pISSN 2005-8063
eISSN 2586-5854
2024. 12. 31.
Vol.16 No.4
pp. 63-72

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2024.16.4.063

CNN-based automatic classification of stuttering: Detection of repetitions and prolongations in stuttered speech

Jin Park[1] · Chang Gyun Lee[2,*]

[1]*Department of Speech Language Rehabilitation, Catholic Kwandong University, Gangneung, Korea*
[2]*Department of Business Administration, Catholic Kwandong University, Gangneung, Korea*

Abstract

This study aims to develop and validate a CNN-based deep learning algorithm to automatically classify repetition and prolongation disfluency types in stuttered speech. The LibriStutter dataset was used, and the speech data were pre-processed into mel-frequency cepstral coefficients (MFCCs) to train a convolutional neural network (CNN) model. With optimized hyperparameters using the GRID search method, the model achieved high performance, with an accuracy of 0.9912 and a loss of 0.0544. Among the fluent speech and four disfluency types (sound repetitions, word repetitions, phrase repetitions, and prolongations), the model demonstrated strong classification performance for sound repetitions and prolongations, while the classification accuracy for word and phrase repetitions was comparatively lower, indicating areas for future improvement. This study demonstrates the feasibility of automated stuttering disfluency assessment and suggests further research to enhance clinical applicability by incorporating diverse datasets and multi-modal approaches.

Keywords: stuttering, artificial intelligence, convolutional neural network, repetitions, prolongations

1. 서론

말더듬은 반복, 연장, 막힘 등의 비유창성이 구어의 흐름을 방해하면서 유창성이 저하되는 말장애의 일종이다(van Riper, 1972). 반복은 음소, 음절, 단어의 일부나 전체가 불수의적으로 여러 번 반복되는 현상을 의미하며(예, “사사사사과를 먹었어”), 연장은 특정 음소가 비정상적으로 길게 이어지는 현상(예, “ㅅ----아과를 먹었어”)이며, 막힘은 말하려는 순간 소리가 나지 않은 ‘끊김 현상’을 말한다(예, “사....과를 먹어”). 이러한 비유창성은 말더듬의 기본적 특성이며, ‘일차행동’ 혹은 모든 말더듬 화자에게 나타나는 특성이라는 의미에서 ‘공통행동’이라고도 한다(Shim et al., 2022).

전통적으로 말더듬 평가는 비유창성의 빈도나 비율을 기반으로 청지각적 판단을 통해 이루어진다. 하지만 이러한 평가는 시간이 많이 소요되며, 평가자간 신뢰도 문제도 발생할 수 있다(Kully & Boberg, 1988; Yaruss, 1997). 따라서 최근에는 인공지능을 이용한 자동화된 말더듬 평가 방법이 주목을 받고 있다. 인공지능을 기반으로 한 분류 알고리즘

* kdmis@cku.ac.kr, Corresponding author.
Received 14 November 2024; Revised 4 December 2024; Accepted 4 December 2024

을 통해 말더듬을 좀 더 신속하고 신뢰성있게 식별하는 기술이 개발 중이며, 다양한 연구에서 유의미한 성과가 도출되고 있다(Alnashwan et al., 2023; Barrett et al., 2022).

인공지능 기반의 말더듬 자동 식별은 대체로 음성 데이터 수집, 전처리(pre-processing)와 라벨링(labelling), 알고리즘 모델 수립, 기계학습 실행, 성능 및 효과 검증을 포함한 5단계로 진행된다(Sheikh et al., 2022). 구체적으로, 먼저 자발화(spontaneous speech)나 읽기 등을 통해 말더듬 화자의 음성 데이터를 수집한다. 전처리 단계에서는 수집된 음성 데이터를 전사(transcription)하고, 음절 또는 단어와 같은 특정 언어 단위로 구분하고, 개별 말더듬에 대한 라벨링 작업을 수행한다. 이후 주파수, 진폭, 길이, 포만트 주파수, MFCCs (mel frequency cepstral coefficients) 등과 같이 개별 말더듬을 특정할 수 있는 음향적 특징을 추출하여, 이를 바탕으로 말더듬 식별 모델을 수립한다. 마지막으로 추가 데이터셋을 통해 수립된 식별기의 성능을 검증하여 실효성을 평가한다.

최근까지 말더듬 자동 식별에 대해 다양한 인공지능 알고리즘을 적용한 연구들이 진행되어 왔다. 체계적 문헌고찰 연구인 Barrett et al.(2022)에서는 보고된 총 27건의 연구 사례 중 ANN(artificial neural network)와 SVM(support vector machine)이 가장 빈번하게 사용된 알고리즘으로 확인된다. ANN은 총 12개의 연구에서 활용되었으며, 주로 복잡한 음성 신호의 패턴을 학습하기 위한 다층 신경망 구조를 채택하였다. 예를 들어, Howell & Sackin(1995)에서는 ANN을 통해 auto-correlation coefficients와 vocoder coefficients 등의 특징을 입력으로 사용하여, 반복과 연장을 82%의 정확도로 식별하였다. Swietlicka et al.(2009)은 MLP(multi-layer perceptron)과 RBF(radial-basis function) 기반의 ANN을 비교하여 유창함과 비유창함을 92%의 정확도로 식별하였다. 이러한 결과는 ANN이 음성 신호의 복잡한 패턴을 인식하는 데 유효함을 시사한다. 한편, SVM은 총 8개의 연구에서 활용되었으며, 주요 특성으로 고차원의 특징 공간에서 데이터 포인트 간의 최대 마진을 찾는 기능을 이용하였다. 예로, Pálfy & Pospíchal(2011)은 여러 커널을 활용하여 SVM 성능을 평가했으며, sigmoid 커널 사용 시 99.05%의 높은 정확도를 보고하였다. 또한, Ravikumar et al.(2009)에서는 SVM과 ANN을 비교 분석하여 SVM이 98.3%의 정확도로 우수한 성능을 나타냄을 보고하였다. 또한 말더듬 관련 음향적 특징으로는 총 17개의 연구에서 채택된 MFCCs가 가장 빈번하게 사용되었다. 기본적으로 MFCCs는 음성 신호의 주파수 대역 정보를 캡처하여 유창성과 비유창성 구별에 기여한다. Ravikumar et al.(2008)에서는 MFCCs를 바탕으로 ANN 모델을 학습하여 유창성과 비유창성을 83%의 정확도로 식별하였다. Fook et al.(2013)은 MFCCs와 LPCs(linear prediction coefficients), LPCCs(linear prediction cepstral coefficients) 등을 SVM에 적용해 반복과 연장을 96.2%의 정확도로 식별하였다.

