pISSN 2005-8063
eISSN 2586-5854
2023. 12. 31.
Vol.15 No.4
pp. 71-80

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2023.15.4.071

AI-based stuttering automatic classification method: Using a convolutional neural network*

Jin Park[1] · Chang Gyun Lee[2,**]

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

Abstract

This study primarily aimed to develop an automated stuttering identification and classification method using artificial intelligence technology. In particular, this study aimed to develop a deep learning-based identification model utilizing the convolutional neural networks (CNNs) algorithm for Korean speakers who stutter. To this aim, speech data were collected from 9 adults who stutter and 9 normally-fluent speakers. The data were automatically segmented at the phrasal level using Google Cloud speech-to-text (STT), and labels such as 'fluent', 'blockage', prolongation', and 'repetition' were assigned to them. Mel frequency cepstral coefficients (MFCCs) and the CNN-based classifier were also used for detecting and classifying each type of the stuttered disfluency. However, in the case of prolongation, five results were found and, therefore, excluded from the classifier model. Results showed that the accuracy of the CNN classifier was 0.96, and the F1-score for classification performance was as follows: 'fluent' 1.00, 'blockage' 0.67, and 'repetition' 0.74. Although the effectiveness of the automatic classification identifier was validated using CNNs to detect the stuttered disfluencies, the performance was found to be inadequate especially for the blockage and prolongation types. Consequently, the establishment of a big speech database for collecting data based on the types of stuttered disfluencies was identified as a necessary foundation for improving classification performance.

Keywords: stuttering, automatic stuttering identification, deep learning model, convolutional neural network (CNN)

1. 서론

말더듬은 반복(repetitions)이나 연장(prolongations), 막힘(blockings)과 같은 핵심행동과 이로 인한 투쟁행동 등으로 구어의 흐름이 비정상적으로 방해를 받아서 유창성이 깨지는 장애를 말한다(Van Riper, 1972). 반복은 말소리, 음절, 낱말의 일부 또는 전체가 불수의적으로 여러 번 반복되는 것을, 연장은 특정 말소리가 비정상적으로 길게 이어지는 현상을 말한다. 막힘은 말을 하려

* This research was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (MOE) in 2023 (2022RIS-005).
** kdmis@cku.ac.kr, Corresponding author
Received 17 November 2023; Revised 12 December 2023; Accepted 12 December 2023

고 하지만 말소리가 나오지 않는 일종의 '끊김 현상'을 의미한다. 이러한 발화 비유창성은 말더듬의 기본행동이기에 일차행동(primary behaviors) 또는 말을 더듬는 모든 사람이 보이는 행동이기에 공통행동(universal behaviors)으로 불리기도 한다(Shim et al., 2022). 일반적으로 말더듬 평가는 빈도나 비율에 기반한 청지각적 판단을 필수적으로 고려하여 수행된다(Tichenor et al., 2022). 하지만 평가 수행에 있어 긴 소요시간과 평가자 간 신뢰도(reliability) 문제가 야기될 수 있다(Kully & Boberg, 1988; Yaruss, 1997). 따라서 말더듬 평가에 있어 소요시간을 단축시키면서도 동시에 평가의 신뢰도를 확보할 수 있는 방안이 필요하다고 할 것이다. 기본적으로 본 연구에서는 말더듬 화자들의 음성 데이터를 통한 인공지능 기반 분류 알고리즘을 개발하고 이를 통해 말더듬을 자동적으로 식별하는 방법을 고안하고자 하였다.

일반적으로 인공지능의 기계학습을 통한 말더듬 자동 식별 연구는 음성 데이터 수집, 음성 데이터 전처리(pre-processing), 연구모델 수립, 기계학습 실행, 연구모델의 성능 검증 단계로 수행되었다(Sheikh et al., 2022). 먼저 음성 데이터는 자발화 또는 읽기과제를 통해 정상 화자의 유창한 발화와 말더듬 화자의 반복, 연장, 막힘과 같은 말더듬 비유창성이 수집되었다. 이후 음성 데이터 전처리 단계에서 음성 데이터를 기록(transcription)하고 이를 자동적으로 분류할 수 있는 언어학적 단위(예, 단어나 어절)로의 컷팅(cutting)과 함께 개별 말더듬 비유창성에 대한 라벨링(labeling) 작업을 수행하였다. 또한 개별 말더듬 비유창성과 연관된 음향학적 특징(예, 길이, 주파수, 진폭, 포먼트 주파수, MFCCs(mel-frequency cepstral coefficients), LPCCs(linear prediction cepstral coefficients)들을 추출하였다. 연구모델 수립과 기계학습 단계에서는 추출된 특징들을 바탕으로 ANN(artificial neural network), HMM(hidden Markov model), SVM(support vector machine) 등의 기계학습 알고리즘을 통해 말더듬 자동 식별기(classifier)를 수립하였다. 최종적으로 추가 데이터셋을 통해 지도학습을 통한 학습 성능과 식별기의 성능을 검증(verification)하였다.

