으며 평균 연령은 6세 4개월이다. 공기역학적 평가를 위하여 Phonatory Aerodynamic System(Model 6600; KayPENDAX, Montvale, NJ, USA)을 사용하였다. 말과제로는 모음 /ㅏ/ 연장발성과 연결발화 과제(말하기, 노래하기)로는 “생일 축하합니다. 생일 축하합니다. 사랑하는 엄마의 생일 축하합니다.” 문장을 사용하였다. 모음 /ㅏ/ 연장발성의 공기역학적 분석은 Maximum Sustained Phonation을 사용하였고 연속발화(말하기, 노래하기)는 Running Speech를 사용하였다. 그 결과, 모음 /ㅏ/ 연장발성 시 성대결절 아동보다 일반 아동이 통계적으로 유의하게 더 긴 최대발성지속시간이 관찰되었다. 이는 성대결절은 대부분 양측성으로 발성 시 모래시계형 모양의 성대접촉패턴을 보이는데 이러한 성대접촉패턴으로 인해 발성 중 성문을 통과하는 호기량이 증가하면서 성대결절 아동이 일반 아동보다 최대연장발성지속시간이 짧아지고 청지각적으로 기식적인 음성을 산출하는 것으로 사료된다. 말하기 과제 동안에는 성대결절 아동이 일반 아동보다 통계적으로 유의하게 더 높은 최고흡기류율(PIF)과 흡기량(IVC)을 보였다. 이는 성대결절로 인한 불완전한 성대의 접촉으로 발성 시 호기가 더 빠져 나가므로 이로 인해 더 빈번한 흡기를 보이는 것으로 사료된다. 노래 과제 동안에는 성대결절아동이 일반 아동보다 통계적으로 유의하게 높은 최고호기류율(PEF)을 보였다. 이는 성대결절 아동이 일반 아동보다 노래하기 과제 시 더 많은 양의 호기류율을 배출한다는 것을 시사하며, 성대결절 아동의 음성 특징 중 기식적인 음성을 보이는 것을 뒷받침할 수 있다. 따라서, 성대결절 아동의 공기역학적 평가에 있어 아동의 호흡능력을 더 잘 이해하기 위해서는 검사 시 사용하는 말과제로 모음연장발성과 함께 연속발화를 이용한 공기역학적 평가를 수행할 필요가 있다. 또한, 공기역학적인 검사에 있어 성대결절과 같은 과기능성 음성장애를 지닌 아동뿐만 아니라 일반 아동들도 말하기 과제보다 노래하기 과제에서 공기역학적 지표들이 더 효율적이었으므로 노래하기 치료기법이 아동의 호흡을 개선시키는데 도움이 될 수 있음을 시사하였다.

언어재활사의 주관적·객관적 음성피로도

전 혜 원, 성 철 재
충남대학교 언어병리학과, 충남대학교 언어학과

The subjective·objective voice fatigue of Speech-Language Pathologist

Hyewon Jeon, Cheoljae Seong
Dept. of Speech & Language Pathology, Linguistics, Chungnam National University
jhw8245@hanmail.net, cjseong49@gmail.com

언어재활사는 직업의 특성상 음성피로에 노출되기 쉽고, 음성장애를 느끼는 주관적 음성 증상 유병률이 다른 직업군에 비해 높다. 본 연구에서는 언어재활사의 주관적·객관적 음성피로도에 대해 알아보고자 하였다. 연구대상은 대전·충남지역 사설 치료실에 근무하고 있는 20-30대 여성 언어재활사였으며, 한국판 음성피로도검사(Korean Vocal Fatigue Index, K-VFI; 강영애 외, 2017)를 포함한 설문조사와 함께 /아/모음 연장 발성과 문단 읽기 과제를 실시하여 음성샘플을 수집하였다. 수집된 음성샘플은 Praat을 사용하여 공명 관련 음향변수, 음성 관련 음향변수, 사운드 스펙트럼 관련 음향변수를 분석하였다. 청지각 평가는 현직 언어재활사 7명이 Praat ExperimentMFC를 이용하여 음성피로의 정도를 1점(낮음)부터 6점(높음)까지 6점척도로 평가하였다.

K-VFI에서 음성피로 및 음성피로로 인한 음성 사용 회피 항목과 음성 사용으로 인한 신체적 피로 항목에서 주관적 음성피로 유무로 나누어진 두 집단 간의 유의미한 차이가 있었다(p<0.01). 그러나 음향학적 특성에서는 주관적 음성피로 유무로 나누어진 두 집단 간의 유의미한 차이가 나타나지 않았고, 각 집단 내에서도 음향학적 특성의 차이가 없었다.

청지각 평가로 나누어진 음성피로 집단과 음성피로 없는 집단의 몇 가지 변수에서 집단 간 유의미한 차이가 있었다. 집단 간 유의미한 차이를 보인 음향변수들을 아래 <표1>에 제시하였다. 그러나 Pearson 상관관계 분석 결과 주관적 음성피로 집단과 청지각적 평가로 나누어진 집단 간의 상관관계는 유의하지 않았다(r=-.160, p=.112).

<표1>청지각 평가로 나누어진 음성피로 집단과 음성피로 없는 집단의 공명·음성·스펙트럼 관련 음향변수에 대한 이원분산분석(Two-way ANOVA) 결과

변수	주효과	df	F	변수	주효과	df	F
F1_dB	집단	1	8.707**	shim_ddq	집단	1	18.600***
mean_energy	집단	1	8.818**	HNR	집단	1	35.963***
jit_local	집단	1	26.490***	NHR	집단	1	25.400***
jit_rap	집단	1	25.102***	cepsPeak	집단	1	16.316***
jit_ppq5	집단	1	32.007***	cpp	집단	1	11.414**
jit_ddq	집단	1	25.114***	rnr	집단	1	8.198**
shim_local	집단	1	19.004***	centroid	집단	1	4.074*

shim_apq3	집단	1	18.601***	tilt	집단	1	7.878**
shim_apq5	집단	1	16.759***	mean_energy	집단	1	9.683**
shim_apq11	집단	1	19.916***				

*p<0.05, **p<0.01, ***p<0.001

본 연구에서는 설문조사 및 음향분석을 통해 언어재활사의 주관적·객관적 음성피로도에 대해 알아보고자 하였다. 연구결과 K-VFI는 개인의 주관적 음성피로도를 반영해 줄 수 있는 검사 도구로 보인다. 그러나 주관적 음성피로도와 음향분석 결과가 일치하지 않았고, 주관적 음성피로 유무로 나누어진 집단과 청지각평가로 나누어진 집단 간의 상관관계도 유의하지 않았다. 이러한 음향분석과 청지각평가의 결과는 대상자의 음성 특성, 즉 '음질'과 관련이 있다고 볼 수 있다. 결과적으로 음성피로감과 같은 주관적 문제를 객관적 데이터로 살펴보는 것에 한계가 있으므로 개인의 음성피로도 평가를 위해서는 다차원적인 평가 방법이 필요함을 시사한다.

참고문헌

[1] 강영애, 장재원, 구본석(2017). 음성장애환자 대상 음성피로와 음성평가 간 상관 및 음성피로도 설문(Voice Fatigue Index)의 임상적용. 대한이비인후과학회지 두경부외과학, 60(5), 232-242.

[2] 김지성, 최성희(2018). 한국 언어재활사의 음성피로 및 음성과 관련된 삶의 질. Communication Sciences and Disorders, 23(4), 1078-1090.

[3] 송윤경, 표화영(2010). 언어치료사의 음성증상 및 한국어판 음성장애지수에 대한 예비연구. 말소리와 음성과학, 2(2), 123-133.

[4] 이은정(2015). 직업적 음성사용자의 음성증상 및 '음성건강' 관련 서비스 인지도 조사. 디지털융복합연구, 13(1), 397-405.

Usefulness of Vocal Fatigue Index for evaluation of laryngeal hypertension

Ji Sung Kim, Dong Wook Lee
(Dept. of Otorhinolaryngology, ChungBuk National University Hospital,
slp2046@naver.com, dwlee@chungbuk.ac.kr

This study compares Vocal Fatigue Index(VFI) scores according to the presence or absence of external laryngeal tension in Hyperfunctional voice disorder. And through this, it is to confirm the usefulness of VFI to hypertension of Extrinsic laryngeal muscles. The subjects were 61 women diagnosed with Hyperfunctional voice disorder (hypertension group 42, none hypertension group 19). The author palpated Extrinsic laryngeal muscles for evaluation of hypertension and classified them as the presence or absence. The voice measurements were jitter, shimmer, Korean-Voice Handicap Index-10(K-VHI-10), and Korean-Vocal Fatigue Index(K-VFI). The voice compared were according to the diagnosis and presence of hypertension only for patients with Hyperfunctional voice disorder. As a result of comparing the voice measurement according to the presence or absence of hypertension, there was no significant difference in the acoustic variables, K-VHI-10 and K-VFI-total, K-VFI-Fatigue. Whereas, K-VFI-Physical(p=.006) and K-VFI-Rest(p=.022) were significantly higher in the Hypertension group. These results indicate that the hypertension group has more physical discomfort and less voice recovery than the group without hypertension. It means that K-VFI can measure the physical discomfort and limitations of voice recovery due to hypertension of the external laryngeal muscle. The VFI can be used as one of the methods to evaluate the hypertension of the external laryngeal muscle in Hyperfunctional voice disorder.

구어처리과제의 발달: 문장폭기억과제를 중심으로

오 소 정
동명대학교 언어치료청각학과

The Development of Performances on Language Processing Tasks: In Case of Competing Language Processing Task

Sojung Oh
Dept. of Speech-Language Pathology ana Audiology, Tongmyong University
sjoh@tu.ac.kr

해외에서 문장폭기억과제(CLPT:Competing Language Processing Task)는 다양한 문화, 언어적 배경을 가진 아동들의 언어학습능력을 평가하는 과제로 자주 사용되어 왔다. 언어처리과제로 널리 사용되는 비단어따라말하기 과제보다 기억을 유지하면서 정오를 판단하는 두 가지 과제를 동시에 수행하여야 하는 과제의 특성상 좀더 명확하게 작업기억을 평가할 수 있는 과제로 인식되어 왔다.

본 연구의 목적은 아동기, 청소년기, 성인기 문장폭기억과제 수행의 발달 양상을 살펴봄으로써 언어학습 잠재력을 확인할 수 있는 평가도구로서의 잠재력을 확인하는 것이다.

언어문제를 경험한 적이 없고 한국어 단일언어를 사용하는 아동부터 성인까지(7세~22세)의 5개 집단 122명이 컴퓨터화된 문장폭기억과제(CLPT)에 참여하였다. 연령집단은 두 학령기 아동 집단(7~9세/10~12세), 두 청소년 집단(13~15세/16~19세), 그리고 성인 집단(20~22세)으로 구성되었다.

컴퓨터화된 문장폭기억과제(CLPT)는 일련의 문장세트에서 각 문장을 듣고 정오를 판단함과 동시에 세트 종료 후 각 문장의 마지막 어절을 기억하는 과제였다. 1~ 6개 문장으로 구성된 6개 세트로 구성되어 기억해야하는 단어의 수가 1개에서 최대 6개까지 증가하도록 구성되어 있었으며, 이전 연구에서 연구자가 개발한 것을 사용하였다.

연령에 의한 주효과가 있는지 SPSS ver.18을 이용하여 일원분산분석(one-way ANOVA)을 실시하고, 어떤 집단에서 연령 간 차이가 나타났는지 LSD 사후검정을 실시하였다.

기술적 통계 결과 문장폭기억과제(CLPT) 낱말 회상은 아동기부터 성인기까지 점진적 발달 양상을 보였으며, 성인들도 100% 수행에 도달하지 못하고 50~60%의 수행 정확도에 머무르는 것으로 나타났다. 일원분산분석 결과, 문장폭기억과제(CLPT) 낱말회상점수는 연령증가에 따라 유의하게 증가하는 것으로 나타났으며($F_{(4,\ 121)}$ = 8.055, p<.001), LSD 사후검정을 실시한 결과 낱말회상점수의 연령집단 간 차이는 아동기 초기인 초등 저학년 아동과 초등 고학년, 중학생, 고등학생, 성인 집단 간 차이가 모두 유의하였다. 나머지 집단 간에는 유의한 차이가 나타나지 않았다.

본 연구를 통하여 문장폭기억과제 수행 정확도가 연령증가에 따라 향상됨을 알 수 있었고, 성인기까지도 100% 정확도에 도달하지는 못하는 것으로 나타났다. 구체적으로 학령기에 저학년 시기를 지나 고학년이 되면서 뚜렷한 수행 향상을 보이고, 이후 청소년기와 성인기에는 큰 변화를 보이지 않는 것으로 나타났다. 즉, 학령기 전기에서 후기 간 차이가 전반적인 연령에 따른 차이를 가져왔음을 확인할 수 있었다. 이런 양상을 고려할 때 후속 연구를 통해 학령전기 아동을 포함하여 문장폭기억과제(CLPT) 수행이 학령전기 아동 어느 시점부터 가능하고 이후 학령기가 되기까지 어떠한 발달 양상을 보이는지를 밝힌다면 평가도구로 유용성을 확인하면서 적용 가능한 대상 연령대를 짐작할 수 있을 것으로 판단된다.

1차원 합성곱 신경망 구글넷 기반 스테레오 음성의 도래 방향 추정

이 정 혁, 김 홍 국
광주과학기술원 전기전자컴퓨터공학부

Estimation of Direction of Arrival of Dual-Microphone Speech Based on 1D-CNN GoogLeNet

Junghyuk Lee, Hong Kook Kim
School of Electrical Engineering and Computer Science, GIST
ljh0412@gist.ac.kr, hongkook@gist.ac.kr

음성의 도래방향은 주로 스테레오 음원의 교차상관도를 통해 계측된다. 그러나, 낮은 SNR의 잡음 환경에서는 교차상관도 상에 잡음의 위상이 크게 추정되며, 잡음제거 알고리즘 적용 시, 교차상관도로 측정된 음성의 위상이 훼손될 수 있다. 또한 마이크 사이 간격에 따라 특정 각도를 추정하기 어려울 수 있다. 본 연구에서는 위와 같은 잡음 제거된 스테레오 음원의 위치 추정을 정확하게 수행하기 위해 교차상관도 및 좌/우 음성 에너지를 입력, 음성 도래방향 분류를 출력으로 하는 데이터 셋을 구성하였으며, 각도별 특징을 정확하게 추출하기 위해, 다양한 필터 크기를 갖는 1차원 합성곱 신경망으로 재구성된 구글넷 모델을 사용하였다. 제안된 모델을 훈련 및 평가 하기 위해 재난 상황에서 드론을 통한 구조작업을 가정한 구조자 음성 및 드론 잡음으로 구성된 데이터베이스를 구성하였으며, 33,726개 테스트 음원에 대해, clean 환경 음원의 경우 99.85%, U-net으로 잡음 제거된 noisy 환경 음원의 경우 94.93% 도래 방향 추정 정확도를 보였다.