나아가 기계학습을 통한 말더듬 자동 식별에 있어 음성 데이터의 충분한 확보가 중요한데, UCLASS(University of College London's Archive of Stuttered Speech)나 LibriStutter와 같은 공개된 데이터셋을 이용하기도 하였다. 예를 들어, Mahesha & Vinod(2013)는 UCLASS를 이용해 MFCCs를 추출하여 SVM 모델을 바탕으로 음절 반복, 단어 반복, 연장에 대한 자동 식별 방식을 개발하였는데, 각각 93%, 73%, 100%의 식별 정확도를 보고하였다. 또한 Kourkounakis et al.(2020)은 LibriStutter를 이용해 순환신경망(recurrent neural network) 모형을 바탕으로 음소 반복, 단어 반복, 구 반복, 연장을 각각 79.80%, 92.52%, 92.52%, 89.44%의 정확도로 식별하였다. 이처럼 비교적 많은 표본으로 구성된 데이터셋을 통해 수립된 말더듬 자동 식별은 내구성(robustness)과 일반화에 있어 상대적으로 큰 장점을 가진다(Jo et al., 2022).

최근 인공신경망과 관련된 기술의 발전과 공통 데이터 획득이 가능해지면서 합성곱층 신경망(convolutional neural network, CNN)을 통한 장애 음성 식별 연구 사례들이 보고되었다(Bhushan et al., 2021; Fang et al., 2019; Jo et al., 2022; Park & Lee, 2023; Wang et al., 2019). 기본적으로 CNN은 이미지와 같은 공간적 데이터 학습에 특화된 DNN (deep neural network)의 응용 알고리즘으로, 장애 음성의 자동 분류 영역에서는 합성곱층(convolutional layer)과 풀링층(pooling layer)을 통해 음성 데이터의 시각적 특징을 추출하고, 완전 연결층(fully-connected layer)을 통해 각 유형별 활성화 함수(activation function)를 생성해 이를 학습하는 방식으로 활용되고 있다(Goodfellow et al., 2016; Lee, 2017). Jo et al.(2022)은 CNN을 기반으로 정상 및 후두장애 음성에 대한 자동 식별 방식을 개발하였으며, 최고 85%의 식별률을 기록하였다. 또한 Wang et al.(2019)은 CNN을 기반으로 구개열 환자의 과대비성을 위한 자동 식별 방식을 개발하였으며, 95%의 정확도를 보고하였다. 말더듬 식별과 관련해 Bhushan et al.(2021)은 CNN 모델을 바탕으로 자동 식별 방법을 개발하였으며, 유창함과 비유창함을 89%의 식별률로 분류하였다. Park & Lee(2023)는 9명의 말더듬 성인의 읽기 샘플을 통한 음성 데이터를 바탕으로 CNN 기반의 말더듬 비유창성에 대한 자동 식별 방식을 개발하였으며, 반복은 71%, 막힘의 경우는 75%의 식별률을 기록하였다.

본 연구에서는 기본적으로 CNN 알고리즘을 기반으로 말더듬에 대한 자동 식별 방법을 개발해 보고자 한다. 단순히 유창함과 말더듬 간의 이분적 구분에 초점을 맞춘 이전 연구(Bhushan et al., 2021)와는 달리 개별 말더듬 비유창성 유형에 대한 자동 식별 방법을 개발하고자 한다. 또한 막힘의 경우, 비교적 짧은 '끊김 현상' 또는 시각적 긴장(tension)으로도 나타날 수 있기에 음성 데이터만을 가지고 정확한 파악이 어려울 수 있다(Guitar, 2019). 따라서 막힘의 정확한 식별을 위해서는 음성 데이터뿐 아니라 턱이나 입술 등과 같은 조음기관의 긴장을 보여주는 시각적 단서를 포함한 종합적 판단이 필요하다(Altinkaya & Smeulders, 2020). 이러한 점을 고려해 본 연구에서는 막힘을 제외한 말더듬 비유창성, 즉 반복과 연장에 대한 자동 식별 방식을 개발하고자

한다. 더불어 비교적 적은 음성 데이터셋을 통한 이전 연구(Park & Lee, 2023)와는 달리 일정량 이상의 데이터셋을 활용해 수립하고자 하는 말더듬 자동 식별 방식의 견고성이나 일반화도 증대시키고자 한다.

본 연구는 CNN 알고리즘을 통한 이전 연구(Bhushan et al., 2021; Park & Lee, 2023)의 제한점을 인지하고 말더듬 비유창성 유형 가운데 (음소 반복, 단어 반복, 구 반복을 포함한) 반복과 연장에 대한 CNN 기반의 자동 식별 모델을 개발하고자 한다. 또한 식별 모델의 견고성이나 일반화를 고려해 총 50명의 말더듬 화자의 20시간 길이의 음성 데이터로 구성된 LibriStutter 데이터셋(Kourkounakis et al., 2021)을 사용하고자 한다. 본 연구를 통해 보다 식별 정확도와 신뢰도가 높은 말더듬 자동 식별 기술의 개발과 함께 이를 통해 고도화된 말더듬 평가 및 중재 관련 서비스가 창출되기를 기대한다.

2. 연구방법

2.1. 음성 데이터 개요

본 연구에서 활용한 음성 데이터는 Queen's University의 공개 LibriSpeech의 ASR(automatic speech recognition) 말뭉치인 LibriStutter 데이터셋(Kourkounakis et al., 2021)이다. 본 데이터셋은 총 50명(남성 23명, 여성 27명)의 말더듬 화자의 총 20시간 길이의 합성된(synthetic) 음성 샘플로 구성되어 있다. 그리고 유창함(즉, 말더듬이 없음)과 아울러 총 4가지의 말더듬 유형(반복과 연장)에 대해 표 1과 같이 분류되어 있다.