현재까지 인공지능 기계학습을 통한 말더듬 자동 식별에 대한 다양한 연구가 진행되었다. 예를 들어, Howell et al.(1997)은 12명의 말더듬 화자(아동)의 발화를 수집하여 반복과 연장에 관련된 최대에너지(energy peak) 갯수와 평균 길이(mean duration) 등을 포함한 총 32개의 음향학적 특징을 추출해 ANN 모델에 기반한 말더듬 자동 식별 방식을 개발하였다. 특히, 이들은 단어를 단위로 하여 유창함을 포함해 개별 비유창성(즉, 반복과 연장)에 대한 라벨링 작업을 수행하였다. 추가 데이터셋을 통한 검증 결과, 유창함과 비유창함을 포함해 92%의 식별 정확도를 보였다. Ravikumar et al.(2009)에서는 15명의 말더듬 화자(성인)를 대상으로 MFCCs를 추출한 SVM 모델을 통해 음절 반복에 대한 자동 식별 방식을 개발하였다. 이들은 음절을 단위로 하여 라벨링 작업을 수행하였으며, 검증 결과, 94.35%의 비교적 높은 정확도를 보고하였다. 비교적 최근 Mishra et al.(2021)은 MFCCs를 추출한 ANN 모델을 기반으로 하여 유창함과 비유창함(즉, 반복, 연장, 막힘 모두 포함)의 이분(binary) 분류에 대한 자동 식별 성능을 연구하였으며, 검증 결과, 86.67%의 정확도를 보였다. 이후에도 다양한 연구들이 진행되었는데, 문헌고찰 연구인 Sheikh et al.(2022)에 따르면 기계학습을 통한 말더듬 자동 식별 연구에 있어서 음성 데이터 전처리를 위해 추출된 음향학적 특성으로는 보고된 34건의 연구사례 가운데 MFCCs가 12건으로, 기계학습 알고리즘으로는 SVM과 ANN이 각각 6건으로 주로 활용되었음을 알 수 있다. 또한 기계학습을 통한 말더듬 자동 식별에 있어 음성 데이터의 충분한 확보가 중요한데, 11건의 연구사례들이 UCLASS(University of College London's Archive of Stuttered Speech)나 LibriStutter Data와 같은 공개된 데이터셋을 이용하기도 하였다. UCLASS의 경우에는 100개 이상의 실제 말더듬 발화샘플로, 반면에 LibriStutter Data는 50명의 말더듬 화자의 총 20시간 길이의 발화샘플과 추가로 합성된(synthesized) 말더듬 샘플로 구성되어 있다. 하지만 공개된 데이터셋을 통한 연구들은 관련 수행 성능을 비교할 때 장점은 있으나 해당 데이터에 한정된 실험이라는 제한성을 보일 수 있다(Jo et al., 2022).

최근에는 인공신경망의 다양한 도구가 보급되고 공통 데이터 획득이 가능해지면서 CNN(convolutional neural network)을 적용한 장애 음성 식별 연구사례들이 보고되었다(cf. Bhushan et al., 2021; Fang et al., 2019; Jo et al., 2022; Prabhu & Seliya, 2022). 이러한 연구들은 주로 장애 음성을 구별하기 위해 CNN 모델을 수립하고 분류의 정확도를 확인하는 과정으로 수행되었다. 예를 들어, Jo et al.(2022)은 후두 장애음성(양성종양과 악성종양)의 자동 식별을 위해 CNN 모델과 기계학습 방법들을 활용하여, 분류 성능에 대한 정확도, 특이도, 정밀도, 민감도를 산출하여 식별기의 성능을 확인하였다. 검증 결과, 최고 85%의 정확도를 나타냈다. 또한 Bhushan et al.(2021)은 CNN 모델을 바탕으로 유창함과 비유창함(즉, 반복, 연장, 막힘 모두 포함)의 이분적 분류에 따른 자동 식별 방법을 수립하였으며, 정확도는 89%로 보고하였다. ANN, SVM의 이후 단계에서 개발된 CNN은 딥러닝 알고리즘의 한 형태로 시각화된 음성 데이터로부터 특징을 추출하는 알고리즘으로 활용되며 합성곱층(convolutional layer), 풀링층(pooling layer)을 통한 음성 데이터의 시각적 특징을 추출하고 완전 연결층(fully-connected layer)을 통한 각 분류별 활성화 함수(activation function)를 생성해 음성 자동 분류 영역에 활용되고 있다(Goodfellow et al., 2016; Lee, 2017). 본 연구에서는 기본적으로 이러한 CNN 알고리즘을 활용한 식별기 모델을 기반으로 한국어를 모국어로 하는 말더듬 화자를 대상으로 비유창성 형태의 자동 식별 방법을 연구하였다. 또한 단순히 유창함과 말더듬 간의 이분적 구분에 초점을 맞춘 연구(Bhushan et al., 2021)와는 달리 유창한 발화와 개별 말더듬 비유창성 유형(즉, 반복, 연장, 막힘)에 대한 자동 식별 방법을 수립, 개발해 보고자 하였다.

2. 말더듬 음성 데이터

2.1. 음성 데이터 개요

연구 개발에 사용한 음성은 말더듬 성인 9명(남성 9명, 평균

연령 34세, 표준편차 5.32세)과 정상화자 9명(남성 9명, 평균연령 37세, 표준편차 7.43세)을 대상으로 일정량의 구절을 읽도록하여 음성 데이터를 수집하였다. 파라다이스 유창성검사(P-FA-II, Shim et al., 2010) 결과, 9명의 말더듬 화자 가운데 중증도가 각각 심함(severe)이 5명, 중간 정도(moderate)가 2명, 약함(mild)이 2명으로 평가되었다. 읽기자료는 총 143어절(404음절)로 고등학교 읽기 수준의 구절 자료(부록)를 활용하였다(Park et al., 2015). 녹음은 방음처리가 된 대학의 음성 실험실에서 실시되었고, 음성 신호는 16 bit, 44.1 kHz로 표본화하여 WAV 파일로 저장하였다.