Acknowledgment

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2021-0-00014, 재난상황 대응을 위한 엣지컴퓨팅 기반 시청각 인지지능 솔루션 개발)

상호의존정보를 이용한 임베딩 기반 음성감정인식

박 순 찬, 김 형 순
부산대학교 전자공학과

Embedding-Based Speech Emotion Recognition Using Mutual Information

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

음성감정인식을 위한 훈련 데이터는 그 특성상 충분히 수집되기 어려우며, 제한된 화자의 발화로 훈련된 음성감정인식 시스템은 훈련 데이터의 일부 화자 특성에 대한 편향이 발생할 수 있다. 본 연구에서는 상호의존정보(mutual information)의 최적화를 통해 음성으로부터 일반, 화자, 감정 특성을 분리하여 각각에 대한 임베딩(embedding)을 추출하고, 이들을 기반으로 음성감정인식 수행하여 그 성능을 개선하였다. 임베딩 추출을 위해 트랜스포머 인코더에 음성 특징과 함께 일반, 화자, 감정 특징 추출을 위한 특수 토큰을 입력으로 하고, 각 토큰에 해당하는 네트워크 출력을 각각에 대한 임베딩 벡터로 사용하였다. 전체 훈련 과정은 상호의존정보 추정을 위한 Mutual Information Neural Estimator(MINE)의 훈련과 임베딩 추출 및 감정 분류 모델의 훈련을 번갈아 가며 수행한다. 임베딩 추출 및 감정 분류 훈련 과정에는 화자 및 감정 분류를 위한 교차 엔트로피와 함께 각각의 임베딩 사이 상호의존정보를 최소화하는 다중작업훈련(multi-task learning)이 적용되어, 각각의 임베딩이 상호 의존적인 정보를 포함하지 않도록 한다. 이때 화자 분류기의 입력은 일반 임베딩과 화자 임베딩, 감정 분류기의 입력은 일반 임베딩과 감정 임베딩을 사용하여 각각의 임베딩이 상호 의존적이지 않으면서 의도한 특성을 추출하도록 유도한다. 효과적인 화자 임베딩 추출을 위해 먼저 VoxCeleb2 데이터를 통해 화자분류에 대한 사전훈련을 수행하고, IEMOCAP 데이터의 기쁨, 슬픔, 화남, 중립의 네 가지 감정 범주의 음성으로 미세조정 및 감정인식 성능을 평가하였다. 상호의존정보를 이용한 다중작업훈련을 적용한 경우 단순 감정 분류 훈련만 적용한 경우 대비 가중정확도가 2.2%, 비가중정확도가 3.1% 개선되는 것을 확인할 수 있었다.

감사의 글

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

도메인 적대적 학습 기반의 억양 음성 인식

나 형 주[1], 장 민 지[2], 나 희 정[2], 박 정 식[1*]
[1]한국외국어대학교 언어공학연구소, [2]한국외국어대학교 영어학과

Accented Speech Recognition based on Domain Adversarial Training

Hyeongju Na[1], Minjee Jang[2], Heejung Na[2], Jeong-Sik Park[1*]
[1]Language Technology Research Institute, Hankuk University of Foreign Studies
[2]Dept. of English Linguistics, Hankuk University of Foreign Studies
skgudwn34@gmail.com, jmj3047@naver.com, 9203l3@naver.com, *parkjs@hufs.ac.kr

심층신경망 기술의 발전과 함께 음성인식 성능이 크게 향상되었지만, 억양이 섞인 발화에 대해서는 성능이 저하되는 문제가 발생한다. 억양이 섞인 발화는 표준 발화와 비교했을 때 언어학적인 차이가 존재하는데, 이러한 차이가 성능 저하의 요인이 된다. 따라서 본 연구에서는 도메인 적대적 신경망 기법을 사용하여 두 발화 간의 차이를 줄이고, 종단 간 기법을 사용하여 음성인식의 과정을 간소화하였다.

도메인 적대적 신경망은 소스 도메인과 타겟 도메인이 적대적으로 학습이 되면서 두 도메인 간의 분포 차이를 줄이는 것을 목적으로 한다. 도메인 적대적 신경망은 특징 추출기, 도메인 분류기, 레이블 예측기 총 3개의 부분망으로 구성되는데, 특징 추출기에는 합성곱 신경망을, 도메인 분류기에는 심층 신경망을, 그리고 레이블 예측기에는 양방향 게이트 순환 유닛을 이용하여 도메인 적대적 신경망을 구성하였다. 또한, 레이블을 예측할 때 종단 간 기법을 활용하여 입력 데이터를 사전 분할하지 않고, 레이블 예측 이후의 후처리 작업을 없애면서 음성인식 과정을 간소화하였다.

본 연구에서 제안한 도메인 적대적 학습 기반의 억양음성인식 성능을 입증하기 위하여 Baseline 모델과 DANN 모델을 만들어 실험을 진행하였다. 실험 데이터로는 Common Voice 코퍼스의 영어 데이터셋을 사용하였고, 그중에서 미국, 호주, 캐나다, 잉글랜드, 인도 억양의 데이터를 실험에 사용하였으며, 미국 억양을 소스 도메인으로 나머지 네 개의 억양을 타겟 도메인으로 정하였다.

실험 결과 호주, 캐나다, 잉글랜드, 인도 억양 모두에서 DANN 모델의 성능이 Baseline 모델보다 높은 성능을 보였다. 하지만 억양에 따라 성능 개선의 차이가 있었으며, 캐나다 억양에 비해 잉글랜드 억양과 인도 억양에서 성능이 눈에 띄게 향상되었다. 이 같은 결과는 잉글랜드 억양과 인도 억양이 소스 도메인으로 사용된 미국 억양 데이터와 언어학적으로 큰 차이가 존재하여 Baseline 모델에서는 성능이 낮았으나, 도메인 적대적 학습 기반의 DANN 모델이 타겟 억양의 특성을 반영함으로써 성능이 크게 개선된 것으로 분석된다. 따라서 도메인 적대적 신경망은 소스 도메인과 타겟 도메인 사이의 분포의 차이를 줄임으로서 억양음성인식의 성능을 향상시킬 수 있음이 확인되었다.

* 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.2020R1A2C1013162)

L2 화자의 유창성 수준에 따른 짧은 발화 음성 대상 언어 식별

나 희 정[1], 나 형 주[2], 장 민 지[1], 박 정 식[2*]
[1]한국외국어대학교 영어학과, [2]한국외국어대학교 언어공학연구소

Language Identification on Short Utterance Speech according to Fluency Level of L2 Speaker

Heejung Na[1], Hyeongju Na[2], Minjee Jang[1], Jeong-Sik Park[2*]
[1]Dept. of English Linguistics, Hankuk University of Foreign Studies
[2]Language Technology Research Institute, Hankuk University of Foreign Studies
9203l3@naver.com, skgudwn34@gmail.com, jmj3047@naver.com, *parkjs@hufs.ac.kr

발화 내에서 언어가 교체되는 현상을 '코드 스위칭(code-switching)'이라 한다. 일반적인 언어식별과 달리 코드 스위칭 발화를 대상으로 하는 언어식별은 두 언어, 즉 모국어와 제 2언어(L2)를 동일 화자의 발화 내에서 구별해야 하므로 언어식별 모델이 화자의 발화 특징을 간과하면서 언어의 차이를 식별할 수 있어야 한다. 또한, 코드 스위칭 발화에서는 단일 언어 세그먼트의 길이가 비교적 짧기 때문에 언어식별 성능이 저하될 수 있다. 제 2언어 발화의 경우, 화자별로 유창성 측면에서 차이를 나타내며, 모국어 억양이 심하게 반영되어 유창성이 낮은 화자 음성의 경우, 모델은 각 언어 특성의 차이를 반영하지 못하기 때문에 제 2언어와 모국어를 구별하는 것이 어려워진다.

본 연구에서는 한국인 화자가 발성한 한국어-영어 코드 스위칭 발화의 언어식별을 위한 기초연구로서 짧은 발화 음성 대상의 언어식별 시스템을 구축하고 한국인 L2 화자의 영어 유창성 수준에 따른 식별 성능의 변화를 확인하였다. 언어식별을 위한 모델은 CNN을 기반으로 하였으며, ETRI 데이터를 활용하였다. 한국인 화자 50명의 영어 유창성 수준을 원어민과 유사한 수준부터 영어 음소를 모두 한국어 음소로 대체하여 발화하는 수준까지 총 4가지 레벨로 분류하고 각각의 언어식별 성능을 관찰하였다. 또한 짧은 발화 길이를 3초, 1초, 0.5초, 0.3초로 세분화하여 실험을 진행하였다. 실험 결과, 유창성 수준 및 발화의 길이에 따라 언어식별 성능이 큰 차이를 보였다. 향후 본 연구 결과를 이용하여 유창성 및 발화길이와 관계없이 자동으로 한국어-영어 코드 스위칭 구간을 탐지하는 연구를 수행할 계획이다.

* 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.2020R1A2C1013162)

다자간 대화 음성인식 시스템 개발

박 전 규, 강 점 자, 동 성 희, 박 기 영, 오 유 리, 이 성 주, 최 우 용
한국전자통신연구원 복합지능연구실

Development of the Multi-speaker Dialogue Speech Recognition System

Jeon Gue Park, Jeom Ja Kang, Sunghee Dong, Yoo Rhee Oh, Kiyoung Park,
Sung Joo Lee, Woo Yong Choi
Integrated Intelligence research Section, ETRI
{jgp jjkang dsh7560 pkyoung yroh lee1862 wychoi4}@etri.re.kr

다자가 참석하는 회의 및 상담 등을 위한 음성인식에 있어서는 일반적으로 개인별로 할당된 마이크로폰 또는 다수의 화자가 공유하는 어레이 마이크로폰을 사용하게 되며, 환경적으로는 다양한 잡음, 반향, 발성 겹침이 존재하고 비정형 구어체 발성 특성이 빈발하는 특성이 있다. 여기에 더하여 근거리 및 원거리 발성을 동시에 고려해야 하는 난이도가 높은 음성인식 태스크가 된다. 이러한 배경에서 본 논문에서는 화자간 발성겹침을 위해 Conv-Tasnet에 기반하는 음원분리 기술, 각 화자의 발성 구간을 시간영역에서 분리하는 UIS-RNN 기반의 화자분리 기술, 입력 음성에 대해 온라인 실시간으로 인식을 수행하는 트랜스포머 기반의 종단형 스트리밍 음성인식 기술, 화자별로 구분되어 인식된 단위 문장에 대한 멀티모달 감정인식 기술 등 다양한 기술 요소로써 통합하는 회의 환경 음성인식 시스템을 제안하고 있다.

[감사의 글] 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2019-0-01376, 다중 화자간 대화 음성인식 기술개발)

Graph Attention Networks를 활용한 프레임 단위 화자 특징 결합

허 정 우[1], 심 혜 진[1], 박 재 한[2], 이 가 희[2], 유 하 진[1]
[1]서울시립대학교 컴퓨터과학부, [2]KT Corperation

Frame-Level Speaker Feature Aggregation Using Graph Attention Networks

Jungwoo Heo[1], Hye-jin Shim[1], Jae-han Park[2], Ga-hui Lee[2], Ha-Jin Yu[1]
[1]School of Computer Science, University of Seoul, [2]KT Corperation
jungwoo4021@gmail.com, shimhz6.6@gmail.com, park.jaehan@kt.com, gahui.lee@kt.com, hjyu@uos.ac.kr

화자 인증은 발성을 비밀번호로 활용하는 사용자 인증 과제로, 발성에 내포된 화자 특징을 벡터로 사상하여 이들 간의 비교를 통해 수행될 수 있다. 일반적인 화자 인증 시스템은 발성의 프레임별 화자 특징을 심층 신경망을 활용하여 추출한 뒤, 이들을 풀링 기법을 통해 하나의 벡터로 사상한다. 효과적인 화자 인증 시스템을 구축하기 위해서는 프레임 단위의 화자 특징들을 효과적으로 종합할 필요가 있다. 본 연구에서는 프레임 단위 화자 특징을 하나의 발성 단위 화자 특징으로 효과적으로 종합하기 위해 Graph Attention Networks (GAT) [1]를 활용하였다. GAT는 그래프 구조의 데이터를 처리하기 위해 고안된 심층 신경망으로, 각 노드 쌍에 대해 서로 다른 가중치 (attention score)를 부여하는 방식을 통해 노드 간의 관계 및 중요도를 모델링 할 수 있는 것으로 알려져 있다. 기존 화자 인증 연구에서의 풀링 기법은 전체 프레임 단위 화자 특징 중 중요한 특징을 선택하거나, 평균·분산 등의 통계값을 종합된 발성 단위 화자 특징으로 사용하였는데, 이러한 방식은 프레임 단위 화자 특징 간의 관계를 모델링하고 분석하고 활용하기 어려울 수 있다. 따라서 GAT를 화자 인증에 적용하여, 각 프레임 단위 화자 특징의 관계를 모델링하여 활용하는 방식을 통해 개별 프레임의 특성만을 고려하는 것 보다 발성 단위 화자 특징 종합에 도움이 될 수 있다. 이를 위해 특징 지도를 시간 축으로 분할하여 노드로 간주하고, 이들을 전결합한 그래프를 정의하여 GAT의 입력으로 사용하였다.

실험에는 발성의 원시파형을 입력으로 사용하는 RawNet2 [2]를 베이스라인으로 사용하였다. 이는 합성곱 연산을 활용하여 원시 발성으로부터 프레임 단위 화자 특징을 추출한 후, Gated Recurrent Units (GRU) 계층을 사용하여 프레임 단위 화자 특징들을 하나의 발성 단위 화자 특징으로 종합한다. 본 논문에서는 GAT를 활용한 프레임 단위 화자 특징 종합의 효과를 확인하기 위해 RawNet2의 GRU 계층을 GAT로 대체하여 성능 변화를 비교하였다. 실험에 사용한 심층 신경망은 VoxCeleb2 데이터 세트를 사용하여 학습되었고, 학습된 신경망의 평가는 VoxCeleb1 공식 트라이얼을 사용하여 진행되었다. 실험 결과, GRU를 사용하여 프레임 단위 화자 특징을 혼합하는 경우 동일 오류율이 2.48%이고, GAT를 사용하여 프레임 단위 화자 특징을 혼합하는 경우 동일 오류율은 2.23%임을 확인하였다. 즉, 제안한 시스템의 성능은 베이스라인 대비 약 10% 개선되었음을 확인하였다.

참고문헌

[1] P. Veli˘ckovi´c, G. Cucurull, A. Casanova et al., “Graph attentionnetworks,”arXiv preprint arXiv:1710.10903, 2017.

[2] J.-w. Jung, S.-b. Kim, H.-j. Shim et al., “Improved rawnet withfeature map scaling for text-independent speaker verificationusing raw waveforms,”in Proc. Interspeech, 2020

Transformer 및 Parallel WaveGAN을 이용한 다화자 음성합성

최 연 주, 엄 지 섭, 김 회 린
카이스트 전기및전자공학부

Multi-Speaker Text-to-Speech Using Transformer and Parallel WaveGAN

Yeunju Choi, Jisub Um, Hoirin Kim
School of Electrical Engineering, KAIST
wkadldppdy@kaist.ac.kr, twiz0311@kaist.ac.kr, hoirkim@kaist.ac.kr

본 연구에서는 Transformer 및 Parallel WaveGAN을 이용한 다화자 음성합성 시스템을 제안한다. 딥러닝의 발전과 함께 음성합성 분야에서 진취적인 성과들이 나오면서, 다양한 화자의 음색으로 음성합성이 가능한 다화자 음성합성 시스템의 수요도 증가하고 있다. 그에 따라, 우수한 음질로 현재 가장 널리 알려져 있는 Transformer 기반 음성합성 모델과 뉴럴 보코더 중 state-of-the-art 모델인 Parallel WaveGAN을 결합하여, 다양한 화자의 목소리로 거의 실시간으로 음성 합성을 할 수 있는 시스템을 구현하였다.

어텐션 메커니즘 기반의 seq2seq 모델에서 특정 구간이 생략 혹은 반복되는 고질적인 alignment 문제가 다화자 음성합성의 경우에 다양한 alignment를 모델링하면서 더욱 심화된다. 이를 직접적으로 해결하고자, 훈련 단계에서 guided attention loss를 적용하였고, 합성 단계에서는 디코더-인코더 어텐션 메커니즘을 통해 alignment를 구할 때 각 디코더 time step마다 볼 수 있는 입력 텍스트의 구간을 제한하였다. 화자 정보를 나타내는 화자 임베딩으로서는, 화자 번호에 따라 학습된 고정된 길이의 벡터를 사용하였다. 화자 임베딩은 Transformer 인코더의 출력 및 디코더 pre-net의 출력마다 각각 더해져서 전체 모델이 훈련되는 동안 함께 학습되도록 하였다.