표 1. LibriStutter 데이터셋 구성
Table 1. The composition of LibriStutter dataset

Code	Types	Example	Frequency
0	유창(clean)	No stutter (i.e., fluent)	135,092
1	음소 반복 (sound repetition)	“th-th-this”	1,606
2	단어 반복 (word repetition)	“why why”	1,597
3	구 반복 (phrase repetition)	“I know I know that”	1,537
4	연장 (prolongation)	“whoooooo is there”	1,564

LibriStutter 데이터셋은 분류 레이블이 포함된 파일, 각 분류 데이터별 음성 파일, 시간에 따라 정렬된 전사본이 포함되어 있고, 음성 데이터는 FLAC(free lossless audio codec) 파일로 저장되어 있고 샘플링 레이트(sampling rate, sr)는 22 kHz로 설정되어 있다. 비유창성 유형의 음성샘플 라벨링별 빈도는 유창(clean)의 경우 135,092개, 음소 반복(sound repetition)은 1,606개, 단어 반복(word repetition)은 1,597개, 구 반복(phase repetition)은 1,537개, 연장(prolongation)은 1,564개로 나타났다.

2.2. 음성 데이터 전처리

기본적으로 음성 데이터는 진폭(amplitude)과 시간(time)으로 구분되는 파형 형태의 데이터로 기록되고 연속형 데이터로 고차원의 여러 주파수가 생성된다. 본 연구에서 음성 데이터 전처리는 라벨링된 음성 데이터를 기준으로 MFCCs 이미지 데이터로 변환하였다. MFCCs는 음성 데이터의 음향적 특징을 추출하여 나타내는 수치로서 이를 시각화된 이미지로 전처리하고자 하였다. 데이터 전처리 시 MFCCs의 데이터 특징의 개수를 정해주는 파라미터인 n_mfcc를 100으로 설정하였다. 그리고 시간 영역 신호를 주파수 영역으로 변화하기 위해 샘플링 레이트(sr)는 원본과 같이 22 kHz로 설정하였고 음성 데이터의 크기 파라미터인 n_fft는 자연어 처리 시 일반적으로 음성을 25 ms 크기를 사용하기 때문에 sr에 frame_length를 곱한 값을 반영하여 550으로 설정하였다. 다음으로 fft창 사이의 겹침을 나타내는 매개변수인 hop_length는 10 ms를 기본으로 하고 sr을 반영하여 220으로 설정하였다. Python Librosa 라이브러리를 사용하여 라벨링된 음성 데이터의 이미지 전처리 후 그림 1과 같이 이미지 파일로 저장하였다.

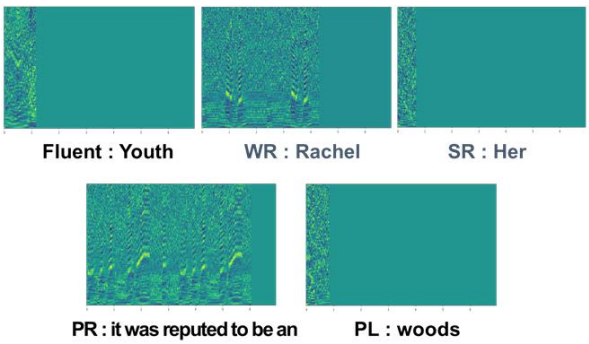
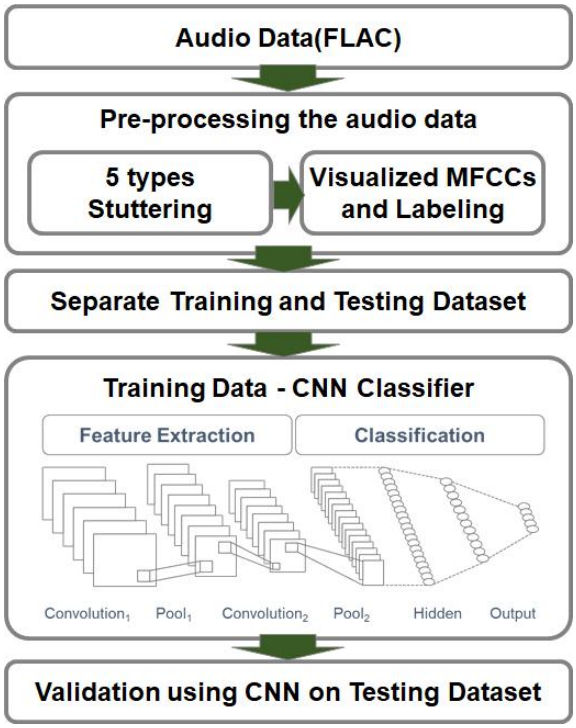


그림 1. 음성 데이터 MFCCs(mel frequency cepstral coefficients) 이미지
Figure 1. Mel frequency cepstral coefficients (MFCCs) image of speech data

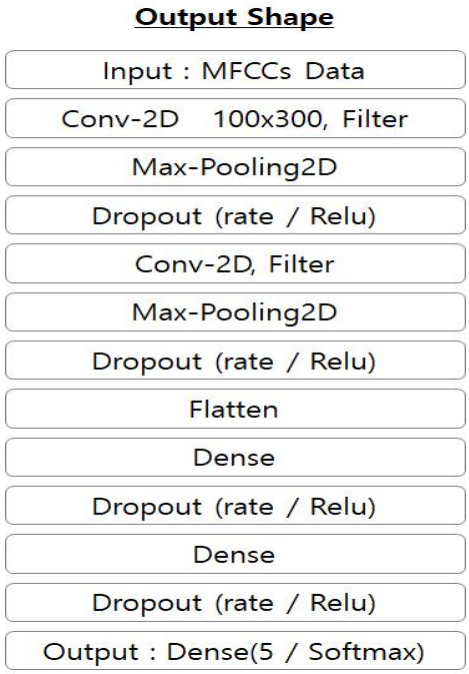
2.3. 말더듬 반복/연장 자동분류 식별 개요

말더듬 자동 분류 식별은 그림 2와 같이 LibriStutter 데이터셋의 라벨링된 음성 데이터를 MFCCs 이미지 데이터로 전처리하였고, 각 음성 데이터 라벨링은 유창, 음소 반복, 단어 반복, 구 반복, 연장의 5가지로 구분하였다. 다음으로 기계학습을 위해 학습데이터셋과 검증데이터셋을 7:3으로 구성하였다. 기계학습은 Python KERAS 라이브러리를 활용하여 학습데이터를 기반으로 CNN 딥러닝 알고리즘을 적용하여 식별기 모델을 수립하고 성능 검증을 실시하였다.