다음으로 읽기구절의 WAV 파일의 음성 데이터를 언어학적 단위(즉, 어절)로 컷팅을 하였다. 이를 위해 Google Cloud STT(speech-to-text) 시스템을 활용하여 단어별로 음성인식을 실시하였다. 어절 단위의 텍스트로 인식 시 타임스탬프를 얻고 타임스탬프를 기준으로 전체 음성 데이터를 어절 단위로 자동 컷팅하였다. 어절 단위로 음성인식 컷팅 결과, 정상 화자의 경우 평균 144어절, 표준편차 1.73으로 나타났고, 말더듬 화자는 평균 133어절, 표준편차 12.31로 나타났다. 총 어절수는 143어절로 정상 화자의 경우 컷팅 어절이 높게 나타난 점은 인식 시 조금 더 많이 컷팅된 것으로 판단되고, 말더듬 화자의 경우 133어절로 음성인식으로 인한 컷팅 결과가 총 어절수보다 낮게 나타났다. 특히 표준편차의 경우 정상 화자보다 말더듬 화자가 높게 나타났는데 말더듬 수준을 고려 시 정상 화자보다 말더듬 화자에 대한 인식수준이 낮은 것을 확인할 수 있었다. 또한, 인식 정확도 측면에서 정상 화자의 경우 110개로 76.923%로 나타났고, 말더듬 화자의 경우 109개로 76.224%로 나타났다. 인식 정확도가 낮은 어절에서는 수작업으로 보정을 실시하였다. 이러한 결과는 음성인식 완전 자동화를 위한 말더듬 화자 및 정상 화자의 음성 인식수준 향상의 필요성을 시사하고 있으나 본 연구에 있어서 기존의 수작업으로 음성 데이터를 컷팅하는 것에 비교해 작업시간을 단축시키는데 유효하다는 점을 확인할 수 있었다.

다음으로 어절 단위로 분류된 음성 데이터에 대한 라벨링 작업을 실시하여 각 음성 데이터를 유창, 반복, 연장, 막힘으로 분류하였다. 라벨링 작업은 말더듬 평가 및 중재 경험이 10년 이상 된 언어재활사(1급 소지자) 3명이 참여하였다. 라벨링 작업 결과는 표 1과 같이 총 545개의 음성 샘플을 확보하였으며 말더듬 발화 데이터에 있어 반복은 37건, 막힘은 26건, 연장은 5건으로 나타났다.

표 1. 말더듬 화자 음성 데이터 데이터세트 구성
Table 1. Frequency of fluent and stuttered disfluencies and its overall rate in the audio dataset

Type of stuttered disfluencies	Frequency	Overall rate (%)
Fluent (F)	477	87.5
Repetition (R)	37	6.8
Blockage (B)	26	4.8
Prolongation (P)	5	0.9
Total	545	100.0

2.2. 음성 데이터 전처리

음성 데이터 전처리는 라벨링된 오디오 데이터를 기준으로 실시하였다. 오디오 데이터 특성은 물체에 진동하면서 발생하기 때문에 진폭(amplitude)과 시간(time)으로 구분되는 파형 형태의 데이터로 기록된다. 오디오 데이터는 연속형 데이터이고 고차원의 여러 주파수가 섞여서 생성된다. 이러한 음성 데이터에서 특징을 뽑기 위해 MFCCs, Chroma, Mel Spectrogram 등의 여러 특성 추출 방법이 활용되었다(Garg et al., 2020). MFCCs는 음성 데이터의 주파수 특성을 표현하기 위해 사용되는 특징 추출 방법이고, chroma는 음악 분석에서 사용되는 방법으로 음악의 음높이 정보를 추출하는데 사용된다. Mel은 주파수 특성이 시간에 따라 달라지는 오디오를 분석하기 위한 특징 추출 기법이다. 본 연구에서는 각 어절로 컷팅 및 라벨링된 WAV 음성 데이터를 전처리하기 위해 Python Librosa 라이브러리를 사용하여 MFCCs 데이터로 변환, 활용하였다.

데이터 변환 시 MFCCs의 경우 데이터 특징의 개수를 정해주는 파라미터인 n_mfcc는 100으로 하여 다양한 음성특징을 추출할 수 있도록 하였고, 시간영역 신호를 주파수 영역으로 변화하기 위해 샘플링레이트(sampling rate, sr)는 16 kHz로 설정하고 컷팅된 음성 데이터의 크기를 맞추기 위한 파라미터인 n_fft는 400으로 설정하였다. 일반적으로 자연어 처리시 음성을 25 ms 크기를 사용하기 때문에 sr에 frame_length를 곱한 값을 반영하였다. 다음으로 hop_length는 fft창 사이의 겹침을 나타내는 매개변수로 10 ms를 기본으로 하고 있어 sr을 반영하여 160으로 설정하였다. 이상과 같이 언어학적 단위로 유창, 반복, 연장, 막힘으로 라벨링된 WAV데이터를 그림 1과 같이 음향학적 특성에 있어 시각적 특징을 추출하기 위해 Python Librosa 라이브러리로 MFCCs 데이터를 시각화한 이미지 데이터로 변환하여 각 폴더, 어절별로 이미지 데이터로 전처리하여 저장하였다.