데이터셋으로는 총 19281 발화 분량의 31명 화자의 개성표현용 다화자 한국어 데이터셋을 이용하였고, 이 중 총 27명의 화자에 대해 16778 발화에 해당하는 데이터 중에서 테스트용으로 화자 별로 10 발화를 제외한 나머지 데이터를 가지고 제안한 모델을 훈련하였다. 훈련이 끝난 뒤 테스트용 270 문장에 대해 합성 속도를 측정한 결과, 1.098의 real time factor를 얻었다.

<감사의 글>
본 연구는 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구입니다 (No. 2021-0-00575, 음성·텍스트 딥러닝 기반 보이스피싱 예방 기술 개발).

Multi-Task 기반의 공감형 발화 분류 모델

김 종 인, 정 민 화
서울대학교 인지과학 협동과정, 서울대학교 언어학과

A Multi-task based Empathetic Utterance Classification

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

공감은 다른 사람의 세계를 주관적으로 인식하여, 그 사람이 경험한 것처럼 내재적을 경험하는 과정을 의미한다. 공감은 사람과 사람간의 대화에서 중요한 요소로 고려되고 있지만, 음성언어처리 분야에서의 공감에 대한 연구는 아직 활발히 이루어지고 있지 않다. 또한 기존 연구에서의 공감은 공감의 구성 요소 측면에서 공감의 정서적 측면을 고려하여, 단순히 감정을 분류하는 관점에서 공감을 모델링한 실험이 대부분이다.

본 연구에서는 음성언어처리에서 분야에서의 공감을 재정의 하고, 이를 반영한 공감 발화 분류 모델 실험을 진행하였다. 본 연구에서 공감은 공감의 구성 요소의 측면에서 정서적 측면 뿐만 아니라, 인지적인 측면을 고려하였으며, 공감 과정의 측면에서 공감 상황을 인지하는 것과 공감 발화를 표현하는 것으로 나누어 설계하였다.

본 실험에서는 Aihub 멀티모달 영상 데이터셋을 이용하였으며, 이는 100시간 분량의 실제 대화 상황을 촬영한 데이터셋이며, 85,113개의 발화와 8,000개의 영상클립으로 이루어져 있다. 본 실험에서 사용된 모델은 동일 발화에 대하여 공감과 관련된 4가지 테스크를 분류하는 모델이며, 이는 Cognitive Empathy Awareness, Affective Empathy Awareness, Empathy Expressive, Question로 구성하였다. 본 실험에서는 멀티 테스크 기반의 공감형 발화 분류 베이스라인 시스템을 제안하였다.

본 연구의 최종 목표는 공감형 대화 음성 챗봇을 개발하여, 정서적 공감과 인지적 공감이 필요한 상황을 인지하고 이에 따라 적절한 공감 발화를 생성하는 것이다.

지각 훈련을 통한 한국어 폐쇄음 음향 신호 가중치의 L2 학습

오 은 진
이화여자대학교 영어영문학부

Learning Acoustic Cue Weights for Korean Stops through L2 Perception Training

Eunjin Oh
Dept. of English Language and Literature, Ewha Womans University
ejoh@ewha.ac.kr

본 연구는 한국어 평음과 기음 대조에 초점을 맞춘 지각 훈련을 통해 한국어 학습자들이 L2 폐쇄음을 대조하는 음향 신호의 지각 가중치를 모국어 값의 방향으로 개선하는지 고찰하였다. 중국어가 모국어인 한국어 학습자 19명과 지각 훈련을 진행한 한국어가 모국어인 교사 2명이 실험에 참가하였다. 사전 테스트 결과에 따라 훈련 집단과 비훈련 집단으로 나누고, 훈련 집단만 5일 동안 지각 훈련을 진행하였다. 폐쇄음 신호의 지각 가중치를 추정하기 위해 음향 신호를 체계적으로 조작한 자극으로 사전 및 사후 테스트를 시행했다. 개별 학습자들의 테스트 결과에 대해 이분형 로지스틱 회귀 분석을 시행해, 말소리 대조를 지각하는 데 해당 음향 신호를 사용한 가중치를 추정하는 지각 β 계수 값을 계산하였다. 두 폐쇄음을 구별하는 주요 신호인 F0의 지각 계수에 대해, 훈련 집단은 사전 테스트 대비 사후 테스트에서 평균 0.451의 통계적으로 유의미한 증가를 보인 반면 비훈련 집단은 유의미하지 않은 0.246의 증가를 나타냈다. 그러나 지각 훈련 후 F0 사용의 변화 패턴이 개별 학습자들 간에 다양하게 나타났다.

The influence of processing levels on the perception - production link of L2 phonotactics

Song Yi Kim, Jeong-Im Han
Dept. of English, Konkuk University
adobedobi04@konkuk.ac.kr, jhan@konkuk.ac.kr

It is widely known that Korean learners of English repair illicit consonant sequences by perceiving an illusory vowel and producing the sequences with an epenthetic vowel. This study investigates whether the perception - production link of illusory vowels in second language (L2) is modulated by processing levels. Korean intermediate to advanced learners of English and English native speakers performed perception and production tasks at the prelexical (AX discrimination and pseudoword read-aloud tasks) and at the lexical levels (lexical decision and picture-naming tasks). It was shown that as compared to the native English speakers, Korean learners of English poorly performed in perception tasks both at the prelexical and lexical processing levels, rarely identifying illusory vowels in consonant clusters. Additionally, Korean speakers produced more words with epenthetic vowels than native speakers. More importantly, however, Figure 1 and Figure 2 illustrate that accuracy in not producing an epenthetic vowel between the two consonants of onset cluster was not significantly associated with accurate perception of the clusters either within or across processing levels in Korean learners of English. This disconnection between perception and production questions the presumption that perception precedes production in L2 acquisition and suggests that production and perception accuracy in L2 phonotactics are independent to a certain extent.

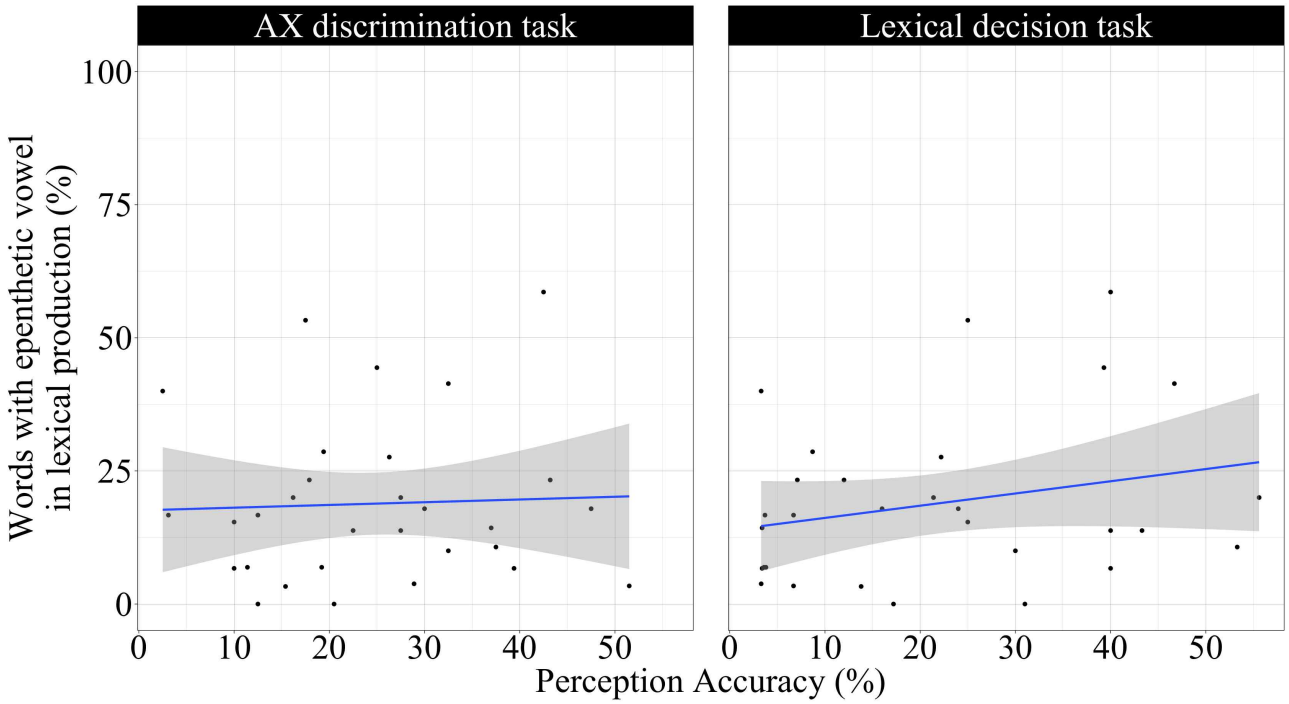


Figure 1. Production of epenthetic vowels over AX discrimination accuracy or lexical decision accuracy in pseudoword read-aloud task

6.

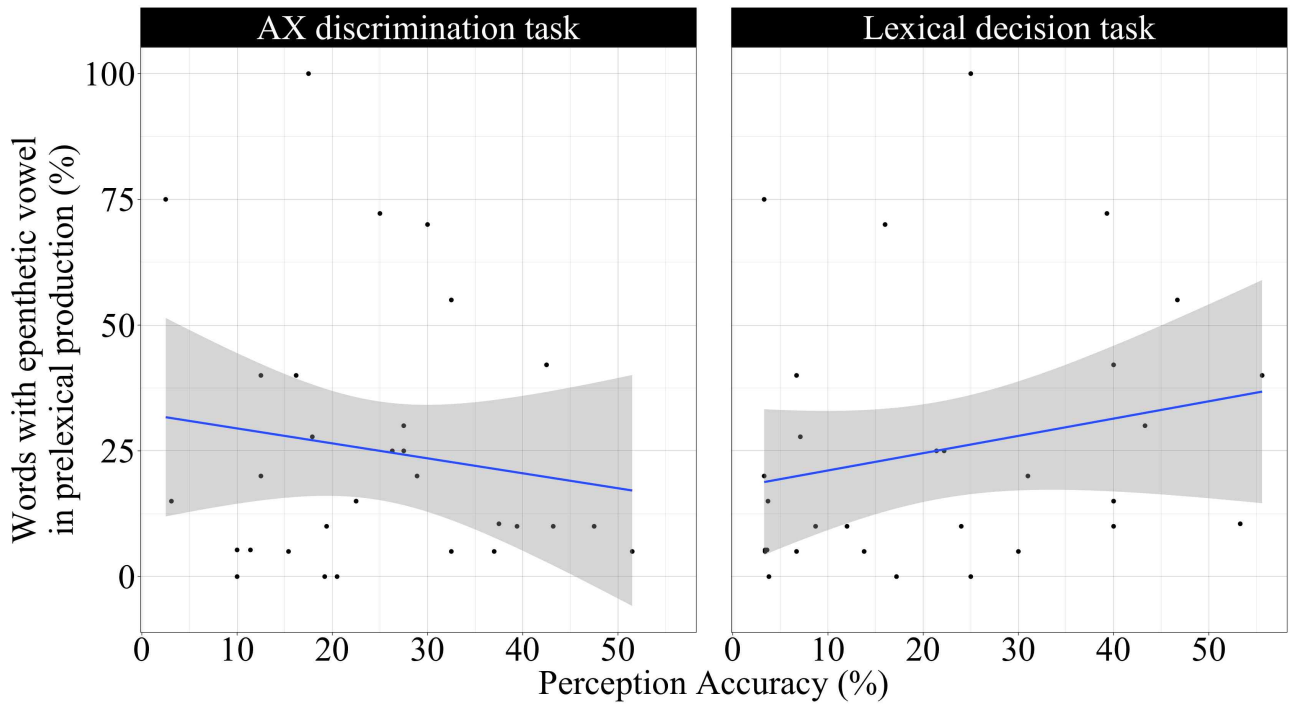


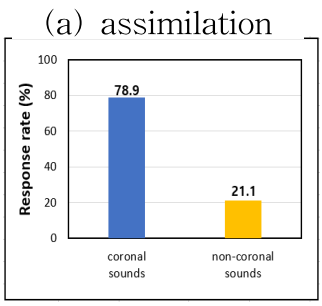
Figure 2. Production of epenthetic vowels over AX discrimination accuracy or lexical decision accuracy in picture-naming task

Word recognition in English coronal place assimilation: Eye tracking study

Eunkyung Sung*, Sehoon Jung**, Sunhee Lee***, Seulgi Oh****
Dept. of English, Cyber Hankuk University of Foreign Studies*
Dept. of English, Kyungsung University**
Dept. of Chinese, Cyber Hankuk University of Foreign Studies***
Dept. of Korean, Hankuk University of Foreign Studies****
eks@cufs.ac.kr, sejung@ks.ac.kr, lishanxi@cufs.ac.kr, youa501@hufs.ac.kr

This study explores the dynamics of lexical activation by comparing the time course of word recognition between assimilated forms (e.g. ca[$^{t}_{p}$] in cat box) and noncoronal forms(e.g. ca[p] in cap box). Specifically, this study investigates how gradient modification in place assimilation context influences word recognition. In addition, the results of the eye movenent and those of the keyboard click are compared. Furthermore, native English and Korean speakers' proportions of fixations on targets and competitors are examined.

A total of 34 participants, including twenty native Korean speakers, as well as fourteen native English speakers as controls, took part in this eye tracking experiment. The stimuli consisted of 190 test trials, including 10 practice, 120 experimental - 30 items involving labialization (e.g. ca/t/ box → ca[$^{t}_{p}$] box) and 30 counterparts with labial codas (e.g. ca/p/ box → ca[p] box); 30 items involving velarization (e.g. ba/t/ cage → ba[$^{t}_{k}$] cage) and 30 counterparts with velar codas (e.g. ba/k/ cage → ba[k] cage) - and 60 filler items. Both English and Korean speakers show higher proportions of fixations on targets (e.g. cat) than on competitors (e.g. cap) in assimilation context (e.g. ca[$^{t}_{p}$] box), as well as higher proportions of fixations on targets (e.g. cap) than on competitors (e.g. cat) in non-assimilation context (e.g. ca[p] box). However, the discrepancy of fixation the Korean speakers. proportions between targets and competitors was more obvious for the English speakers than for the Korean speakers.