FLAC, Free Lossless Audio Codec; MFCCs, mel frequency cepstral coefficients; CNN, convolutional neural network.

그림 2. 말더듬 자동분류 식별기 구조
Figure 2. The structure of an automatic stuttering classifier



MFCCs, mel frequency cepstral coefficients.

그림 3. 반복/연장 CNN(convolutional neural network) 모델 기반 식별기
Figure 3. Repetition/Prolongation convolutional neural network (CNN)-based classifier

2.4. 반복/연장 자동분류 식별기 모델

반복/연장 자동 분류 최적 식별기 설정을 위해 Grid Search(GS) 방식을 적용하였다. GS 방식은 모든 가능한 하이퍼파라미터 조합을 구성하고 각 조합별 시도를 통해 가장 좋은 성능을 보이는 하이퍼파라미터값을 선택하는 방식으로, 본 연구에서는 CNN 알고리즘을 기반으로 하이퍼파라미터 설정을 위해 필터(filters)는 32, 64로 설정하였고, 커널사이즈(kernel_size) 는 (3,3), (5,5), 풀사이즈(pool_sizes)는 (2,2), (3,3)으로 설정하였고, 덴스층(dense_units)은 32, 64, 128로 설정하였고, 드랍아웃율(dropout_rates)은 0.25, 0.5로 설정하였다. 학습방식은 10회의 epoch를 설정하고 배치 사이즈(batch_size)를 64로 설정하고 딥러닝 학습을 실시하였다. GS 방식을 통해 최적화한 결과 필터는 64, 커널사이즈는 (5,5), 풀사이즈는 (3,3), 덴스층은 128, 드랍아웃율은 0.25로 도출되었고 이를 바탕으로 반복/연장 자동분류 식별기 모델을 그림 3과 같이 설계하였다.

반복/연장 자동 식별기는 MFCCs로 변환된 이미지의 시각적 특성을 도출하기 위해 각 단어별 음성 길이 차이로 인한 이미지 크기를 300으로 하고 2차원 배열로 변환하였다. Layer 구성은 2개의 합성곱층, 2개의 완전 연결층, 1개의 출력층으로 구성하였고, GS 방식을 통해 도출된 하이퍼파라미터값을 기반으로 반복/연장 자동 식별기를 구성하였다. 자동 식별기 딥러닝 학습의 loss 척도는 'sparse_ categorical_ crossentropy'로 설정하였고, 최적화 옵션은 'adam(adaptive moment estimation)'을 적용하였다. 식별기 훈련은 100회의 epoch를 설정하였고, 배치사이즈는 64로 하였고, 하이퍼파라미터들은 GS 방식에 따른 최적 결과로 적용하였다.

2.5. 반복/연장 자동분류 식별기 성능 평가

반복/연장 자동 식별기 분류 성능 평가를 위해서 혼동행렬(confusion matrix)을 기반으로 하는 정확도(accuracy), 정밀도(precision), 재현율(recall), F1-score로 확인하였다. 정확도는 식별기 모델이 전체 샘플에서 정확히 분류 예측한 샘플(TP+FN)의 비율(1)을 의미하고, 정밀도는 모델이 참으로 예측한 샘플(TP+FP) 중에서 실제로 참(TP)인 비율(2)을 의미한다. 재현율은 모델이 실제 참인 샘플(TP+FN) 중에서 올바르게 예측한 샘플(TP)의 비율(3)을 의미한다. 일반적으로 재현율이 높으면 정밀도는 낮아지는 경향을 보이고 있어서 좋은 식별기 모델은 재현율과 정밀도가 높은 수치를 나타내는데 이러한 수치는 F1-score로 확인할 수 있다. F1-score는 정밀도와 재현율의 조화 평균(4)을 나타내는 지표로서

종합적인 분류 성능을 의미하는 측정지표를 의미한다(Park & Lee, 2023).

$$Accuracy = \frac{True\ Positive + False\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (1)$$

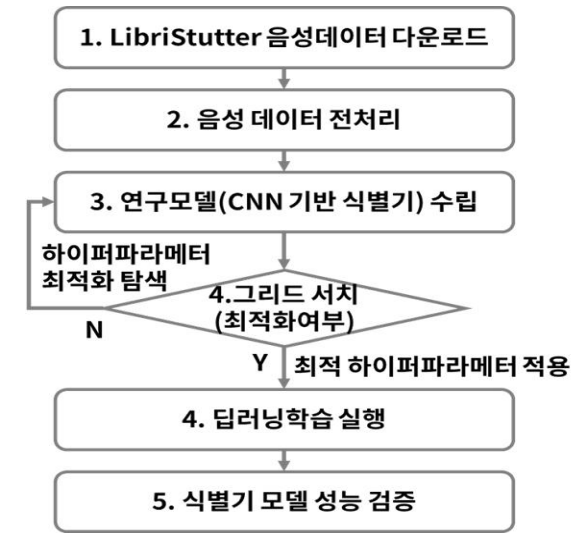
$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

2.6. 연구절차

상기한 본 자동 식별 개발 절차는 그림 4와 같이 요약될 수 있다.



CNN, convolutional neural network.

그림 4. 자동 식별 개발 절차
Figure 4. The procedure of automatic classifier development

3. 연구결과

3.1. 반복/연장 자동분류 식별기 성능 평가 결과

반복/연장 자동분류 식별기 성능 평가를 총 100회 학습을 실시한 추이 결과는 그림 5와 같다. 식별기 성능평가 최종 결과 정확도는 0.9912, 손실은 0.0544로 나타났다.