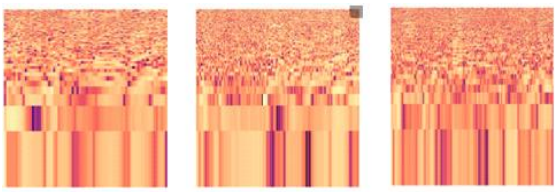


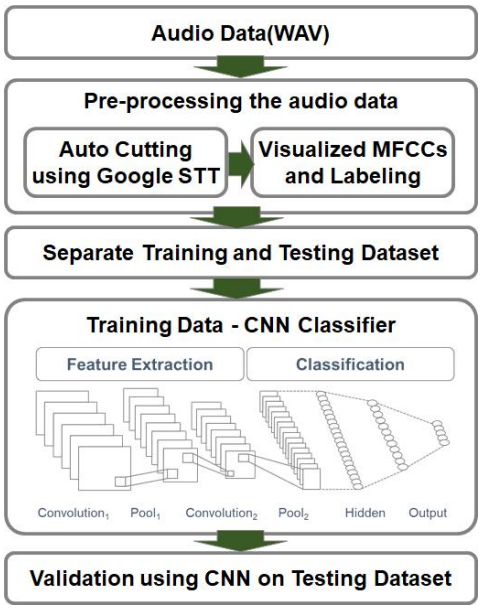
그림 1. 오디오 데이터 MFCCs 시각화(예, '김덕레라고', F, 유창; R, 반복; B, 막힘)
Figure 1. The visualized data (MFCCs) for the audio sample (e.g., 'kimdʌgɾɛrago'; F, fluent; R, repetitions; B, blockages)

3. 말더듬 자동분류 식별기

3.1. 자동분류 식별기 개요

본 연구에서 말더듬 자동분류 식별기를 개발하기 위해 그림 2와 같은 구조를 설계하였다. 음성 데이터를 Google Cloud STT 기술을 활용하여 어절 단위로 자동 컷팅하고 말더듬 비유창성 형태가 라벨링된 음성 데이터를 MFCCs 이미지 데이터로 전처

리하였다. 다음으로 학습데이터세트와 검증데이터세트을 7:3으로 구성하고, 학습데이터세트을 기반으로 CNN 알고리즘을 활용하여 식별기를 설계하였다. 식별기 설계는 Python KERAS 라이브러리를 사용하였고, 학습된 식별기 성능 검증을 실시하였다.



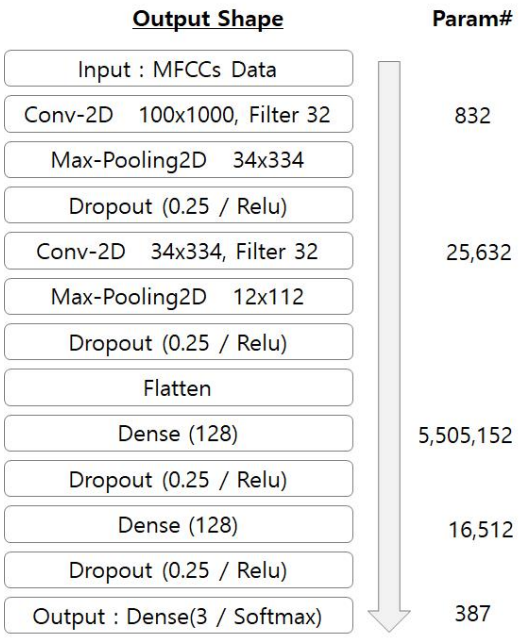
STT, Speech-To-Text; MFCC, mel frequency cepstral coefficients; CNN, convolutional neural networks.

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

3.2. 자동분류 식별기 모델

말더듬 자동분류 식별기는 CNN 알고리즘을 기반으로 설계하였다. 식별기 설계에 앞서 본 연구에서는 말더듬 유형을 유창, 막힘, 연장, 반복으로 분류화하는 것을 목적으로 하였으나 데이터 수집 결과, 연장의 경우 5개 어절로 나타나 식별기 설계에서 제외하였다. 다음으로 최적 식별기 설정을 위해 GS(Grid Search) 방식을 적용하였다. CNN 알고리즘을 기반으로 하이퍼파라미터 설정을 위한 GS의 조합은 필터(filters)는 32, 64로 설정하였고, 커널사이즈(kernel_size)는 (3,3), (5,5), 풀사이즈(pool_sizes)는 (2,2), (3,3)으로 설정하고, 덴스층(dense_units)은 32, 64, 128, 256으로 설정하고, 드랍아웃율(dropout_rates)은 0.25, 0.5로 설정하였다. 학습방식은 10회의 epoch를 설정하고 배치 사이즈(batch_size)를 64로 설정하고 기계학습을 실시하였다. 최적화 결과 정확도는 0.95679로 나타났고 조합결과는 필터는 32, 커널사이즈 (5,5), 풀사이즈 (3,3), 덴스층은 128, 드랍아웃율은 0.25로 나타났다.