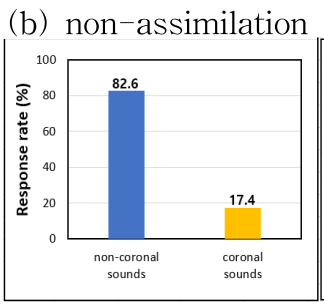
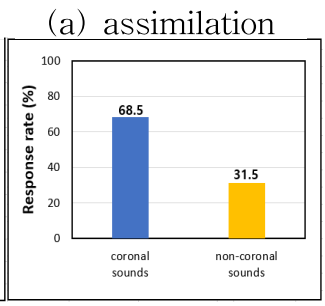


Figure 1. (a) English speakers' response rates of coronal (e.g. cat) and noncoronal codas (e.g. cap) in assimilation context (e.g. ca[$^{t}_{p}$] box), and (b) response rates of noncoroal (e.g. cap) and coronal codas (e.g. cat) in non-assimilation context (e.g. cap box)



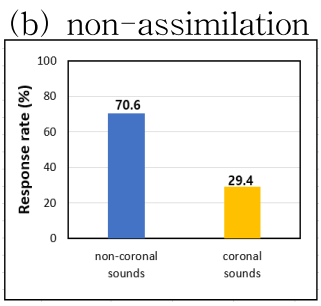


Figure 2. (a) Korean speakers' response rates of coronal (e.g. cat) and noncoronal codas (e.g. cap) in assimilation context (e.g. ca[$^{t}_{p}$] box), and (b) response rates of noncoronal (e.g. cap) and coronal codas (e.g. cat) in non-assimilation context (e.g. cap box)

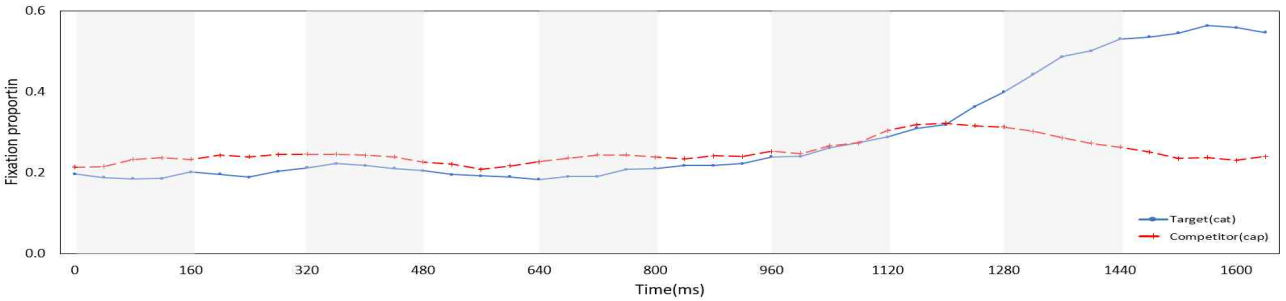


Figure 41. English speakers' proportions of fixations on coronal targets (e.g. cat) and noncoronal competitors (e.g. cap) as a function of processing of assimilated auditory stimuli (e.g. ca[$^{t}_{D}$] box)

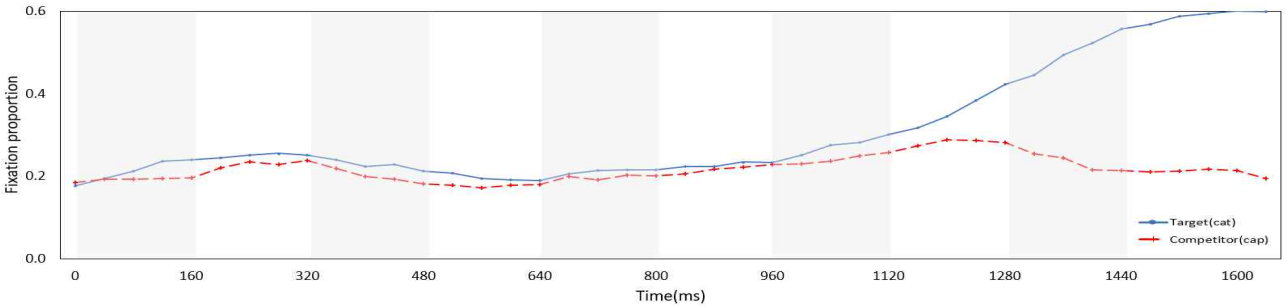


Figure 42. English speakers' proportions of fixations on noncoronal targets (e.g. cap) and coronal competitors (e.g. cat) as a function of processing of non-assimilated auditory stimuli (e.g. cap box)

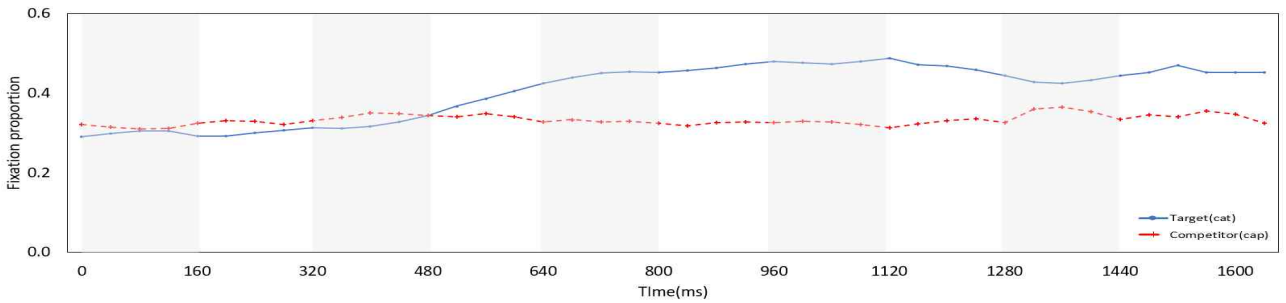


FIgure 43. Korean speakers' proportions of fixations on coronal targets (e.g. cat) and noncoronal competitora (e.g. cap) as a function of processing of assimilated auditory stimuli (e.g. ca[$^{t}_{D}$] box)

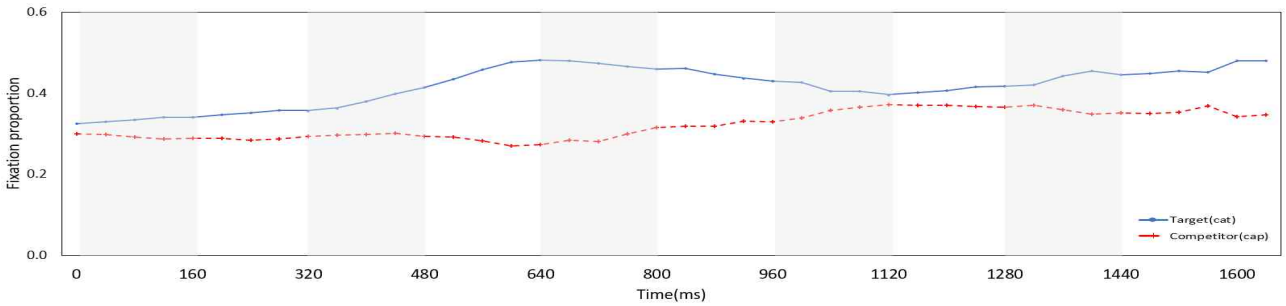


Figure 44. Korean speakers' proportions of fixations on noncoronal targets (e.g. cap) and coronal competitors (e.g. cat) as a function of processing of non-assimilated auditory stimuli (e.g. cap box)

<Acknowledgment>

This research is supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2019S1A5A2A01047388).

프랑스어 강세 음절에 오는 비강 모음의 음향적 특징

박 혜 숙[1], 김 선 희[2]
[1]서울대학교 외국어교육과, [2]서울대학교 불어교육과

Acoustic characteristics of nasal vowels in French stressed syllables

Hye-Sook Park[1], Sunhee Kim[2]
[1]Dept. of Foreign Language Education, [2]Dept. of French Language Education,
Seoul National University
cielcine@snu.ac.kr, sunhkim@snu.ac.kr

일반적으로 강세 위치에 나타나는 모음은 비강세 위치의 모음에 비하여 길게 발화되는 것으로 알려져 있는데, 그 높이와 세기의 실현에 대해서는 의견이 일치하지 않은 경향을 보인다(Schweitzer & Dodane, 2020 ; Abry & Veldeman-Abry, 2007 ; Delattre & Monnot, 1968). 프랑스어에는 한국어의 음소 체계에 없는 비강 모음 [ɑ̃], [ɛ̃], [ɔ̃]이 있다. 비강 모음은 구강 모음과 달리, 연구개가 하강하면서 구강 뿐 아니라 비강으로도 공기가 유출되는 조음적 특징을, 비강으로 공기가 유출되는 시점부터 반전 공명 Antiformant이 나타나는 음향적 특징을 나타다(Clements et al., 2015; Delvaux et al., 2002). 본 연구는 프랑스어의 비강 모음이 강세음절에 나타날 때 그 음향적 특징을 규명하는 것을 목적으로 한다. 데이터는 온라인 프랑스어 교수학습 용 단어 오디오 가운데 강세 음절(1음절)에 비강 모음을 가진 163개(CṼ.CV)와 비강세 음절(2음절)에 비강 모음을 가진 140개(CV.CṼ)의 오디오 총 303개를 수집하였다. 강세음절과 비강세음절의 비강 모음에 대하여 지속시간, F0, 강도 및 F1, F2, F3를 추출하였다(Borel, 2015 ; Ghio & Pinto, 2007 ; Vaissière, 2007 ; Takemoto et al., 2006 ; Smith, 1951). 통계분석 결과, 비강세음절에 비하여 강세음절의 지속 시간은 세 개의 모음 모두 더 길게 나타났다. F0도 강세음절에서 세 모음 모두 높게 나타났다. 강도 또한 강세 음절에서 비강세 음절보다 높게 실현되었다. F1의 경우는 [ɔ̃]을 제외하고 다른 두 개의 모음이 모두 강세음절에서 비강세 음절에 비하여 낮게 실현되었고, F2의 경우는 세 모음이 모두 강세 음절에서 낮게 실현되었다. F3의 경우는 [ɔ̃]을 제외하고 다른 두 개의 모음이 모두 강세음절에서 비강세 음절에 비하여 낮게 실현되었다. 즉, 프랑스어의 비강모음은 강세음절에서 혀의 높이가 높아지고 혀의 위치는 좀 더 후설로 이동하는 경향을 보이며, 이때 원순성도 커진다고 볼 수 있다.

코로나19로 인한 마스크 착용이 아동 말소리 발달에 미치는 영향 연구

박 범 준, 윤 수 연
충남대학교 언어학과

The Impact of Face Mask Use on Child Language Development in the COVID-19 Pandemic

Beomjoon Park, Suyeon Yun
Dept. of Linguistics, Chungnam National University
0925qjawns@naver.com, suyeon.yun@cnu.ac.kr

코로나19 대유행으로 인한 마스크 착용은 얼굴 아래쪽 절반을 가림으로써 발화시 음향 신호의 전달을 방해하고 입술 모양의 시각적 정보를 차단한다. 이에 따라 언어발달기에 있는 아동의 언어지연이 우려되고 있으며, 한 조사에서는 어린이집 원장 및 교사의 약 75%가 마스크 사용으로 인해 언어 노출 및 발달 기회가 감소했다고 응답했다(정춘숙, 사교육없는세상 2021). 그러나 이를 뒷받침할 경험적 증거는 아직까지 제시된 바 없다. 본 연구는 언어발달기에 주변인이 마스크를 착용한 채 오랜 시간을 보내는 아동의 언어발달 척도를 검사 및 분석하고 그 결과를 기존 연구의 코로나19 이전 아동의 발달 척도와 비교함으로써 마스크 착용이 말소리 발달에 유의미한 영향을 미치는지 알아본다. 코로나19 대유행 시작 이후 약 1년 6개월 동안 보육교사가 상시 마스크를 착용해 온 경기도 소재 어린이집 2개소의 아동 23명(2;5~4;7, 남 9명·여 14명)에 대해 '우리말 조음·음운검사2(U-TAP2; 김영태 외 2018)'의 단어검사를 실시하고 그 발화를 녹음하여 분석했다. 검사 단어에 대한 음성 전사를 바탕으로 개정자음정확도(PCC-R) 점수를 산출한 결과, 기존 연구(하지완 외 2019)에서 보고된 코로나19 이전 동일 연령대 아동의 점수 평균보다 낮은 점수를 기록한 참가 아동의 비율이 2;5-2;10 그룹에서 100%(1/1), 3;0-3;5 그룹에서 100%(9/9), 3;6-3;11 그룹에서 87.5%(7/8), 4;0-4;7 그룹에서 80%(4/5)로, 대부분의 실험 참가 아동이 평균 이하의 점수를 보였으며, 이들 중 평균보다 10점 이상 낮은 점수를 보인 아동의 비율은 각각 100%(1/1), 88.9%(8/9), 71.4%(5/7), 75%(3/4)이었다. 이는 마스크 착용으로 인해 실제로 아동의 말소리 발달이 지연되고 있을 가능성을 시사한다.

참고문헌

김영태, 신문자, 김수진, 하지완. 2018. *우리말 조음·음운검사2(U-TAP2)*. 인싸이트.
정춘숙, 사교육없는세상. 2021. 코로나19가 아동 발달에 미친 영향과 그 해법을 모색한다. 온라인 국회토론회 자료.
하지완, 김수진, 김영태, 신문자. 2019. 자음정확도와 단어단위 음운지표를 이용한 일반아동의 말소리 산출 능력에 대한 발달 연구. *Communication Sciences & Disorders*, 24(2): 469-477.

Different phonetic reduction patterns of English function words between native English and Korean speakers

Wooji Park, Seok-Chae Rhee
Dept. of English Language and Literature, Yonsei University
woojipark@yonsei.ac.kr, schree@yonsei.ac.kr

The present study explored the reduction degrees of English function words produced by 12 native English and 24 Korean speakers. In addition, it examined whether subcategories of English function words and Koreans' varying English proficiency could affect their function word reduction patterns. Based on the previous research, this study sub-categorized function words into a strong function word class (including determiners, Wh-words, and demonstratives) and a weak function word class (including articles, prepositions, conjunctions, modal auxiliaries, finite auxiliaries, relative pronouns, and pronouns). The findings revealed that native English and Korean participants reduced function words differently depending on subcategories, and such differences were seen in duration and intensity measurements. Also, native English and Korean participants with higher English proficiency reduced function words in a strong function word class less than those in a weak function word class. However, Korean participants with lower English proficiency did not reduce conjunctions as much as native English and Korean participants with higher English proficiency but lengthened the most among ten subcategories. These results have important implications for understanding that function words are reduced differently depending on subcategories, and varying English proficiency levels of Korean speakers create variances in the representation of English function word phonetic reduction.

Korean learners' production of English /l/: Ultrasound and acoustic analyses

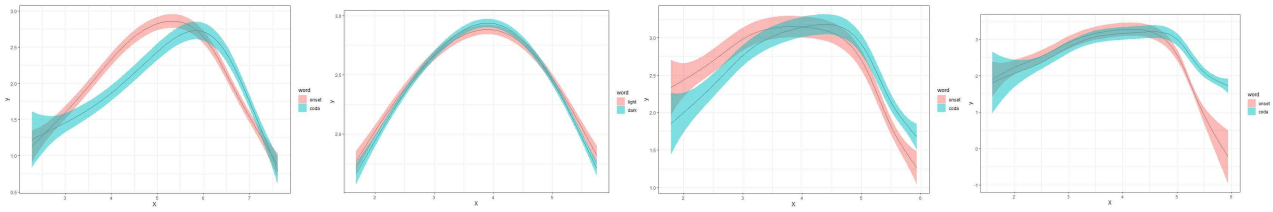
Joo-Kyeong Lee
Dept. of English Language and Literature, University of Seoul

jookyeong@uos.ac.kr

The current study presents ultrasound and acoustic analyses for Korean learners' production of English /l/ in onset and coda positions. Five Korean learners of English, who were rated as intermediate in English proficiency from a Foreign Accentedness (FA) task, participated in the ultrasound recording of the tongue. Their ultrasound data and temporal measurements of [l] were compared with those of an American English speaker. Both native English speaker and Korean learners of English read a carrier sentence 'Please say ___ for me.' where one syllable real English words containing [l] in onset or coda positions were embedded. The target [l] sounds were annotated first based on waveforms and spectral configurations and their duration was measured. Sproat & Fujimura (1993) reported that not only dark-[ɬ] but also clear-[l], to some extent, can be articulated with darkening (tongue body lowering and tongue dorsum retraction) in American English. Clear-[l] and dark-[ɬ] were categorically different, but the degree of darkening was affected by prosodic boundaries within each category of [l] in various dialects of British English (Turton, 2017). Spline coordinates (over 42 points) were extracted, for onset [l], from the frame where the tongue tip moved the most forward and, coda [l], the frame where the tongue dorsum was retracted the most. For statistical testing, mixed-effect regression was used for durational measures, while SS-ANOVA was used for tongue contours.