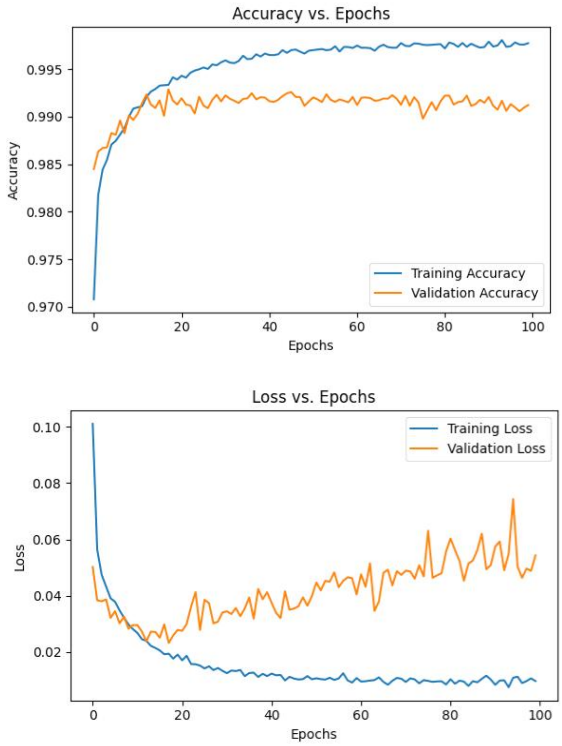


그림 5. 반복/연장 자동분류 식별기 훈련 추이 결과
Figure 5. Training results of the repetition/prolongation automatic classifier

100회 학습한 결과 정확도에 있어 epoch가 40회 이상 구간에서 학습데이터셋의 정확도와 검증데이터셋의 정확도가 0.998에서 0.990 사이로 나타나 학습데이터셋 정확도의 분산이 안정적으로 나타났고 검증세트의 정확도 분산도 동일하게 나타났다. 또한 학습세트의 정확도가 검증세트에 비해 0.008 높은 것으로 나타났다. 그리고 Loss에 있어서 학습데이터셋은 epoch가 40회 이상 구간에서 Loss 분산이 안정적으로 나타났으며 검증데이터셋의 경우 20회 이상 구간에서 점점 발산하는 형태로, 40회 Loss가 0.0413에서 95회 0.0744로 높은 Loss를 나타내고 있었다. 이는 훈련데이터셋으로 생성된 모델이 과적합되어 일반화 능력이 저하되고 있으며 최적화 알고리즘을 통한 지역 최적화를 찾지 못해 발산되고 있는 것으로 판단된다.

다음으로 반복/연장 자동분류 식별기의 분류 성능에 대한 결과분석을 실시하기 위해 유창, 음소 반복, 단어 반복, 구 반복, 연장별 정확도, 정밀도, 재현율, F1-score 결과를 표 2와 같이 확인하였다.

표 2. 반복/연장 자동식별 분류 성능 결과
Table 2. Classification performance results of the repetition/prolongation automatic detector

Types	Precision	Recall	F1-score	Accuracy
유창(clean)	1.00	1.00	1.00	0.99 (macro avg 0.86)
음소 반복 (sound repetition)	0.94	0.85	0.89	
단어 반복 (word repetition)	0.74	0.68	0.71	
구 반복 (phrase repetition)	0.77	0.78	0.78	
연장(prolongation)	0.90	0..98	0.94	

반복/연장 자동분류 식별기 분류 성능은 F1-score 0.99로 높은 수준으로 나타났고, 평균 분류 성능은 0.86으로 나타났다. 유창, 음소 반복, 단어 반복, 구 반복, 연장의 각 정밀도, 재현율, F1-score의 경우 유창은 모두 1.00으로 나타나 유창에 대한 분류 성능이 높은 것으로 확인되었다. 다음으로 음소 반복의 경우 정밀도가 0.94, 재현율이 0.85, F1-score가 0.89로 나타나 학습데이터에 대한 분류 성능이 높고 검증데이터의 분류 성능이 낮은 것으로 확인되었다. F1-score는 0.89로 단어 반복, 구 반복에 비해 반복 유형 중 가장 높은 분류 성능으로 나타났다. 다음으로 단어 반복의 경우 정밀도 0.74, 재현율 0.68, F1-score 0.71로 나타나 분류 성능이 반복 중에 가장 낮은 것으로 확인되었다. 다음으로 구 반복의 경우 정밀도 0.77, 재현율 0.78, F1-score 0.78로 정밀도와 재현율이 균형을 이루고 있음이 나타났다. 마지막으로 연장의 경우 정밀도 0.90, 재현율 0.98로 나타나 학습데이터보다 검증데이터의 분류 성능이 더 좋은 것으로 나타났다. 연장의 경우 반복보다 더 높은 분류 성능을 나타내고 있었다.

그림 6은 반복/연장 자동분류 식별기별 혼동행렬에 대한 결과를 나타내는 도표로서 유창의 경우 자동 분류 식별기가 정확히 분류한 건은 40,512건으로 나타났고, 음소 반복이 33건, 단어 반복이 30건, 구 반복이 9건, 연장이 5건으로 오분류 되었다. 즉 유창의 경우 음소 반복과 단어 반복에 대한 분류 성능이 구 반복과 연장에 비해 낮은 것으로 나타났다. 다음으로 음소 반복의 경우 403건이 정확하게 분류되었고 단어 반복이 21건 유창이 5건, 구 반복이 2건으로 오분류되었다. 단어 반복의 경우 317건이 정확하게 분류되었고 구 반복이 76건, 음소 반복이 24건, 유창이 9건, 연장이 1건으로 오분류되었다. 다음으로 구 반복의 경우 348건이 정확하게 분류되었고, 단어 반복이 90건, 유창이 7건, 음소 반복이 4건, 연장이 2건으로 오분류되었다. 마지막으로 연장의 경우 467건이 정확하게 분류되었고 유창이 23건, 구 반복 12건, 단어 반복 11건, 음소 반복이 8건으로 나타났다.