GS방식을 통해 최적화한 말더듬 자동분류 식별기 모델은 총 학습가능한 파라미터 5,548,515개이며 그림 3과 같이 식별기를 설계하였다. 이 식별기는 MFCCs로 변환된 데이터로부터 받은 이미지를 시각적 특성을 도출하기 위해 각 단어별 음성 길이 차이로 인한 이미지 크기를 처리하기 위해 각 단어별 MFCCs 데이터의 배열 크기를 1,000으로 하여 2차원 배열로 변환하여 처리하였다. Layer 구성은 2개의 합성곱층(convolutional) layer, 2개의 완전히 연결된(fully-connected) layer, 1개의 출력 layer로 구성하였다. 첫 번째 layer Conv2D는 입력 이미지로부터 32개의 필터를 사용하고 832개의 파라미터를 생성하고 Max_pooling2d에서 3×3 필터를 사용하여 특성 맵의 크기를 풀링하였고 드랍아웃 활성화함수를 relu로 하여 0.25로 설계하였다. 두 번째 Conv2D는 32개의 필터를 사용하여 25,632개의 파라미터를 생성하고 Max_pooling 2d에서 3×3 필터를 사용하여 특성 맵의 크기를 풀링하였고 드랍아웃율은 동일한 기준을 적용하였다. 세 번째 layer는 다차원 데이터를 1차원으로 평탄화하고, 128개 뉴런으로 하여 5,505,152개의 파라미터 생성 및 드랍아웃은 동일하게 설계하고 네 번째 layer도 128개 뉴런으로 16,512개의 파라미터를 생성하고 드랍아웃을 동일하게 적용하였다. 마지막 출력 layer에서는 말더듬 형태별 분류를 하기 위해 유창, 막힘, 반복의 3개 뉴런으로 하여 활성화 함수를 softmax로 설계하였다.



MFCC, mel frequency cepstral coefficients; CNN, convolutional neural networks.

그림 3. MFCCs 데이터 - CNN 식별기
Figure 3. CNN classifier based on MFCCs data

4. 실험결과

4.1. 연구모델 성능평가 개요

말더듬 자동분류 식별기 성능평가를 실시하기 위해 loss척도는 'sparse_ categorical_crossentropy'로 설정하였다. 사용된 loss 척도는 다중 클래스 분류 문제에서 자주 사용되는 손실함수로서 주로 정수 형태의 레이블로 표현된 클래스를 가진 데이터에 적용하고 있다. 다음으로 최적화 옵션은 'adam(adaptive moment estimation)'을 적용하였고 식별기 훈련에는 100회의 epoch를 설

정하였고, 배치사이즈는 64로 하였고 GS방식을 통한 하이퍼파라미터들은 최적화는 식별기 모델결과를 적용하였다.

분류성능 평가를 위해서는 혼동행렬(confusion matrix)을 기반으로 하는 정확도(accuracy), 정밀도(precision), 재현율(recall), F1-score로 확인하였다. 혼동행렬이란 분류 모델의 성능을 평가하기 위한 표로서 정확도, 정밀도, 재현율, F1-score를 산출하기 위한 기초 통계표이다. 정확도는 식별기 모델이 전체 샘플에서 정확히 분류 예측한 샘플(TP+FN)의 비율 (1)을 의미한다. 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 측정지표이다.

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

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

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

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

4.2. 연구모델별 성능평가 결과

말더듬 자동분류 식별기 성능평가 결과는 그림 4와 같다. MFCCs-CNN 말더듬 자동분류 식별기의 정확도는 0.95679, 손실은 0.35755로 나타났다.

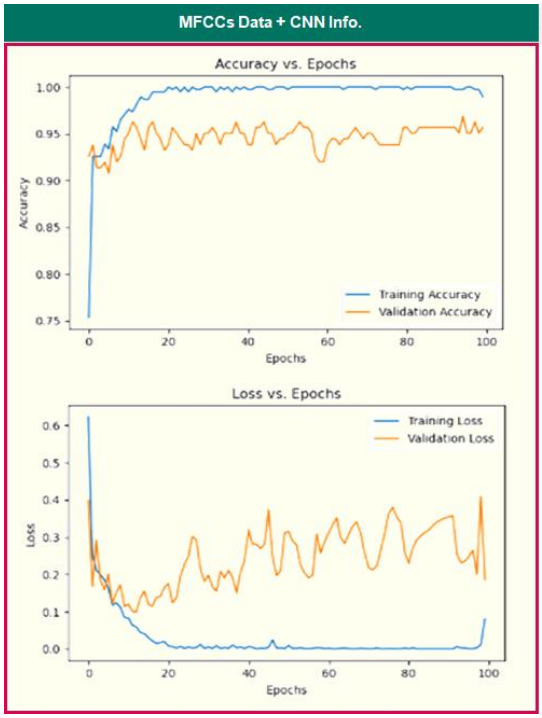


그림 4. 식별기 검증 손실 및 정확도 추이
Figure 4. Verification loss and accuracy trends by classifier

총 100회의 학습을 수행하면서 변화되는 학습곡선을 살펴보면 학습세트의 성능이 검증세트의 성능보다 높은 정확도를 나타내고 있으며 epoch이 20회 이상구간에서 학습세트의 정확도의 분산이 안정적이나 검증세트의 정확도의 분산이 학습세트에 비해 높은 것으로 나타나 과대적합에 대한 유의가 필요하다.

다음으로 분류성능에 대한 분석을 실시하기 위해 각 식별기별 정확도, 정밀도, 재현율, F1-score 결과를 표 2와 같이 확인하였다.