Results showed that three Korean learners of English did not show any difference in tongue contours between onset and coda-[l], but that two learners articulated onset and coda-[l] somewhat differently. None of them, however, produced onset and coda [l] sounds as distinctively as the native speaker; tongue body lowering and tongue dorsum retraction were not clearly observed for the coda-[l] as illustrated in Figure 1. More specifically, three of the Korean learners did not register darkening for coda-[l] but consistently produced clear-[l] both in onset and coda positions as shown in Figure 1 (b). Two of them implemented darkening gesture to some degree, but they did not articulate tongue dorsum retraction without tongue body lowering as shown in Figure (c) and (d). The duration measurements were normalized by dividing the duration of /l/ by the duration of the vowel /ej/ of *say* in the sentence '*Please say ____ for me*.' The duration data were submitted to a linear mixed effect analysis in R. Figure 2 shows that both onset and coda [l] sounds were significantly longer in the native speaker's production than those of Korean speakers ($\chi^2(1)=7.4024$, $p<0.01$). While two variants of /l/ were not significantly different in length for the native speaker ($p=0.663$) while the duration of onset-[l] significantly longer than that of coda-[l] for all Korean learners ($p<0.001$). Even though Korean learners made a tongue tip contact with the alveolar ridge for onset-[l] as native English speakers did, the duration of the contact was not sufficiently long in the case of Korean learners' production and

therefore resulted in a flap [ɾ]. The two Korean learners who articulated the darkening gesture for coda-[l] did not make darkening as extended as the native speaker. It might be attributed to the shorter duration; there was no sufficient time for tongue body lowering and tongue dorsum retraction.



(a) (b) (c) (d)

Figure 1. Smooth splines of [l] in onset and coda positions: (a) native speaker, (b-d) Korean learners.

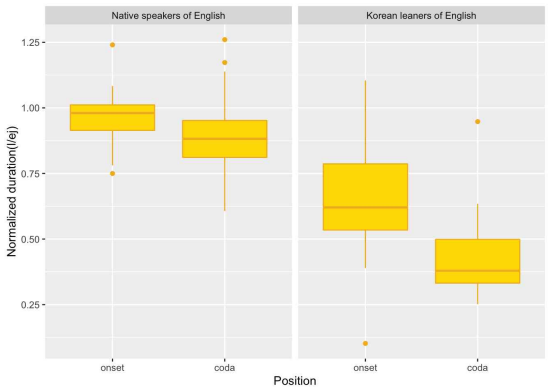


Figure 2. Normalized duration of [l] for native speaker and Korean learners.

References

Sproat, Richard & Fujimura, Osamu (1993) Allophonic variation in English /l/ and its implications for phonetic implementation, *Journal of Phonetics*, 21, 291-311.

Turton, Danielle (2017) Categorical or gradient? An ultrasound investigation of /l/-darkening and vocalization in varieties of English, *Laboratory Phonology*, 8(1): 1-31.

< Acknowledgment >

This work was supported by the research fund given to Joo-Kyeong Lee from the University of Seoul.

특강 II

좌장: 박상희(대구사이버대)

언어병리학 음성장애 영역에서의 AI 연구 최신 경향 및 임상 특성

(김근효, 부산대병원)

언어병리학 음성장애 영역에서의 AI 연구 최신 경향 및 임상 특성

김 근 효
부산대학교병원 이비인후과 음성언어치료실

Latest AI research and clinical characteristics in the field of voice disorder

Geunhyo Kim
Dept. of Otorhinolaryngology, Pusan National University Hospital
kimgeunhhyo@gmail.com

본 내용은 언어병리학 음성장애 분야에서 진행되고 있는 인공지능 관련 연구들과 임상현장에서 어떻게 적용되고 있는지에 대한 내용이다. 성대 영상, 음성 샘플과 같은 많은 데이터가 생성되고 있으며 이러한 데이터를 잘 활용하여 음성장애의 조기 진단 및 중재에 중요한 역할을 할 수 있다. 최근 발표된 머신러닝, 딥러닝 관련 연구들을 살펴보며 음성학, 언어병리학 분야에서 시도할 수 있는 내용들을 알아보고자 한다.

구두 발표 Ⅱ: 음성학 및 음운론

좌장: 오은진(이화여대)

An acoustic study of Korean mothers' vowel space

Eon-Suk Ko, Sunghye Cho
Department of English Language and Literature, Chosun University
Linguistic Data Consortium, University of Pennsylvania
eonsukko@chosun.ac.kr, csunghye@sas.upenn.edu

This study evaluates opposing claims on the didactic function of infant-directed speech (IDS) in infants' learning of vowel categories as reflected in Korean mothers' IDS. According to the tutorial function hypothesis of IDS, the formants in IDS might be produced in a way that could enhance vowel contrasts, but it might be that enhanced classification of vowel categories is not the key mechanism behind infants' learning of vowel contrasts. 23 Korean mothers participated in a word-teaching task with their 9- and 16-month old infants, which were recorded and transcribed. Despite some indication for local enhancement effects in the formant values, consistent across-the-board enhancement of vowel categories in the IDS were not confirmed. Consistent with recent findings in English IDS, we found greater within-category variation in IDS compared to ADS.

서울말 내포문 wh-섬 제약의 지각 및 반응시간 연구

윤 원 희
계명대학교 영어영문학전공

Perception and Response Time of the Wh-island Constraint in Seoul Korean

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

내포문에 포함된 의문사 구는 모문의 동사가 비교량 동사일 경우 모문의 작용역을 갖을 수 없는 wh-섬 제약이 존재하는 것으로 알려져 왔으나 의문사 억양으로 문장이 발화되면 모문 작용역 해석도 가능하다는 선행 연구가 보고되었다. 그러나 기존의 실험이 소수의 실험 참가자일 경우에 국한하거나, 실험 절차의 문제점이 발견되어 wh-섬 제약의 위배가 보편적으로 나타나는 것인지에 대한 의문이 제기되었다. 이에 49명의 서울말 화자를 대상으로 서울말 내포문 의문사 구를 모문 작용역으로 해석하는지에 대한 청취 실험을 진행하였고, 이와 함께 반응시간을 분석하였다. 실험 결과 피험자의 83.7%가 의문사 억양으로 읽은 내포문 의문사 포함 문장의 의문사 작용역을 모문으로 받아들였고, 의문사 억양으로 발화된 문장의 반응시간은 대답의 정오를 떠나 긴 시간이 소요됨을 보여주었다. 세 번의 청취 실험을 통해 50명의 화자 중 1명만이 모문 작용역 해석을 전혀 받아들이지 않았고, 이와 별도로 첫 회 8명에서 나타난 wh-섬 제약이 2, 3회의 실험을 거치며 각각 6명과 3명으로 줄어든 것으로 보아 서울말에서 이 제약이 완전히 위배 된다고는 볼 수 없으며 다만 약한 wh-섬 제약이라고 말할 수 있다.

Research on English Speakers' Production of Word Boundary Stops Based on Voicing Found in Spontaneous Speech

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

The current study investigates how the voicing types influence durations of homorganic word boundary stops. For this purpose, Buckeye speech corpus, which contains numerous spontaneous speech materials, were used. To measure the word boundary stops such as /p#p/, /b#p/, /p#b/, and /b#b/, stop closures of both word-final stops and those plus VOTs of word-initial stops were selected. The results were analyzed based on voicing types such as /voiceless#voiceless/, /voiced#voiceless/, /voiceless#voiced/, and /voiced#voiced/, and showed that word-boundary stops with word-initial voiceless stops are longer than those with word-initial voiced stops due to the VOT differences, and those with alveolar stops are shorter than those with bilabial and velar stops. In addition, speech rate masked the closure duration differences of word-final stops.

구두 발표 Ⅱ: 말장애 및 음성의학

좌장: 이옥분(대구사이버대)

마비말장애 음성인식 성능에 영향을 미치는 언어학적 특징 분석

여 은 정, 김 선 희, 정 민 화
서울대학교 언어학과

Analysis of linguistic characteristics affecting ASR performance for dysarthric speech

Eun Jung Yeo[1], Sunhee Kim[2], Minhwa Chung[1]
Dept. of Linguistics, Seoul National University[1]
Dept. of French Language Education, Seoul National University[2]
ej.yeo@snu.ac.kr, sunhkim@snu.ac.kr, mchung@snu.ac.kr

마비말장애란 중추 신경계 및 자율 신경계의 손상으로 말소리 산출과 관련된 근육이 마비되거나 약해지면서 발생하는 말운동장애이다. 마비말장애인은 대개 사지 움직임의 불편함을 동반하기 때문에 음성인식 기반 인터페이스를 사용하는 것이 유용하다. 그러나 마비말장애 음성은 비장애 음성과 상이한 음성학적 특성을 가지기 때문에 비장애인 대상 음성인식기를 사용할 시 낮은 음성인식 성능이 나타난다. 본 연구는 어떤 언어학적 요인들이 마비말장애 음성인식 성능에 영향을 미치는지 상관분석과 회귀분석을 통해 알아본다. 분석에 사용된 특징은 다음과 같다: 음질의 jitter, shimmer, HNR, voice breaks 개수, voice breaks 정도; 운율-발화 속도의 말속도, 조음속도, 쉼의 개수; 운율-음높이의 F0 평균값, 표준편차, 최솟값, 최댓값, 중앙값, 25 분위수, 75 분위수, 범위; 운율-리듬의 %V, Varco-V, Varco-C, nPVI-V, nPVI-C, 발음-모음왜곡도의 모음공간면적, FCR, VAI, F2-Ratio. 상관분석 결과, F0의 평균, 최솟값, 25 분위수, 75 분위수를 제외한 모든 요인과 음성인식기의 SER(Syllable Error Rate) 간 유의미한 상관 관계가 나타났다. 회귀분석 결과, voice breaks 개수(β=.26, p<.05), 말속도(β=-.24, p<.05), 모음공간면적(β=-.22, p<.05), Varco-V(β=-.20, p<.05), shimmer(β=.19, p<.05), Varco-C(β=.18, p<.05), nPVI-C(β=.14, p<.05), nPVI-V(β=.11, p<.05)가 SER을 유의미하게 예측하였다. 본 연구는 마비말장애 음성인식기의 성능 향상의 방향성을 탐색했다는 점에서 그 의의가 있다. 특히, 음성인식 성능 향상을 위한 음성 개선(voice conversion) 연구의 우선순위를 정하는 것에 있어 도움이 될 것으로 기대된다.

연축성 발성장애 환자에서 발화범위 프로파일의 특성

이 승 진[1], 김 재 옥[2], 임 성 은[3], 임 재 열[4]
1) 한림대학교 언어청각학부 및 청각언어연구소
2) 강남대학교 교육대학원 언어치료교육전공
3) 강남세브란스병원 이비인후과 음성언어치료실
4) 연세대학교 의과대학 이비인후과학교실

Characteristics of speech range profile in patients with spasmodic dysphonia

Seung Jin Lee[1], Jaeock Kim[2], Sung-Eun Lim[3], Jae-Yol Lim[4]
1) Division of Speech Pathology and Audiology, Research Institute of Audiology and Speech Pathology, Hallym University
2) Major in Speech Pathology Education, Graduate School of Education, Kangnam University
3) Voice Clinic, Department of Otorhinolaryngology, Gangnam Severance Hospital
4) Department of Otorhinolaryngology, Yonsei University College of Medicine
sjl@hallym.ac.kr, jaeock@gmail.com, selim@yuhs.ac, jylimmd@yuhs.ac

목적: 연축성 발성장애(adductor spasmodic dysphonia, AdSD)는 후두의 불수의적 근긴장 이상을 특징으로 하는 신경학적 음성장애이며, 발성하는 음도에 따라 그 증상이 상이하게 발현되는 특징을 보일 뿐만 아니라, 평가 과제에 따라 감별진단 가능 여부가 달라질 수 있다. 따라서 이 연구에서는 AdSD 환자에서 발화범위 프로파일(speech range profile, SRP) 과제를 시행하여 정상 음성과 구분되는 음역대 변수를 확인하고, AdSD의 예측요인을 확인하고자 하였다.

방법: 연구대상은 서울 소재 대학병원 이비인후과에서 AdSD로 진단된 여성 환자 25명(연령=34.7±14.5세)과 정상 성인 여성 44명(연령=32.6±10.9세)였다. 각 대상자는 '불이야' 문단을 활용한 SRP 과제를 시행하였으며, 최고 및 최저 음도와 Hz 및 반음 단위 음도 범위, 최고 및 최저 에너지와 에너지 범위를 측정하여 집단 간 비교를 하고, 로지스틱 회귀분석을 통해 예측 요인을 확인하였으며, ROC 곡선 분석을 통해 절단점을 확인하였다.

결과: SD군의 최고 및 최저 음도, 최고 에너지는 정상군에 비해 낮았으며, Hz 및 반음단위 음도 범위, 에너지 범위가 정상군에 비해 좁았다. AdSD의 유의한 예측요인은 최저 음도와 최고 에너지였다. AUC가 0.9 이상인 변수는 최고 음도(≤415.3Hz)와 최고 에너지(≤107dB)였다.

결론: AdSD군과 정상군 간 발화범위 프로파일 측정치의 차이는 단순한 모음 연장과제가 아닌 다양한 음도와 음량을 유도하는 발화범위 프로파일 과제 수행 시 음도와 음량의 특성이 AdSD에 대한 감별진단 도구로서 임상적 유용성을 가질 수 있음을 시사한다.

말소리장애아동에 대한 인공지능과 언어재활사의 음성인식 비교

천 시 온, 최 성 희, 이 경 재, 최 철 희
대구가톨릭대학교 언어청각치료학과

Comparison of speech perception between AI and speech-language pathologists for children with speech sound disorders

Sion Cheon, Seong Hee Choi, Kyoungjae Lee, Chul-Hee Choi
Dept. of Audiology & Speech Language Pathology, Daegu Catholic University,
songofcon@gmail.com, shgrace@cu.ac.kr, kjlee0119@cu.ac.kr, cchoi@cu.ac.kr

최근 인공지능을 이용한 음성인식 기술이 삶에 편리성을 더하고 있다. 본 연구는 현재의 상용화된 인공지능을 대상으로 인공지능 유형별로 아동의 발음 명료도에 따라 정상아동과 말소리장애아동의 말소리 인식 정도와 말소리 자극어(낱말 vs. 문장)에 따른 음성인식 정도를 살펴보았다. 또한 인공지능의 말소리 인식 능력과 언어재활사가 평가한 말소리 인식을 비교함으로써 현재 장애아동 말소리에 대한 인공지능의 음성인식 현황을 살펴보고자 하였다.

따라서 본 연구는 아동 총 20명(정상아동 10명, 말소리장애아동 10명)이 참여하였으며, U-TAP2를 사용하여 아동이 산출한 목표 낱말과 문장 발화를 녹음한 음성 파일을 사용하였다. 본 연구에서 사용한 음성인식 기능을 갖춘 인공지능의 유형의 경우 Papago, Google, Naver, Apple(iPhone), Samsung(A9) 모델을 사용하였으며, 아동의 음성 파일에 대한 인공지능과의 음성인식률 비교와 언어재활사 간 음성인식률에 대한 상관성을 알아보기 위해 언어재활사 2명이 참여하였다.

본 연구 결과, 음성인식기능을 갖춘 인공지능 유형 5가지에 따라 정상아동 집단과 말소리장애아동 집단의 음성인식의 인식률의 경우 정상아동 집단이 말소리장애아동 집단에 비해 상대적으로 높은 음성인식률을 보였으며, 유의한 차이가 나타났다. 또한, 인공지능 유형 5가지와 말소리 자극어(낱말 vs. 문장)에 따른 음성인식률에 차이를 알아본 결과, 인공지능 유형에 따라 음성인식률에 유의한 차이가 나타났으나, 자극어에 따른 인공지능 유형별 음성인식률은 유의한 차이를 보이지 않았다.