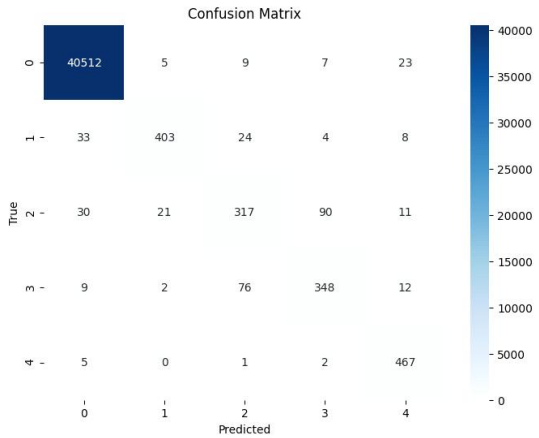


그림 6. 반복/연장 자동분류 식별기 혼동행렬(0=유창, 1=음소 반복, 2=단어 반복, 3=구 반복, 4=연장)
Figure 6. Confusion matrix for the repetition/prolongation automatic classifier (0=fluent, 1=sound repetitions 2=word repetitions, 3=phrase repetitions, 4=prolongations)

4. 논의 및 결론

본 연구는 CNN 기반의 딥러닝 알고리즘을 활용하여 말더듬의 비유창성 유형 중 반복과 연장을 자동으로 분류하는 모델을 개발하고, 그 성능을 평가하였다. 최적화된 하이퍼파라미터를 적용한 반복 및 연장 자동 분류 식별기의 최종 성능은 정확도 0.9912, 손실 0.0544로 나타나 비교적 우수한 성과를 보였다. 분류 성능 평가에서 유창, 음소 반복, 단어 반복, 구 반복, 연장의 다섯 가지 유형에 대한 평균 F1-score는 0.86으로 나타났으며, 특히 음소 반복과 연장에서는 상대적으로 높은 분류 성능을 보였다. 하지만 단어 반복과 구 반복의 분류 성능은 다소 낮아 향후 개선이 필요한 것으로 판단되었다.

이러한 결과를 바탕으로 몇 가지 논의를 하자면 다음과 같다. 첫째, LibriStutter를 이용한 이전 연구인 Kourkounakis et al.(2020)의 결과와 비교해보면, 본 연구에서는 음소 반복(94% vs. 79%)과 연장(90% vs. 89%)에서는 상대적으로 높은 정확도를, 반면에 단어 반복(74% vs. 92%)과 구 반복(77% vs. 92%)에서는 낮은 정확도를 보였다. 본 연구에서 CNN 기반 모델이 단어 반복과 구 반복 간의 분류 성능이 상대적으로 낮은 이유는 두 유형의 음향적 유사성에 기인한 것으로 판단된다. 단어 반복과 구 반복은 발음 속도나 강세, 음소 간 간격 등에서 유사한 특성을 가지며, 이러한 미세한 차이는 모델이 학습하기 어려운 음향 패턴이 될 수 있다. 단어 반복과 구 반복 모두 발화가 반복되는 현상으로 인식되지만, 반복 단위가 다르기 때문에 음향적으로 매우 비슷한 특징을 공유하면서도 작은 차이로 인해 혼동이 발생될 수 있다. 이를 해결하기 위해서는 MFCCs와 같은 음향적 특징 외에도 발화 패턴에서 차이를 명확히 구분할 수 있는 다양한 음성 특징을 추가하여 모델이 두 유형을 구별하도록 개선할 필요가 있다. 예를 들어, 발화의 시작과 끝 지점

에서 음소의 강세 변화나 주파수 대역 변화 수치, 발화 길이 등 정형화된 데이터의 추가적인 특징을 모델에 제공하면 단어와 구 반복의 차이를 좀 더 뚜렷이 학습시킬 수 있을 것이다. 또한, 데이터 증대 기법(Yang et al., 2021)을 사용하여 단어와 구 반복 발화에 대한 음향 데이터를 증가시키고, 이를 통해 데이터 다양성을 높여 모델의 학습 범위를 확장할 필요가 있다. 이러한 접근을 통해 두 유형 간의 구분이 더욱 정교하게 이루어질 수 있을 것으로 기대된다.

둘째, 본 연구에서는 MFCCs 기반의 음성 데이터와 같은 정형적 특징만을 사용했으나, 다중 양식(multi-modal) 데이터 접근 방식을 도입함으로써 모델의 성능을 더욱 향상시킬 수 있다. 다중 양식 접근은 기존의 음향적 정보 외에 비정형 데이터를 추가로 수집하여, 비유창성 발화 시 발화자의 신체적, 생리적 반응을 기반으로 분류 성능을 높이는 방법론이다(Das et al., 2022). 말더듬 발생 시 화자의 신체적 긴장이나 심리적 불안은 얼굴 표정, 시선 움직임, 심박수, 뇌파 등의 생리적 반응으로 나타나며, 이러한 반응 데이터를 모델에 추가함으로써 발화 패턴 이외의 정보도 함께 고려할 수 있다. 예를 들어, 비유창성 발화가 발생할 때 발화자의 눈 깜빡임 빈도나 입술 근육의 긴장도는 특정 유형의 말더듬 발화와 연관될 수 있다. 비정형 데이터를 활용하면 반복과 연장 유형이 주는 음향적 정보 외에 말더듬의 정서적, 생리적 특성을 모델이 학습할 수 있어 정확도가 높아질 것이다. 이러한 다중 양식의 데이터를 추가로 학습시켜 CNN 모델에 적용하면 말더듬 유형 분류 성능을 더욱 향상시킬 수 있을 것으로 기대되며, 이는 임상 환경에서의 실제 적용 가능성도 높여줄 수 있을 것이다.

셋째, 본 연구 결과에서는 학습데이터에서의 높은 정확도에 비해 검증데이터에서의 성능이 다소 낮아졌으며, 이는 과적합 문제로 인한 성능 개선의 여부를 검토할 필요가 있다. 과적합 문제는 학습데이터의 특정 패턴에 과도하게 맞춰진 모델이 검증데이터나 새로운 데이터에 대해 일반화 성능이 떨어지는 문제이다. 이를 해결하기 위해 드롭아웃(dropout; Hinton et al., 2012)이나 조기 종료(early stopping; Caruana et al., 2000)와 같은 정규화 방법을 도입할 수 있다. 드롭아웃은 학습 과정에서 무작위로 일부 뉴런을 비활성화하여 모델이 특정 패턴에 의존하지 않도록 유도하는 방식으로, 모델의 일반화 능력을 높이는 데 효과적이다. 또한, 조기 종료는 모델의 학습이 특정 횟수를 초과하여 손실이 증가할 경우 학습을 중단하여 과적합을 방지하는 방법이다. 이러한 정규화 방법 외에도, 다양한 환경의 발화 데이터를 사용하여 학습데이터셋을 확장함으로써 모델이 다양한 패턴을 학습할 수 있도록 하는 것도 중요한 방안이 될 수 있다. 추가적인 데이터셋 확보가 어려운 경우, 데이터 증강 기법을 사용하여 데이터의 양과 다양성을 높여 모델이 과적합되지 않도록 하는 것도 고려해볼 만한 접근이다. 이러한 방법들은 모델의 일반화 성능을 강화하며, 더 많은 데이터셋에서 일관된 성능을 기대할 수 있도록 할 것이다.