표 2. 식별기 분류 성능 결과
Table 2. Results of the classification performance

Types of fluent and stuttered disfluencies	Precision	Recall	F1-score	Accuracy
Fluent	0.99	1.00	1.00	0.96
Blockages	0.75	0.60	0.67	
Repetition	0.71	0.77	0.74	

말더듬 자동분류 식별기의 분류 성능은 유창의 경우 정밀도는 0.99, 재현율은 1.00, F1-score는 1.00로 나타났다. 막힘의 경우 정밀도는 0.75, 재현율은 0.60, F1-score는 0.67로 나타났고, 반복의 경우 정밀도는 0.71, 재현율은 0.77, F1-score는 0.74로 막힘에 비해 분류 성능이 높은 것으로 나타났다.

그림 5는 식별기별 혼동행렬을 나타낸 것으로 말더듬 자동분류 식별기가 유창의 경우 정확하게 분류된 건이 139건으로 나타났다. 막힘의 경우 6건이 정확하게 분류되었고 반복 4건으로 총 4건이 오분류 되었다. 반복의 경우 10건이 정확하게 분류되

었고 유창 1건, 막힘 2건으로 총 3건이 오분류 되었다. 분류 성능평가 결과 막힘과 반복에 대한 분류 성능이 낮은 점은 향후 분류 성능을 높이기 위한 충분한 데이터 수집을 통한 실험 연구가 필요한 것으로 판단된다.

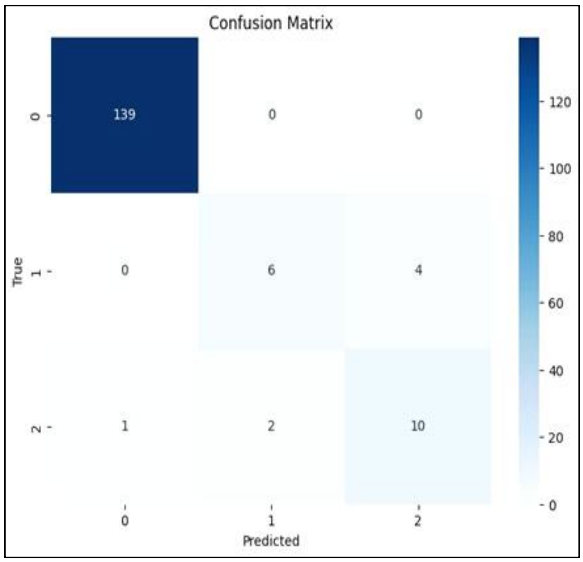


그림 5. 식별기 혼동행렬(0=유창, 1=막힘, 2=반복)
Figure 5. Classifier's confusion matrix (0=fluent, 1=blockages, 2=repetitions)

5. 논의와 결론

본 연구는 기본적으로 말더듬 화자의 유창한 발화를 포함해 비유창성 유형을 반복, 연장, 막힘의 형태로 자동으로 분류하여 말더듬 평가 수행에 있어 긴 소요시간과 평가자 간의 신뢰성 문제를 해결하고자 하는 인공지능 기반의 말더듬 자동분류 식별기 연구의 일환으로 진행되었다. 이를 위해 본 연구에서는 공개된 영어권 말더듬 화자 데이터(예, UCLASS, LibriStutter)가 아닌 한국어 화자의 발화를 활용하여 연구를 진행하였다. 또한 말더듬 비유창성 관련 특징으로 MFCCs를 추출해 CNN 알고리즘 모델을 기반으로 한 말더듬 자동분류 식별기를 설계하였으며 이러한 방식의 유효함(effectiveness)도 확인하였다. 하지만 막힘, 반복의 경우 식별기의 성능이 유창에 비해 상대적으로 낮은 것으로 나타났고, 연장의 경우 샘플 데이터 부족으로 인한 식별기 모델 적용이 어려운 것으로 나타났다. 향후 분류 성능 개선을 위한 충분한 말더듬 유형별 데이터 수집을 통해 추가 검증이 필요함을 확인할 수 있었다.

본 연구 결과를 바탕으로 몇 가지 논의를 하자면 다음과 같다. 첫째, 본 연구에서 유창함과 말더듬 비유창함(반복, 막힘) 모두를 포함한 전체 인식 정확도가 95.679%가 나타났다. 이중에서 유창함의 정확도와 재현율이 각각 0.99와 1.00으로 F1-score가 1.00으로 나타났다. 이는 본 연구에서 활용한 CNN 기반 말더듬 식별기를 통한 유창한 발화의 인식 성능이 매우 유효함을 보여주는 결과라 할 수 있다. 이전에 수행된 ANN, HMM, SVM 등의 다양한 기계학습을 통한 말더듬 자동식별 연구에서도 일관적으로 보여주는 결과이기도 하다(Barrett et al., 2022). 반면, 말더듬 비유창성에 대한 자동 식별 성능은 상대적으로 낮은 편으로 나타났다. 구체적으로, 반복에 대한 식별 정확도와 재현율은 각각 0.71와 0.77으로 F1-score는 0.74로 나타났다. 막힘에 대한 정확도와 재현율은 각각 0.75와 0.60으로 F1-score는 0.67로 나타났다.