언어재활사 2명 간 20명의 아동의 음성인식률에 대한 상관성을 알아본 결과, 높은 정적상관관계가 나타났다. 또한, 인공지능 프로그램과 언어재활사집단 간 음성인식률의 차이를 비교한 결과, 언어재활사 집단이 인공지능에 비해 통계적으로 유의하게 매우 높은 음성인식률을 나타내었다.

현재까지 정상아동과 말소리장애아동의 말소리에 대한 인공지능의 음성인식률은 매우 낮은 편이었으며 특히, 언어재활사가 평가한 자음정확도와 비교했을 때 인공지능의 음성인식률은 현저한 차이가 있는 것으로 나타났다. 인공지능의 경우 환경소음과 아동음성과의 변별오류의 제한점을 극복하고자 최대한 소음이 차단된 음성을 조사하였으나 향후 소음환경에서 아동음성인식으로 확대된 연구가 필요할 것이다.

현재 국내에서 다양한 인공지능이 접목된 언어치료용 애플리케이션이 연구, 개발되고 있으나 정상아동 및 말소리장애아동의 음성인식에 대한 기초자료는 전무한 실정이다. 본 연구는 이러한 현재 국내 상황에서 정상아동 및 말소리장애아동이 사용할 수 있는 음성인식 기반 학습이나 치료용 도구 개발을 위하여 본 연구결과가 기초 자료로서 제공되는데 의의가 있다. 본 연구결과를 토대로 향후 말소리장애아동 집단과 정상아동 집단 외에 다양한 말과 언어에 문제를 가진 장애아동 집단을 대상으로 음성인식에 대한 후속 연구가 진행되기를 기대한다.

구두 발표 Ⅱ: 음성공학

좌장: 유하진(서울시립대)

Attention 인공신경망을 통한 한국어 방언의 억양 패턴 학습 및 방언 식별

이 주 영[1], 김 경 화[2], 정 민 화[1]
서울대학교 언어학과[1], 대검찰청[2]

Intonation Modeling and Dialect Identification on Korean Dialects with an Attention Neural Network

Jooyoung Lee[1], Kyungwha Kim[2], Minhwa Chung[1]
Dept. of Linguistics, Seoul National University[1]
Supreme Prosecutor's Office[2]
excalibur12@snu.ac.kr[1], savoix@spo.go.kr[2], mchung@snu.ac.kr[1]

억양 지표는 한국어 방언을 구분하는 척도 중 하나이다. 한국어 방언 연구는 소규모의 데이터를 대상으로 방언 전문가가 방언별 특징을 기술하고 차이점을 밝히는 방향으로 진행되어 왔다. 본 연구에서는 대검찰청의 한국인 표준 음성 DB[1, 2]의 방언 음성에서 나타나는 억양 패턴을 딥러닝 모델이 학습하여, 학습의 결과로서 방언 특징이 잘 나타나는 구간을 분석하고, 이를 모델의 방언 식별 성능과 연결지어 설명한다. 이 때 방언의 음향 특징의 패턴 변화를 모델이 스스로 파악할 수 있도록 Attention LSTM 모델을 사용한다. 또한, 방언 특색이 상대적으로 두드러지게 나타나는 50대 이상의 화자를 대상으로 하며, 방언 특징 구간 비교를 위해 화자별로 동일한 문장을 녹음한 낭독발화에서의 학습 결과를 살펴본다.

본 연구에서 사용한 Attention 알고리즘은 분류하고자 하는 정답 방언과 입력 음향 패턴 사이에서 방언 특징이 두드러지게 나타나는 구간에 모델이 스스로 가중치를 부여하는 방식이다. 따라서, 모델의 Attention 가중치 결과를 살펴보면 모델이 학습한 방언별 특징 패턴을 알 수 있다. 본 연구에서 다루는 방언은 한국어 방언들 중 억양 패턴이 잘 나타나는 경상 방언으로, 경남권과 경북권 방언 분류를 통해 두 지역의 학습 패턴 차이를 살펴보았다. 또한, 학습에 사용한 음향 특징은 20차원 MFCC, 20차원 MFCC-delta, 20차원 MFCC-delta-delta, 그리고 1차원의 F0를 합한 61차원으로 구성하였다.

학습은 171명의 화자로부터 자동 분절 후 추출한 10,000개의 발화로 하였으며, 테스트는 24명의 화자로부터 분절한 1,624개 발화에 대해 진행하였다. 분류 성능은 F1 점수 기준 56.6%였다. [그림 1]은 동일한 문장을 낭독한 경남권과 경북권 60대 화자 각각의 Attention 가중치 결과를 나타낸 것이다. [그림 1]에서 상단 그림은 파형, 하단 그림은 방언 음성에 대한 학습 모델의 Attention 가중치를 의미한다. 화자마다 Attention 가중치의 분포는 조금씩 다르지만 대체로 경남권에서는 음성 전반에 걸쳐, 그리고 경북권은 음성 구간 중간중간에 높은 가중치를 부여한 것을 확인할 수 있었다.

본 연구에서는 경남권 및 경북권의 방언 분류 및 각 방언권에 대한 Attention LSTM 모델의 Attention 가중치 분포를 살펴보았다. 본 연구 방법론 및 결과를 토대로 타 지역 방언권과의 차이도 알아볼 예정이다.

트랜스포머 기반 실시간 한국어 음성인식

오 유 리, 박 기 영*
한국전자통신연구원 인공지능연구소 복합지능연구실

Streaming a Transformer based end-to-end Korean speech recognition

Yoo Rhee Oh, Kiyoung Park
Artificial Intelligence Research Laboratory
Electronics and Telecommunications Research Institute (ETRI)
yroh@etri.re.kr, pkyoung@etri.re.kr

딥러닝(deep learning) 기술, 빅 데이터(big data), 하드웨어 성능 등의 발달과 함께, 음성인식(automatic speech recognition) 성능이 크게 개선되고 있으며 이에 따라 다양한 응용분야에서 음성인식을 적용할 수 있게 되었다. 기존의 음성인식과 비교하여, connectionist temporal classification (CTC)가 결합된 트랜스포머(Transformer) 기반 종단형(end-to-end) 음성인식은 높은 음성인식 성능을 보인다. 하지만, 어텐션(attention)을 트랜스포머 기반 종단형 음성인식은 입력음성의 전체데이터를 필요로 하기 때문에 스트리밍(streaming) 디코딩(decoding)이 어려운 단점을 가지고 있다.

본 발표에서는 스트리밍 디코딩이 가능한 트랜스포머 기반 종단형 한국어 음성인식 소개에 초점을 두고, 스트리밍 디코딩을 위하여 변경된 트랜스포머 인코더(encoder) 구조[1]와 디코딩 알고리즘[2]을 소개한 후, 한국어 음성인식 성능 비교를 보인다.

참고문헌

[1] E. Tsunoo et al., "Transformer ASR with contextual block processing," in Proceedings of ASRU, 2019, pp.427-433.
[2] E. Tsunoo et al., "Streaming Transformer ASR with blockwise sychronous beam search," in Proceedings of SLT, 2021, pp.22-29.

감사의 글

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

* 교신저자

실시간 시청각 발화구간검출 알고리즘

정 세 영, 박 형 민
서강대학교 전자공학과

Real-Time Audio-Visual Voice Activity Detection Algorithm

Se-Yeong Jeong, Hyung-Min Park
Dept. of Electronic Engineering, Sogang University
jsydshs@sogang.ac.kr, hpark@sogang.ac.kr

음성 기술은 사용자의 손을 자유롭게 하고, 편리함을 극대화하는 인간 친화적 가치를 창출할 수 있다. 일상의 삶과 업무, 교육과 학습, 여가와 여행 등 다양한 영역에서 사람과 기계 간 양방향성 인터페이스로 사용할 수 있어 연구에 대한 관심도가 높아지고 있다. 특히 최근 IoT(Internet of Things)가 여는 새로운 세상 속에서 인식 기술은 보다 정확하게 정보를 주고받으며 대화를 나눌 수 있는 성능이 요구된다. 이 과정에서 음성인식 성능 향상을 위해 발화구간을 정확히 검출하는 기술은 매우 중요하다.

현재의 발화구간검출은 보통 마이크로폰에 입력된 음성을 통해서 활성화를 구분하는 방식으로 조용한 환경에서 비교적 높은 정확도를 가지고 있다. 하지만 주변 잡음이 심한 환경에서는 발화구간을 정확히 검출하는 것이 매우 어렵다. 본 연구에서는 마이크 입력 신호 외에 카메라를 이용하여 발화시 화자 입술의 움직임을 검출하여 잡음 환경에서도 강인한 발화구간을 검출하는 기술을 제안하며, 발화구간검출에 사용되는 딥러닝 모델의 경량화를 통해 모델 파라미터 수를 감소시켜 실시간 구동을 가능하게 하였다.

Keyword: 발화구간검출, 딥러닝, 시청각정보처리, 실시간 알고리즘

특강 Ⅲ

좌장: 정현성(한국교원대)

AI 시대, 음성학의 방향

남호성(고려대)

AI 시대, 음성학의 방향

남 호 성
고려대학교 영어영문학과

The AI era, where to go?

Hosung Nam
Dept. of English Language and Literature, Korea University
hnam@korea.ac.kr

20세기 중반 DARPA의 언어 지능에 대한 천문학적인 투자를 시작으로 음성학은 공학과 함께 AI 시대의 서막을 함께 했다. 그렇게 공학과 긴밀한 협업뿐 아니라 20세기말까지 언어학의 여러 분야 중에서는 가장 과학과 공학에 맞닿아 있는 학문분야로서 연구와 교육에서 인문학의 선도적 모범이 되어왔다. 그럼에도 불구하고 2000년을 전후로 하여 음성 인식과 같은 언어 지능이 투자 대비 참담한 성능으로 기대에 미치지 못하자 그간의 투자와 관심이 급격히 줄어 들었다. 그리고 일부의 책임을 참여했던 음성학자를 포함한 언어학자들에게 묻기도 했다. 그렇게 음성학과 음성공학은 서로 다른 길을 가게 된다. 음성공학은 음성인식에 대한 한계가 드러나면서 더이상 음성인식에 대한 교육과 연구가 중단되기에 이르렀고, 공학과 멀어지게 된 음성학은 음운론과 같은 이론의 증명이나 언어교육과 같은 대중 학문 등으로 집중하는 경향이 있었다.

인공지능으로 대변되는 4차 산업혁명 시대, AI는 기계와의 궁극적 대화를 목표로 한다. 다른 여러 분야에서의 AI 기술의 괄목할 만한 발전에도 불구하고 기계와 소통하는 언어 지능의 영역은 여전히 기대에 못 미친다. 이것은 언어 지능에 대해 언어라는 특수성과 복잡성을 배제한 채 공학적으로만 접근했기 때문이라고 볼 수도 있다. 언어는 문학, 문화, 심리, 과학, 철학, 공학 등의 다양한 학문의 융복합체임에도 불구하고 여전히 현재의 언어 지능은 공학자의 전유물인 것처럼 보인다. 언어 지능, 그 중에서도 음성 지능의 컨텐트의 주인인 음성학은 이 문제에 대한 어떤 해결책을 제시하고 있기는 커녕 현재의 기술에 대한 이해마저 부족한 실정이다. AI의 시대, 주인이 주인 노릇을 못하고 있는 셈이다. 융합이 키워드인 이 시대. 인간의 음성에 대한 과학과 공학은 그 어떤 분야보다 융합적이다. 하지만, 음성학은 3차에서 4차로 가는 시대의 변화에서 음성학의 그 융합적 특성에 부흥하지 못하였다. 누군가가 써놓은 교과서 (예: A course in phonetics) 에 갇히어 그 속에서 경쟁하고 생존하기에 급급했다.

그 결과는 참담하기 그지 없다. 전세계적으로 불어닥친 4차 산업 혁명. 기술이 어떠한 가치보다 우위에 있는 시대. 인문학 전반에 불어닥친 이 위기는 앞으로 다가올 5차, 6차 산업 혁명에서 더 나아지지 않을 것 같다. 엎친데 덮친격 한국은 학령인구의 급감으로 대학은 정원을 못 채워 사라져 가고 있고 그 찬바람의 제일 앞에 있는 학과가 음성학을 전공으로 둔 어문계열이라는 것을 부정하기 힘들다. 이것은 전공 지식이 당장의 취업과 직접적 관련이 없어서일 것이다. 그것을 해결하지 못하면 학교와 학과는 유적지가 될 날이 멀지 않은 것 같다. 혹자는 이야기한다. AI 시대, 음성학에 물 들어왔다고 하지만, 우리는 정작 저을 노가 없는 것 같다. 인공지능과 데이터 관련 취업 시장에서 첨단 기술을 가지고 있고 플러스 알파로 언어

학, 음성학적인 지식을 가지고 있다면 더할 나위 없는 스펙이다. 하지만 그 플러스 알파만으로는 전공을 살리는게 힘든 실정이다. 드물게 전공을 살린 취업을 하더라도 중심 역할보단 보조 역할에 만족해야 하는 형편이다.

더이상 학생들이 찾지 않는 전공. 학생이 없는 연구 실적은 공허한 메아리임을 우리는 잘 알고 있다. 어떻게 하면 학생들이 다시 찾는 전공이 될 수 있을까에 대한 희망으로 우리가 보지 못한 '음성'이라는 주제가 내재하고 있는 융합의 요소를 공유하고 3가지 없애 볼 것과 3가지 가져 볼 것, 즉 3무와 3유의 길을 조심스럽게 공유하고자 한다.

구두발표 III: 음성학 및 음운론

좌장: 성은경(사이버한국외대)

A longitudinal study of individual and age-related differences in perception-production links in second language speech

Donghyun Kim
Department of Liberal Arts & Teacher Training, Kumoh National Institute of Technology
dkim@kumoh.ac.kr

The present study examines (1) how second language (L2) speech sound contrasts are acquired longitudinally and (2) how L2 speech perception and production are linked within an individual learner. The study addresses these issues by examining developmental changes in perception and production of English vowel contrasts by Korean learners of English during their first year of immersion in Canada. The study first explores changes in production of English vowel contrasts by Korean adult and child learners. Then, it further investigates how and to what extent developmental changes in production are associated with those in perception. Results revealed that L2 learners showed improvement in the production of L2 vowel contrasts, producing spectrally more distinct vowel contrasts at later time points. Results also indicated that this improvement manifested more in child L2 learners than in adult L2 learners. Importantly, a close link between perception and production was found in child learners after 8 months of exposure when their production abilities had also significantly improved. These results corroborate previous findings that accurate production of L2 speech contrasts is contingent on perceptual abilities at early stages of L2 speech acquisition (Flege et al., 1999).

References

Flege, J. E., MacKay, I. R. A., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *Journal of the Acoustical Society of America, 106*(5), 2973-2987.

The Effects of Focus-on-Form Pronunciation Instruction on The Evaluation of High School Students' Utterances

Ji Sun Yuk
Sejong Global High School
kskimhn@naver.com

The purpose of this study is to incorporate focus-on-form (FonF) pronunciation instruction into the integrated EFL education paradigm in Korea by drawing out the effects of FonF pronunciation instruction during the evaluation of high school students' utterances in specialized English classes, called Public Speaking and Presentation (PSP), which find out how FonF pronunciation instruction and different types of assessment (such as prepared speech tasks and impromptu speech tasks) affects the evaluation (such as accentedness, comprehensibility, and intelligibility, rhythm, fluency, and the pronunciation of English vowels). 64 Korean high school students, with their average English proficiency ranging from an intermediate to an advanced level, participated in this study. Findings were that their overall accentedness, comprehensibility, intelligiblity, rhythm, and fluency were improved through the FonF pronunciation instruction. Moreover, their pronunciation of English vowels became much clearer and more typical after the FonF pronunciation instruction. Types of assessment affected the students' accentedness, comprehensibility, and fluency, while intelligibility and rhythm were not directly affected by the assessment type. Thus, it suggests meaningful implications for teaching pronunciation, emphasizing the effects of FonF pronunciation instruction.