넷째, 본 연구에서 개발된 반복 및 연장 자동 분류 식별기는 임상 환경에서 사용 가능한 자동화된 비유창성 평가 도구로 발전할 가능성을 가지고 있다. 서론에서 언급한 것처럼, 기존의 청지각적 판단에 기반한 말더듬 평가는 평가자의 주관적 판단에 의해 일관성이 떨어질 수 있으며, 시간과 비용이 많이 소요된다는 단점이 있다. 이에 반해, 자동화된 평가 도구는 데이터 기반으로 일관된 결과를 제공할 수 있으며, 평가 시간이 단축됨으로써 임상 환경에서의 활용성을 크게 높일 수 있다. 예를 들어, 본 연구의 모델을 임상용 소프트웨어나 애플리케이션으로 구현하여 말더듬 화자의 발화를 자동으로 분석하고, 결과를 즉시 제공할 수 있도록 한다면 진단과 중재 계획 수립에 있어 효과적인 도구가 될 수 있다. 이와 같은 자동화 평가 도구는 객관적인 데이터 기반의 평가를 통해 말더듬 화자에 대한 정밀한 진단과 효과적인 중재 계획 수립을 가능하게 하여, 임상 현장에서 중요한 역할을 할 수 있다. 더 나아가, 반복 및 연장 외에도 막힘과 같은 다른 말더듬 유형을 추가하여 더 폭넓은 말더듬 평가가 가능하도록 연구를 확장할 수 있으며, 이러한 자동화된 도구는 실제 임상 환경에서 많은 관심을 받을 것으로 기대된다. 이를 위해 모델의 성능을 향상시키는 것과 더불어, 실제 임상 환경에서의 파일럿 테스트를 통해 활용 가능성을 검증하는 후속 연구가 필요할 것으로 사료된다.

마지막으로 한국어 말더듬 화자 음성 데이터셋 구축은 언어 특성 반영, 기술 개발 지원, 임상 진단 및 치료의 질 향상 등의 측면에서 중요한 의미를 지닌다. 한국어는 음운론적 및 구문적 구조에서 영어와 다른 특성을 가지고 있으며, 이는 말더듬 현상의 양상에 있어서도 차별적인 특성을 나타낼 수 있다. 한국어 말더듬의 특성을 정확히 반영하기 위해서는 한국어 화자의 음성 데이터를 바탕으로 한 데이터셋이 필요하다. 인공지능(AI)과 음성 인식 분야의 발전으로 인해 말더듬 화자의 음성을 인식하고 이를 분석하는 기술의 수요가 증가하고 있다. 그러나 현재 말더듬 연구의 상당 부분은 영어 데이터를 바탕으로 이루어지고 있어 한국어 화자를 대상으로 한 기술 개발에는 한계가 있다. 따라서 한국어 말더듬 데이터셋을 구축함으로써 보다 정확한 분석과 인식이 가능해질 것이다. 또한 한국어 말더듬 화자의 음성 데이터를 확보함으로써 임상 현장에서 보다 신뢰도 높은 진단과 치료가 가능하다. 한국어 말더듬 특성에 최적화된 데이터셋은 임상적 유용성과 효과성을 높일 수 있을 것이다.

본 연구는 인공지능 기반의 딥러닝 알고리즘(CNN)을 활용하여 말더듬 비유창성 중 반복과 연장의 자동 분류 식별기를 개발하고 그 성능을 검증하였다. 높은 분류 성능을 통해 반복과 연장의 자동화된 평가가 가능함을 확인하였으며, 이는 임상 및 연구 환경에서 말더듬 평가의 효율성과 신뢰도를 높이는 데 기여할 수 있을 것이다. 향후 연구에서는 다양한 데이터와 다중 양식 접근 방식을 적용하여 반복과

연장의 자동 분류 성능을 강화하고, 임상 적용 가능성을 넓혀갈 필요가 있을 것이다.

References

Alnashwan, R., Alhakbani, N., Al-Nafjan, A., Almudhi, A., & Al-Nuwaiser, W. (2023). Computational intelligence-based stuttering detection: A systematic review. *Diagnostics*, *13*(23), 3537.

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

Barrett, L., Hu, J., & Howell, P. (2022). Systematic review of machine learning approaches for detecting developmental stuttering. *IEEE/ACM Transactions on Audio*, *Speech*, *and Language Processing*, *30*, 1160-1172.

Bhushan, P., Vani, H. Y., Shivkumar, D. K., & Sreeraksha, M. R. (2021). Stuttered speech recognition using convolutional neural networks, *International Journal of Engineering Research & Technology*, *9*(12), 250-254.

Caruana, R., Lawrence, S., & Giles, C. L. (2000). Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In Leen, T., Dietterich, T., & Tresp, V. (Eds.), *Advances in Neural Information Processing Systems 13 (NIPS 2000)*. Denver, CO.

Das, A., Mock, J., Irani, F., Huang, Y., Najafirad, P., & Golob, E. (2022). Multimodal explainable AI predicts upcoming speech behavior in adults who stutter. *Frontiers in Neuroscience*, *16*, 912798.

Fang, S. H., Tsao, Y., Hsiao, M. J., Chen, J. Y., Lai, Y. H., Lin, F. C., & Wang, C. T. (2019). Detection of pathological voice using cepstrum vectors: A deep learning approach. *Journal of Voice*, *33*(5), 634-641.

Fook, C. Y., Muthusamy, H., Chee, L. S., Yaacob, S. B. & Adom, A. H. B. (2013). Comparison of speech parameterization techniques for the classification of speech disfluencies. *Turkish Journal of Electrical Engineering & Computer Sciences*, *21*(7), 1983-1994.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, UK: MIT Press.

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

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv*. https://doi.org/10.48550/arXiv.1207.0580

Howell, P., & Sackin, S. (1995, August). Automatic recognition of repetitions and prolongations in stuttered speech. *Proceedings of the First World Congress on Fluency Disorders* 2 (pp. 372-374), Munich, Germany.

Jo, C., Wang, S. G., & Kwon, I. (2022). Performance comparison on vocal cords disordered voice discrimination via machine learning methods. *Phonetics and Speech Sciences*, *14*(4), 35-43.