Barrett et al.(2022)에서는 총 27개의 연구사례에 대한 식별 정확도를 보고하였는데, 70%(Wiśniewski et al., 2007)에서 98.24%(Hariharan et al., 2012)까지 발화 표본의 크기, 식별 대상으로써 유창함과 개별 말더듬 비유창성의 형태, 활용된 기계학습 모델의 차이에 따라 정확도의 차이를 보이고 있지만 비교적 90% 이상의 높은 정확도를 보고하였다. 이러한 이전 연구 결과와 비교해 볼 때 본 연구의 말더듬 비유창성에 대한 식별 정확도는 상대적으로 낮은 것으로 판단된다. 이는 총 9명의 말더듬 화자를 대상으로 읽기과제를 통한 발화 자료 수집, 특히 말더듬 비유창성 수집의 제한성에 기인한 것으로 사료된다. 향후 식별 성능 개선을 위한 말더듬 비유창성의 충분한 유형별 데이터 수집을 통한 추가 검증이 필요하다 할 것이다. 말더듬 화자의 발화 빅데이터 확보를 통해 보다 신뢰성있는 말더듬 자동 식별 기술의 개발은 물론 이를 통한 고도화된 평가 및 중재 관련 서비스 제공이 가능해지도록 관련 연구가 지속되기를 기대해 본다.

둘째, 본 연구에서 반복에 비해 막힘에 대한 식별 정확도가 낮게 나타났다. 이는 상기한 것처럼 기본적으로 적은 발화 표본에 기인한 것이라 할 수 있다. 하지만 막힘의 경우, 비교적 짧은 '끊김' 또는 단순한 시각적 긴장(tension)만으로 나타날 수 있기에 음성 데이터만을 가지고 정확한 파악이 어려울 수 있다(Guitar, 2019). 따라서 막힘의 정확한 식별을 위해서는 말더듬 화자의 음성 데이터뿐 아니라 턱이나 입술 등과 같은 조음기관의 긴장을 보여주는 시각적 데이터 등을 포함한 기계학습 수행이 필요할 것으로 판단된다(Altinkaya & Smeulders, 2020). 반면, 연장의 경우에는 특정 말소리의 '지속'(continuation)이라는 측면에서 음성 데이터를 통한 기계학습이 막힘보다는 용이하다 할 수 있다. 하지만 말더듬 성인의 경우 상대적으로 연장이 반복이나 막힘에 비해 출현빈도가 적다는 측면을 고려할 때(Jeon & Jeon, 2015), 우선은 말더듬 유형에 따른 충분한 데이터 수집이 필요하다 할 것이다. 또는 이러한 데이터 불균형 문제를 해결하기 위한 재표집(resampling)이나 가중치 재부여(reweighting)와 같은 방법도 고려될 수 있을 것이다(Yang et al., 2021). 이와 관련해 향후 연구가 진행되기를 기대해 본다.

셋째, 본 연구는 반복, 연장, 막힘의 비유창성으로 대변되는 말더듬의 핵심행동에 대한 평가와 관련해 인공지능 기반의 말더듬 자동 식별 기술 개발의 일환으로 진행되었다. 이를 통해 임상전문가(언어재활사)에 의한 전통적인 평가 방식의 긴 소요시간이나 평가자 간 신뢰도 문제를 해결할 수 있는 하나의 대안을 제안하고자 하였다. 하지만 말더듬은 발화에서 나타나는 비

유창성뿐 아니라 관련 신체 행동인 부수행동 또는 이차행동을 수반한다(Shim et al., 2022). 예를 들어, 말을 더듬는 순간 머리와 목 부분에 긴장을 보인다든지, 입술과 턱에서 떨림(tremor) 현상을 보이기도 한다. 또는 말더듬을 멈추게 하기 위해 눈을 자주 깜박거리거나, 고개를 젓히거나 손을 움직이는 행동을 보이기도 한다. 이러한 부수행동은 핵심행동과 아울러 중요한 말더듬 평가 항목으로, 특히 말더듬 진전(development)에 대한 중요한 표지가 되기도 한다(Riley, 1972). 따라서 말더듬 비유창성 관련 음향학적 특징들을 추출해 인공지능 기반의 자동 식별 기술을 개발하는 단계를 넘어 관련 시각적 특징 등을 포함하는 다중양식 학습(multi-model learning)을 통해 부수행동에 대한 식별 기술 연구도 필요할 것으로 사료된다(Das et al., 2022).

마지막으로 인공지능 기반 기계학습을 통한 말더듬 자동 식별 기술은 기본적으로 임상전문가(언어재활사)에 의해 행해지는 전통적인 말더듬 평가 수행에 있어 유용하게 활용될 수 있을 것으로 판단된다. 더불어 중재 효과를 파악할 수 있는 하나의 방식으로도 활용될 수 있을 것이다(Bayerl et al., 2022). 나아가 향후 지속적인 정보통신기술의 발전으로 인공지능, 빅데이터, 사물인터넷, 가상현실 등 다양한 기술들이 접목되면서 말더듬의 평가 및 중재 영역에 활용될 것은 부인할 수 없는 사실일 것이다. 하지만 현재까지 이러한 기술들은 보완적인(augmentative) 측면에서 활용되고 있으며 지속적인 기술 발전이 필요한 상황이다(Sheikh et al., 2022). 이러한 기술들이 임상 현장에서는 분명 유용하게 활용되겠지만 말더듬의 종합적인 평가와 치료를 위해서는 단순히 음향적 특징만으로 파악할 수 없는 부수행동이나 말더듬 화자의 심리나 태도에 대한 고려가 반드시 필요하다는 점이 결코 간과되어서는 안될 것이다.

감사의 글

본 연구에 참여해 음성 데이터를 제공해 주신 모든 대상자분들에게 감사의 말씀을 전합니다.

References

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.

Bayerl, S. P., von Gudenberg, A. W., Hönig, F., Nöth, E., & Riedhammer, K. (2022, June). KSoF: The Kassel state of fluency dataset – A therapy centered dataset of stuttering. *Proceedings of the 13th Conference on Language Resources and Evaluation* (pp. 1780-1787). Marseille, France.