프랑스인 한국어 L2 학습자의 평음, 격음, 경음의 지각 습득 과정과 패턴 - 선택적 주의를 중심으로

이 보 람
소르본 누벨 대학교, 음성음운 연구소

Perception of Korean stops with a three-way laryngeal contrast by French learners - The role of cue-weighting

Boram LEE
Université Sorbonne Nouvelle - Paris 3, LPP
lee.boram@sorbonne-nouvelle.fr

하나의 말소리에는 말소리의 특성을 결정하는 여러 음향 단서(acoustic cue)가 존재하지만, 청자는 모든 음향적 단서를 바탕으로 소리를 변별하지 않는다. 다시 말해, 하나의 말소리에 여러 음향 단서가 존재할 때 청자는 가중치를 두고 있는 음향 단서에 주의를 기울여 말소리를 변별하는데, 이를 선택적 주의(cue weighting, perceptual weighting, selective attention)라고 한다. 예를 들어, 프랑스어의 유성음 /b/와 무성음 /p/ 변별에서 가장 중요한 음향 단서는 VOT이다. 반면에, 한국어의 경음 /ㄷ/, 평음 /ㅌ/, 격음/ㅍ/를 변별하는 가장 중요한 단서는 VOT와 F0이다. 이처럼 이 선택적 주의는 언어마다 다르게 적용되며, 제2 언어(L2) 음운 습득 시 중요한 역할을 한다.

본 연구는 프랑스인 한국어 L2 학습자의 한국어의 삼중 대립 파열음의 습득 과정과 패턴을 선택적 주의 관점에서 살펴보고자 한다. 이를 위해 종단 연구의 일환으로 프랑스 대학교 한국어과 1학년에 재학 중인 프랑스인 L2 한국어 학습자 20명은 2020년 9월부터 2021년 5월까지 매달 총 8번(12월 제외)의 온라인 지각 실험에 참여하였다. 실험은 음운 식별 실험(Identification task)으로 학습자는 /다/, /타/, /따/ 중에 한 소리를 듣고 3가지 선택지 중 한 음을 선택하였다.

실험 결과, 먼저, 식별의 정확성을 살펴보면 /타/(69%) 와 /따/(52%) 에 비해 /다/(26%) 의 지각이 현저히 낮다. 이는 프랑스 학습자가 경음을 다른 소리와 구별하기 위해 F0 음향 단서에 주의를 기울이지 않음을 시사한다. 더욱이 1년 동안 학습자는 /타/(첫 번째 실험 51% vs. 여덟 번째 실험 69%)와 /따/(첫 번째 실험 34% vs. 여덟 번째 실험 52%)의 식별 정확성이 향상된 것에 반해서 /다/의 식별 정확성(첫 번째 실험 30% vs. 여덟 번째 실험 26%)은 향상되지 않았다. 이러한 결과는 프랑스인 학습자가 한국어의 삼중 대립 파열음을 인지하기 위해 모국어에서 가장 중요한 음향 단서인 VOT에만 의존하고 있음을 알 수 있다. 다음으로 /다/, /타/, /따/의 인지 향상성의 패턴이 다르게 나타났다. /타/의 식별 정확성은 첫 실험부터 51%로 높게 나타났지만, /따/의 식별 정확성은 3번째 실험부터 점차 향상되었다. 이는 프랑스 학습자에게 평음이 가장 쉽게 인지되는 음운이며 이러한 결과를 자/모음 교수 시 활용할 수 있을 것이다.

구두발표 III: 말장애 및 음성의학

좌장: 김지연(우송대)

안면 마스크 착용이 발화의 음향학적 특성에 미치는 영향

김 덕 애, 최 성 희, 이 경 재, 최 철 희
대구가톨릭대학교 언어청각치료학과

Effect of facial mask on the speech acoustics

Deokae Kim, Seong Hee Choi, Kyoungjae Lee, Chul-Hee Choi
Dept. of Audiology & Speech Langauge Pathology, Daegu Catholic University
miracle5034@naver.com, shgrace@cu.ac.kr, kjlee0119@cu.ac.kr, cchoi@cu.ac.kr

전 세계적으로 확산되어 온 코로나바이러스감염증-19로 인하여 마스크 착용이 생활화되었다. 본 연구는 마스크 착용이 일상생활의 대화에서 화자의 음성에 어떠한 영향을 미치는지 살펴보고자 마스크 착용 유무에 따른 음향학적 특성을 살펴보았다. 또한, 현재 많이 사용되고 있는 마스크의 유형에 따라 음향학적 특성의 차이를 비교하였다. 이를 통해 마스크 착용과 종류가 말소리에 미치는 영향을 제공함으로써 일상생활의 대화 시나 음성 평가 시 마스크 착용과 관련된 시사점을 제공하고자 한다.

본 연구에는 음성과 청각에 문제를 보이지 않는 정상 성인 총 40명이 참여하였으며, 마스크 종류는 KF 94, 수술용 마스크, 면 마스크를 사용하였고, 말 과제는 모음 연장 발성과 연결 발화('가을' 문단의 두 번째 문장, 모두 유성음으로 된 문장) 총 3가지로 실시하였다. 음향학적 특성을 살펴보기 위하여 모음 연장 발성에서는 다차원 음성분석기(Multi-Dimensional Voice Program)을 사용하여 변동률 분석을 실시하였으며, Real-Time Pitch(RTP)를 사용하여 성별, 말 과제, 마스크 착용 유무 및 마스크 종류에 따른 강도의 차이를 살펴보았다. 또한, Analysis of Dysphonia in Speech and Voice(ADSV)를 사용하여 모음연장발성과 연결발화 시 마스크 착용 유무와 마스크 종류에 따른 음향학적 특성을 비교하였다.

그 결과, 모음 연장 발성과제에서 두 성별 집단 모두 마스크 착용 전보다 후에 Jitter, Shimmer, NHR 측정치가 유의하게 증가하였다. 따라서, 마스크 착용이 음성 신호를 저하시키며, 음질에 영향을 주는 것으로 나타났다. 또한, 세 가지 발화과제에서 모두 마스크 착용 후의 평균 강도가 착용 전에 비해 유의하게 낮아졌다. 이러한 결과는 마스크 착용이 화자의 강도나 음질 조절에 영향을 줄 수 있음을 시사한다. 특히, 본 연구에서는 마스크 종류에 따라 강도에 통계적으로 유의한 차이를 보였는데, 수술용 마스크에서 모음 및 연결발화 모두 강도가 가장 컸으며, KF 94 마스크에서 음성 강도가 가장 작았다. 캡스트럼 분석 결과, CPP는 모음 연장 발성 과제에서는 마스크 착용 및 종류에 따라 영향을 받지 않았으나, 연결 발화 과제에서는 모두 착용 전보다 착용 후 CPP가 증가하였다. 또한, 연결발화과제에서는 두 문장과제 모두 마스크 착용 전보다 착용 후 CPP F_0와 CPP SD, CPP Max, CPP Min이 모두 유의한 차이가 있었다. 이와 더불어 마스크 종류에 따라 CPP 값에 유의한 차이를 보였는데, 모든 말 과제에서 면 마스크가 가장 낮은 CPP 값을 보였다. 말 과제별로 살펴보았을 때, 마스크 종류에 따른 말 과제의 CPP 값은 유의한 차이가 없었으나, 마스크 종류에 따른 말 과제의 L/H Ratio 값은 유의한 차이가 있었으며, KF 94, dental, 면 마스크 순으로 KF 94가 가장 높았다.

따라서 본 연구결과, 마스크 착용 유무에 따라 의사소통에 영향을 미칠 수 있음을 시사하였다. 또한, KF 94 마스크는 비말차단을 목적으로 가장 효과가 있다고 알려져 있으나, 수술용 마스크나 면 마스크에 비하여 말소리의 음향학적 특징에 영향을 주므로 의사소통에 더 영향을 미칠 수 있음을 시사하였다.

변성기 남성의 발화범위 및 음성범위 프로파일

김 재 옥*, 이 승 진**
*강남대학교 교육대학원 언어치료교육전공, **한림대학교 언어청각학부

Speech & Voice Range Profile in Puberty Men

Jaeock Kim*, Seung Jin Lee**
*Major in Speech Pathology Education, Graduate school of Education, Kangnam University
**Division of Speech Pathology and Audiology, Hallym University
jaeock@gmail.com, sjl@hallyn.ac.kr

음역대를 평가하는 음성범위 프로파일(Voice Range Profile, VRP)은 최저부터 최고 범위의 음도와 음성강도의 정보를 통합하여 살펴보는 방법으로 다양한 대상자들의 최대발성능력을 평가한다. VRP는 일반적으로 모음을 연장 발성하는 동안 평가하기 때문에 발화를 산출하는 동안의 발성능력을 살펴보기 어렵다[1]. 이에 음역대를 평가할 때는 기능적 말산출 수행능력을 측정할 수 있는 발화범위 프로파일(Speech Range Profile, SRP)을 함께 살펴보는 것이 바람직하다. 특히 사춘기 이후 후두의 구조적인 변화로 인해 음도와 강도의 변화가 있는 변성발성 과정에 있는 남성의 SRP와 VRP가 변성 발성 전이나 후에 비해 어떠한 차이가 있는지 살펴볼 필요가 있다. 이에 최근 개발되어 타당성이 검증된 『불이야』문단을 이용한 SRP[2]와 축약된 VRP[3]를 사용하여 SRP와 VRP를 측정하고 비교하였다.

본 연구에서는 10~18세의 변성 전 20명(평균연령 10.20세), 변성기 13명(평균연령 13.00세), 변성 후 14명(평균연령 16.86세)의 남성을 대상으로 SRP와 VRP을 비교한 결과, SRP는 기본주파수관련 변인인 최고기본주파수($F0_{max}$), 최저기본주파수($F0_{min}$), 1기본주파수범위($F0_{range}$)가 집단 간에 유의한 차이를 보였고, VRP는 기본주파수관련 변인(F0표, $F0_{min}$n, $F0_{range}$) 및 음성강도관련 변인인 최대음성강도(I_{max}), 최소음성강도(I_{mi}n) 및 음성강도범위(I_{range}) 모두에서 집단 간에 유의한 차이가 있었다. 즉 변성기를 전후로 연령이 증가함에 따라 SPR와 VRP 모두 음도와 음성강도 모두 최대음역대는 높아지는 반면, 최저음역대는 감소함을 알 수 있었다.

감사의 글

본 연구는 2018년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(2018S1A5A2A03032902).

참고문헌

[1] D'Alatri, L. & Marchese, M. R. (2014). The speech range profile (SRP): an eary and useful tool to assess vocal limits. *Acta otorhino-laryngologica italica, 34*(4), 253-258.

[2] 김재옥, 이승진(2019). 발화범위 프로파일 과제 개발 및 타당성 검증. *말소리와 음성과학, 11*(3), 77-87.

[3] 정원정(2018). *음성장애 환자를 위한 축약된 음성범위프로파일(Voice Range Profile, VRP) 검사법 개발과 타당도*. 대구가톨릭대학교 석사학위논문.

발화 환경에 따른 한국어 모음 /오, 우, 으/의 포먼트 특성*

박 지 연, 성 철 재
충남대학교 언어병리학과, 충남대학교 언어학과

Formant characteristics of Korean vowels /ㅇ, ㅜ, ㅡ/ according to the utterance environment

Jiyeon Park, Cheoljae Seong
Dept. of Speech & Language Pathology, Linguistics, Chungnam National University
bpn0525jy@gmail.com, chseong49@gmail.com

모음 /오/는 후설 원순 중모음 또는 중고모음, /우/는 후설 원순 고모음으로 다른 모음들과 달리 원순성이 특징적이다. 우리말에서 /오/와 /우/는 후설성의 정도가 유사하되 높낮이의 차이가 있는 소리로 규정되어 있다(윤지현 & 성철재, 2013). /오/와 /우/는 입술을 동그랗게 말고 돌출시켜 조음하기 때문에 전체적인 성도의 길이가 길어지며 다른 단모음에 비해 포먼트 주파수가 낮게 측정된다. 또한 이러한 조음적 특성으로 다른 단모음보다 조음하기 위한 노력이 더 필요하다. 일반적으로 단순히 모음만 산출하는 환경보다는 발화 환경이 복잡해질수록 조음의 편의성을 위해 이러한 노력이 완화되는 경향을 관찰할 수 있다. 따라서 발화 환경에 따라 /오, 우/의 포먼트 특성을 살펴보고, 포먼트 산점도에서 비교적 근접한 중설고모음 /으/와 함께 비교하고자 하였다.

연구 방법은 다음과 같다. 정상 음성을 산출하는 20~30대 성인 80명(남 40명, 여 40명)을 대상으로 음성을 수집하였다. 녹음 과제는 {V}, {V+다}, 운반구('나는 __를 좋아해요'에 모음을 포함하여 발화하도록 함.) 과제로 총 3가지 유형의 과제를 통해 모음 /오/, /우/, /으/의 데이터를 수집하였다. Praat을 이용하여 과제 유형마다 세 가지 모음의 포먼트(F1, F2)를 측정하였다. 포먼트 측정 시 박지연&성철재(2019)의 변수 세팅을 참조하여 모음에 따라 최대 포먼트, 포먼트 개수 변수를 설정하였다. 그리고 SPSS 26을 이용하여 측정한 F1, F2가 발화 환경과 모음에 따라 차이가 있는지 이원분산분석을 실시하였다.

<표 1> 모음별 발화 환경에 따른 F1 및 F2의 기술통계

모음	발화 환경	남성		여성	
		F1(Hz)	F2(Hz)	F1(Hz)	F2(Hz)
/오/	V	361.5(±40.5)	650.8(±71.1)	407.3(±51.5)	684.9(±114.0)
	V+다	361.3(±47.4)	660.9(±261.4)	405.6(±46.4)	689.1(±123.4)
	운반구	371.8(±68.2)	640.3(±111.2)	398.1(±40.1)	677.9(±101.1)
/우/	V	323.7(±39.9)	766.8(±93.2)	403.0(±46.6)	796.7(±98.2)
	V+다	331.5(±33.0)	762.5(±148.4)	386.8(±38.7)	785.5(±112.7)
	운반구	333.1(±61.1)	794.8(±213.6)	377.9(±37.8)	796.2(±117.7)
/으/	V	356.6(±45.7)	1387.9(±247.4)	424.7(±41.7)	1531(±184.5)
	V+다	356.6(±48.1)	1310.3(±274)	424.1(±42.0)	1510.6(±220.5)
	운반구	346.6(±61.9)	1247.6(±308.6)	415.0(±35.6)	1477.1(±278.4)

* 본 연구는 2021학년도 대학혁신지원사업으로 지원된 연구임.

<표 1>에 모음별 발화 환경에 따른 F1 및 F2의 평균과 표준편차를 제시하였다. 통계 분석 결과, 남성의 경우 모음에 따라 F1 측정치에 통계적으로 유의한 차이가 있는 것으로 나타났다(p=.000). 반면 발화 환경에 따라서는 F1의 차이가 없었고(p=.875), 모음과 발화 환경 상호작용에서도 차이가 없는 것으로 나타났다(p=.612). Bonferroni를 적용한 사후 검정 결과, {V} 환경에서 /오/와 /우/, /우/와 /으/ 간 F1 측정치에 각각 차이가 있었으며(p<.05), {V+다} 환경에서 /오/와 /우/ 간 차이가 있는 것으로 나타났다(p<.05). 운반구 환경에서도 /오/와 /우/ 간 차이가 통계적으로 유의하였다(p<.05). 모음 별 발화 환경 간 차이도 살펴보고자 했으나 모든 경우에서 차이가 없는 것으로 나타났다. F2 역시 모음에 따라 통계적 차이가 유의하였다(p=.000). 발화 환경에 따라서는 F2의 차이가 없었으며(p=.315), 모음과 발화 환경 상호작용에서도 유의한 차이가 없는 것으로 나타났다(p=.115). 사후 검정 결과, {V} 환경과 운반구 환경에서 /오/와 /우/, /우/와 /으/, /으/와 /오/ 모든 경우에서 F2 값에 통계적으로 유의한 차이가 있었다(p<.05). {V+다} 환경에서는 /오/와 /으/, /우/와 /으/ 간 F2 값에 차이가 있는 것으로 나타났다(p<.05). /우/와 /오/ 모음에서 발화 환경 간 비교를 실시했을 때 모든 경우에서 통계적 차이가 없었으나 /으/의 경우, {V}와 운반구 환경 간 F2 측정치에 통계적으로 유의한 차이가 있는 것으로 나타났다(p<.05).