Kourkounakis, T., Hajavi, A., & Etemad, A. (2020, May). Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory. *ICASSP 2020－2020 IEEE International Conference on Acoustics*, *Speech and Signal processing (ICASSP)* (pp. 6089-6093). Barcelona, Spain.

Kourkounakis, T., Hajavi, A., & Etemad, A. (2021). FluentNet: End-to-end detection of stuttered speech disfluencies with deep learning. *IEEE/ACM Transactions on Audio*, *Speech*, *and Language Processing*, *29*, 2986-2999.

Kully, D., & Boberg, E. (1988). An investigation of interclinic agreement in the identification of fluent and stuttered syllables. *Journal of Fluency Disorders*, *13*(5), 309-318.

Lee, Y. H. (2017). Speech/audio processing based on deep learning. *Broadcasting and Media Magazine*, *22*(1), 47-58.

Mahesha, P., & Vinod, D. S. (2013). Classification of speech dysfluencies using speech parameterization techniques and multiclass SVM, *Proceedings of the International Conference on Heterogeneous Networking for Quality*, *Reliability*, *Security and Robustness* (pp. 298-308). Berlin, Heidelberg.

Pálfy, J., & Pospíchal, J. (2011, September). Recognition of repetitions using support vector machines. *Signal Processing Algorithms Architectures*, *Arrangements*, *and Applications* (pp. 1-6). Poznan, Poland.

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

Ravikumar, K. M., Rajagopal, R., & Nagaraj, H. C. (2009, June). Stuttered speech using MFCC features. In *ICGST International Journal on Digital Signal Processing 9* (pp. 19-24), Wilmington, DE.

Ravikumar, K. M., Reddy, B., Rajagopal, R., & Nagaraj, H. C. (2008). Automatic detection of syllable repetition in read speech for objective assessment of stuttered disfluencies. *International Journal of Electrical and Computer Engineering*, *2*(10), 2142-2145.

Sheikh, S. A., Sahidullah, M., Hirsch, F., & Ouni, S. (2022). Machine learning for stuttering identification: Review, challenges and future directions. *Neurocomputing*, *514*, 385-402.

Shim, H. S., Shin, M. J., Lee, E. J., Lee, K. J., & Lee, S. B.

(2022). *Fluency disorders: Assessment and treatment*. Seoul, Korea: Hakjisa.

Świetlicka, I., Kuniszyk-Jóźkowiak, W., & Smołka, E. (2009). Artificial neural networks in the disabled speech analysis. *Advances in Intelligent and Soft Computing*, 347-354.

van Riper, C. (1972). *Speech correction: Principles and methods* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Wang, X., Yang, S., Tang, M., Yin, H., Huang, H., & He, L. (2019). HypernasalityNet: Deep recurrent neural network for automatic hypernasality detection. *International Journal of Medical Informatics*, *129*, 1-12.

Yang, B., Wu, J., Zhou, Z., Komiya, M., Kishimoto, K., Xu, J., Nonaka, K., ... Takishima, Y. (2021, October). Facial action unit-based deep learning framework for spotting macro- and micro-expressions in long video sequences. *Proceedings of the 29th ACM International Conference on Multimedia* (pp. 4794-4798). Chengdu, China.

Yaruss, S. J. (1997). Utterance timing and childhood stuttering. *Journal of Fluency Disorders*, *22*(4), 263-286.

• **박진 (Jin Park)**
가톨릭관동대학교 언어재활학과 교수
강원특별자치도 강릉시 범일로 579번길 24 (내곡동)
Tel: 033-649-7737
Email: gatorade70@cku.ac.kr
관심분야: 유창성장애, 음성장애

• **이창균 (Chang Gyun Lee)** 교신저자
가톨릭관동대학교 경영학과 교수
강원특별자치도 강릉시 범일로 579번길 24 (내곡동)
Tel: 033-649-7266
Email: kdmis@cku.ac.kr
관심분야: 인공지능, 빅데이터, 사물인터넷, 데이터사이언스

CNN 기반 말더듬 자동 분류: 말더듬 반복과 연장 인식

박 진[1] · 이 창 균[2]

[1]가톨릭관동대학교 언어재활학과, [2]가톨릭관동대학교 경영학과

국문초록

본 연구는 CNN 기반의 딥러닝 알고리즘을 활용하여 말더듬 화자의 반복 및 연장 비유창성 유형을 자동으로 식별하는 방법을 개발하고, 그 성능을 검증하는 것을 목적으로 한다. 연구에 사용된 데이터는 LibriStutter 데이터셋으로, 해당 음성 데이터를 MFCC(mel frequency cepstral coefficients)로 전처리하여 CNN(convolutional neural network) 모델의 학습에 사용하였다. 그리드 방식을 활용한 최적화된 하이퍼파라미터를 적용하여 반복과 연장 식별 모델을 구축한 결과, 0.9912의 정확도와 0.0544의 손실을 나타내며 우수한 성능을 보였다. 네 가지 비유창성 유형(음소 반복, 단어 반복, 구 반복, 연장) 중 음소 반복과 연장에서는 높은 분류 성능을 확인하였으나, 단어 반복과 구 반복 간의 분류 성능이 상대적으로 낮아 향후 개선이 필요한 것으로 판단되었다. 본 연구는 자동화된 비유창성 평가가 가능함을 보여주며, 향후 다양한 데이터셋과 다중 양식(multi-modal) 접근을 통해 임상적 적용 가능성을 높이는 연구가 필요할 것이다.

핵심어: 말더듬, 인공지능, 합성곱층 신경망, 반복, 연장

참고문헌

박진, 이창균(2023). 인공지능 기반의 말더듬 자동분류 방법: 합성곱신경망(CNN) 활용. *말소리와 음성과학*, *15*(4), 71-80.

심현섭, 신문자, 이은주, 이경재, 이수복(2022). *유창성장애: 평가와 치료*. 서울: 학지사.

이영한(2017). 딥러닝 기반의 음성/오디오 기술. *방송과 미디어*, *22*(1), 46-57.

조철우, 왕수건, 권익환(2022). 기계학습에 의한 후두 장애음성 식별기의 성능 비교. *말소리와 음성과학*, *14*(4), 35-43.