여성의 경우 모음에 따라 F1 측정치에 통계적으로 유의한 차이가 있었으며(p=.000), 발화 환경에 따라서도 차이가 있는 것으로 나타났다(p<.05). 반면 모음과 발화 환경 상호작용에서는 차이가 없었다(p=.673). Bonferroni를 적용한 사후 검정 결과, {V} 환경에서 모음 간 모든 경우에서 통계적으로 유의한 차이가 없었으나 {V+다} 환경과 운반구 환경에서 /우/와 /으/ 간 F1 측정치에 각각 차이가 있는 것으로 나타났다(p=.000). /오/와 /으/ 모음에서 발화 환경 간 비교를 실시했을 때 모든 경우에서 통계적 차이가 없었으나 /우/의 경우, {V}와 운반구 환경 간 F1 값에 차이가 통계적으로 유의하였다(p<.05). F2에 대한 검정을 실시했을 때, 모음에 따라 F2 측정치에 유의한 차이가 있는 것으로 나타났다(p=.000). 그러나 발화 환경에 따라서는 차이가 없었고(p=.617), 모음과 발화 환경 상호작용에서도 차이가 없는 것으로 나타났다(p=.822). 사후 검정 결과 {V}, {V+다}, 운반구 환경에서 /오/와 /우/, /우/와 /으/, /으/와 /오/ 모든 경우 통계적으로 유의한 차이가 있었다(p<.05).

본 연구에서 3가지 유형의 발화 환경({V}, {V+다}, 운반구)을 제시하여 모음 /오/, /우/, /으/의 F1, F2를 측정하고 발화 환경에 따라 포먼트의 차이가 있는지 확인해보았다. 남성과 여성 모두 모음에 따라 F1, F2 측정치에서 통계적으로 유의한 차이가 있었으나 발화 환경에 따라서는 성별, 모음별로 다른 양상을 확인하였다. 자연발화에서 남성, 여성의 단모음을 분석한 김순옥 & 윤규철(2015), 윤규철 & 김순옥(2015) 연구 결과와 F2에 대한 모음 집단 간 차이에 대한 결과는 상당히 유사하나 F1에 대해서는 차이가 있었다. 이는 발화를 수집하는 방법이 다르다는 것을 감안하고 해석해야 할 것이다. 추후 /오/와 /우/, /우/와 /으/, /으/와 /오/ 간 discrimination 인지실험을 실시하여 청지각 실험 결과를 함께 종합하여 발표할 예정이다.

참고문헌

[1] 김순옥, & 윤규철. (2015). 한국어 자연발화 음성코퍼스의 남성 모음 포먼트 연구. 말소리와 음성과학, 7(2), 95-102.

[2] 박지연, & 성철재. (2019). 성인 포먼트 측정에서의 최적 세팅 구현: Praat software 와 관련하여. 말소리와 음성과학, 11(4), 97-108.

[3] 윤규철, & 김순옥. (2015). 한국어 자연발화 음성코퍼스의 남녀 모음 포먼트 비교 연구. 말소리와 음성과학, 7(2), 131-138.

[4] 윤지현, & 성철재. (2013). F1/F2 의 변화가 한국어/오/,/우/모음의 지각판별에 미치는 영향. 말소리와 음성과학, 5(3), 39-46.

구두발표 III:
음성공학

좌장: 김경화(대검찰청)

전체 맥락과 깊이별 분리 합성곱을 이용한 Conformer 기반 음성 인식기의 순방향 모듈 개선

정 승 훈, 김 홍 국
광주과학기술원 AI대학원

Modification of Feed-Forward Module of Conformer-Based Speech Recognition Using Global Context and Depthwise-Separable Convolution

Seunghun Jeong, Hong Kook Kim
Graduate School of AI, GIST
zldzmfopq12@gm.gist.ac.kr, hongkook@gist.ac.kr

최근 Transformer 기반 음성인식 모델은 Transformer에 합성곱 레이어를 적용한 Conformer를 적용함으로써 개선된 성능을 보였다. 또한, 합성곱 레이어 기반의 Squeeze-and-Excitation (SE)과 깊이별 분리 합성곱(Depthwise-Separable Convolution) 모듈을 적용한 ContextNet은 합성곱 레이어 기반 음성인식 모델의 성능을 개선하였다.

이러한 연구 동향에 힘입어, 본 논문에서는 Transformer 기반 음성 인식 모델에 이와 같은 합성곱 레이어 기반의 모듈을 적용하여 음성 인식 성능이 향상된 모델을 제안하였다. 우선, Transformer 기반의 Joint Connectionist Temporal Classification-Attention 인코더-디코더 모델을 기반으로, 인코더 부분을 12개의 Conformer 블록으로 변경한다. 이때, 각 Conformer 블록은 채널 수와 헤드 수가 가각 256과 4로 구성된 자기-멀티헤드 주의 모듈, 합성곱 모듈, 두 개의 순방향 모듈로 구성된다. 또한, 합성곱 모듈의 합성곱층은 필터크기는 15로 설정하였으며, 선형층은 깊이별 분리 합성곱 모듈로 치환하였다. 또한 Conformer의 순방향 모듈이 발화의 전체 맥락을 활용할 수 있도록 순방향 모듈 말단부에 SE 모듈을 추가하였다.

제안된 음성인식 모델의 성능평가를 위해서 Zeroth-Korean 데이터베이스를 사용하였으며, Transformer 기반의 음성인식 모델과 문자오인식률(CER)을 비교하였다. 입력 특징으로는 Transformer 기반 및 제안된 모델 모두 80차 Mel-filterbank를 사용하였다. 또한, 기존의 Transformer 기반 모델은 채널 수와 헤드 수를 각각 인코더와 디코더에서 512와 8로 증가시켰다. 성능평가 결과, 기존의 Transformer 기반 모델은 1.2%의 CER을 보이는 반면, 본 논문에서 제안된 전체 맥락과 깊이별 분리 합성곱을 이용한 Conformer 기반의 모델은 0.8%의 CER로 상대적으로 33%를 개선하였다. 게다가, 제안된 모델의 파라미터 수는 기존의 Transformer 기반 모델 대비 44%로 경량화되었다.

Acknowledgment

본 연구는 문화체육관광부 및 한국컨텐츠진흥원의 2020년도 문화기술연구개발 지원사업(R2020060002)으로 수행되었음.

영어-한국어 대화체 자동통역을 위한 Cascade 및 End-to-End 접근 방식 비교

방 정 욱, 이 민 규, 윤 승, 김 상 훈
한국전자통신연구원, 복합지능연구실

A Comparison of Cascade and End-to-End Approaches for English-Korean Conversational Speech Translation

Jeong-Uk Bang, Min-Kyu Lee, Seung Yun, Sang-Hun Kim
Electronics and Telecommunications Research Institute (ETRI)
{jubang, mk, syun, ksh}@etri.re.kr

본 연구에서는 영어-한국어 대화체 자동통역(speech translation; ST)을 위한 두 가지 접근법을 비교한다. 먼저, Cascade 자동통역(CAS-ST)은 영어 음성인식(automatic speech recognition; ASR) 모듈과 영어-한국어 기계번역(machine translation; MT) 모듈을 각각 구축한 다음, ASR 모듈에 영어 음성을 입력하여 나온 영어 전사문을 다시 영어-한국어 MT 모듈로 입력하여 최종 한국어 번역문을 생성한다. 반면에, 최근 제안된 End-to-End 자동통역(E2E-ST)은 ASR 및 MT 모듈을 하나의 ST 모델로 통합한 모델을 사용하여, 입력된 영어 음성으로부터 한국어 번역문을 곧바로 생성한다. E2E-ST는 ASR 모듈의 오류가 MT 모듈로 전파되는 문제를 막을 수 있어 많은 기대를 받고 있지만, 세계적으로도 아직 연구 초기 단계이며, 특히 영어-한국어 대화체 음성에 관해서는 많은 연구가 필요한 상황이다.

영어-한국어 대화체 자동통역 실험을 위해, 약 2천 시간의 대화체 영어 음성, 영어 전사문 및 한국어 번역문으로 구성된 통역 데이터를 사용한다. 실험에서는 트랜스포머(Transformer) 구조에서 CAS-ST와 E2E-ST의 성능을 비교하고, 일반적인 E2E-ST 성능개선 기법인 다중작업 학습(multi-task learning)과 사전학습(pre-training)을 채택하여 영어-한국어 E2E-ST에서의 효과를 검증한다. 마지막으로, 추가적인 학습 데이터를 사용하여 대용량 통역 데이터에서의 CAS-ST와 E2E-ST의 성능 변화를 확인한다. 2천 시간의 학습 데이터를 사용한 실험 결과에서 E2E-ST는 CAS-ST보다 낮은 성능을 보이며, 다중작업 학습보다 사전학습이 E2E-ST의 성능개선에 더욱 효과적임을 보여준다. 또한, 대량의 학습 데이터에서 E2E-ST는 CAS-ST와 근접한 품질의 번역문을 생성할 수 있음을 보여준다.

* 본 연구는 한국전자통신연구원 연구운영비지원사업의 일환으로 수행되었음[21ZS1100, 자율성장형 복합인공지능 원천기술 연구].

딥러닝 기반 단일 클래스 분류를 사용한 한국어 키워드 검출

이 승 현, 박 형 민
서강대학교 전자공학과

Korean Key-Word Spotting Using Deep One-Class Classification

Seung-Heyon Lee, Hyung-Min Park
Dept. of Electeronic Engineering, Sogang University
ee_seung@u.sogang.ac.kr, hpark@sogang.ac.kr

키워드 검출이란 사용자가 미리 정의된 특정 단어나 어구를 발화했을 때, 이를 검출하여 장치를 구동시키는 기능이다. 보다 긴 배터리 사용 시간이 관건인 모바일 장치의 특성 상, 상당한 양의 메모리, 연산 자원을 소모하는 음성인식 등의 기능을 항상 켜놓을 수 없기 때문에 키워드 발화를 기준점으로 삼아야 할 필요성이 있다. 이 때, 키워드 검출 기능 자체는 항상 켜진 상태를 가정하기 때문에, 제한된 리소스에서 서버를 거치지 않고, 장치 내에서 구현될 수 있어야 한다. 기존의 키워드 검출 연구들은 대체로 키워드를 포함하는 여러 개의 클래스에 대한 분류로 해당 문제를 접근하였다. 하지만, 이러한 방식으로 학습된 모델들은 학습 과정에 포함되지 않은 키워드와 다른 소리에 대해서 높은 오검출률을 보인다는 한계가 있다. 또한, 키워드 소리에 대비해 키워드와 다른 소리의 데이터 양이 매우 많아서 이진 분류로 학습할 경우 데이터 불균형 문제가 크게 대두된다. 이러한 문제점들을 해결하고자, 우리는 키워드 검출 문제를 오직 키워드 소리로만 학습하는 준지도 딥러닝 방법으로 접근하고자 한다. 보다 구체적으로, 이상 검출(anomaly detection)에 주로 사용되는 단일 클래스 분류 모델인 deep SVDD (suppor vector data description) 모델을 활용하여 한국어 키워드 검출 모델을 개발하였다.

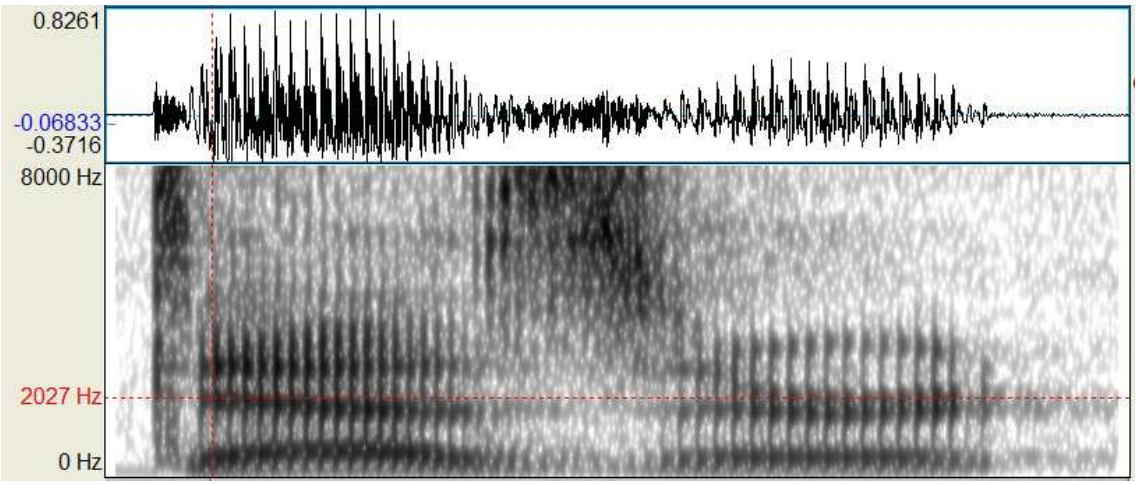
키워드: 키워드 검출, 단일 클래스 분류, 딥러닝

사사: 본 연구는 행정안전부/국토교통과학기술진흥원의 지원으로 수행되었음
(과제번호 21PQWO-153358-03)

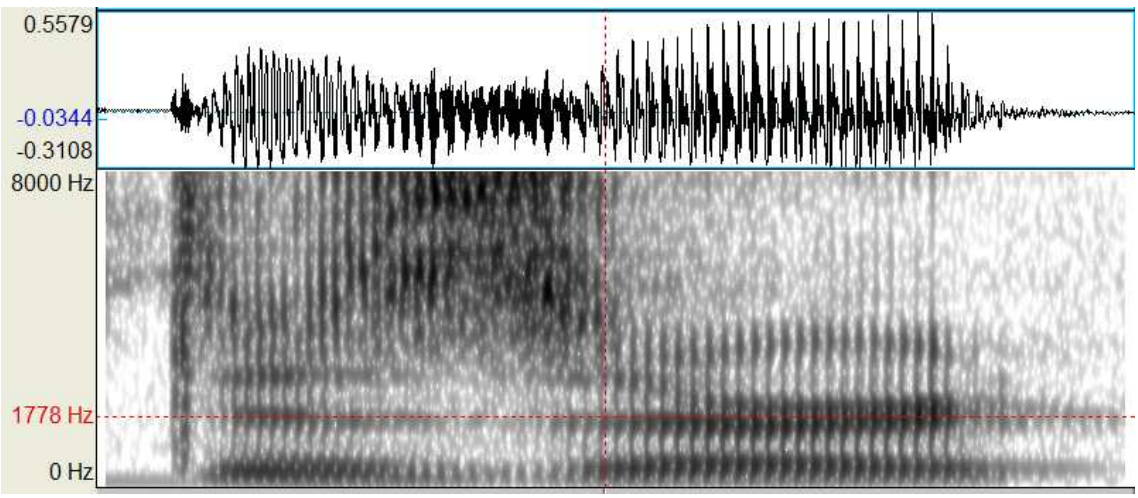
Proceedings

of the 2021 Fall Conference
of the Korean Society of Speech Sciences

(a)



(b)



November 19 & 20, 2021
The Korean Society of Speech Sciences