Bhushan, P. S., 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.

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.

Garg, U., Agarwal, S., Gupta, S., Dutt, R., & Singh, D. (2020, September). Prediction of emotions from the audio speech signals using MFCC, MEL and Chroma. *Proceedings of the 12th International Conference on Computational Intelligence and Communication Networks (CICN).* Bhimtal, India.

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.

Hariharan, M., Chee, L. S., Ai, O. C., & Yaacob, S. (2012). Classification of speech disfluencies using LPC based parameterization techniques. *Journal of Medical Systems*, *36*(3), 1821-1830.

Howell, P., Sackin, S., & Glenn, K. (1997). Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: II. ANN recognition of repetitions and prolongations with supplied word segment markers. *Journal of Speech, Language, and Hearing Research*, *40*(5), 1085-1096.

Jeon, H. S., & Jeon, H. E. (2015). Characteristics of disfluency clusters in adults who stutter. *Journal of Speech-Language & Hearing Disorders*, *24*(1), 135-144.

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.

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.

Mishra, N., Gupta, A., & Vathana, D. (2021). Optimization of stammering in speech recognition applications. *International Journal of Speech Technology*, *24*(2), 679-685.

Park, J., Oh, S. Y., Jun, J. P., & Kang, J. S. (2015). Effects of background noises on speech-related variables of adults who stutter. *Phonetics and Speech Sciences*, *7*(1), 27-37.

Prabhu, Y., & Seliya, N. (2022, December). A CNN-based automated stuttering identification system. *Proceeding of the 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. Nassau, Bahamas.

Ravikumar, K. M., Rajagopal, R., & Nagaraj, H. C. (2009). Stuttered Speech Using MFCC Features. *ICGST International Journal on Digital Signal Processing*, *9*, 19-24.

Riley, G. D. (1972). A stuttering severity instrument for children and adults. *Journal of Speech and Hearing Disorders, 37*(3), 314-322.

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. (2010). *Paradise Fluency Assessment-II (P-FA-II)*. Seoul: Paradise Welfare Foundation.

Shim, H. S., Shin, M. J., Lee, E. J., Lee, K. J., & Lee, S. B. (2022). *Fluency disorders: Assessment and treatment*. Seoul: Korea.

Tichenor, S. E., Constantino, C., & Scott Yaruss, J. (2022). A point of view about fluency. *Journal of Speech, Language, and Hearing Research*, *65*(2), 645-652.

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

Wiśniewski, M., Kuniszyk-Jóźkowiak, W., Smołka, E., & Suszyński, W. (2007). Automatic detection of disorders in a continuous speech with the hidden Markov models approach. In M. Kurzynski, E. Puchala, M. Wozniak, & A. Zolnierek (Eds.), *Computer recognition systems 2: Advances in soft computing* (pp. 445-453). Berlin, Heidelberg: Springer.

Yang, B., Wu, J., Zhou, Z., Komiya, M., Kishimoto, K., Xu, J., Nonaka, K., ... Horiuchi, T. (2021, October). Facial action unit-based deep learning framework for spotting macro- and micro-expressions in long video sequences. *Proceedings of the 29*th *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)
가톨릭관동대학교 언어재활학과
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)
가톨릭관동대학교 경영학과
Tel: 033-649-7266
Email: kdmis@cku.ac.kr
관심분야: 인공지능, 빅데이터, 사물인터넷, 데이터사이언스

<부록>

안녕하세요 저는 육십이 넘은 할머니 김덕례라고 합니다. 한글을 배우고 있는 학생입니다. 한글을 배운 지 2년이 넘었는데도 아직도 어려운 받침이 있는 글자는 다 틀리니 이 노릇을 어쩌면 좋을지 모르겠습니다. 우리 연배들은 세상이 어려울 때 태어나서 못 배우기도 했지만 내가 어렸을 때에는 여자애들이 글자를 배우면 팔자가 세진다고 아버님이 절대 못 배우게 했습니다. 그것이 두고두고 내 평생에 한이 될 줄을 몰랐습니다. 우리 집 양반도 이제껏 아무 불편 없이 잘 살았으면서 왜 새삼스럽게 그런 걸 배우려고 하냐며 제가 한글 공부하는 것을 탐탁지 않게 생각했습니다. 그 설움은 아무도 모릅니다. 혹시나 누가 글씨라도 쓰라고 할까 봐 사람 많이 모인 곳은 가지 않았고 은행에 가서 돈 한 번 찾아보지 못했습니다. 한마디로 눈 뜬 장님이었습니다. 어떤 아줌마가 은행에 갈 때마다 일부러 손에 붕대를 감고 가서 다쳐서 글씨를 못쓰는 것처럼 다른 사람에게 글씨 부탁을 했다는 피눈물 나는 그 얘기는 바로 제 심정과 똑같습니다. 그러나 지금 전 세상 사는 게 재미있고 즐겁고 항상 행복합니다.

인공지능 기반의 말더듬 자동분류 방법: 합성곱신경망(CNN) 활용*

박 진[1] · 이 창 균[2]

[1]가톨릭관동대학교 언어재활학과, [2]가톨릭관동대학교 경영학과

국문초록

본 연구는 말더듬 화자들의 음성 데이터를 기반으로 하여, 인공지능 기술을 활용한 말더듬 자동 식별 방법을 개발하는 것을 주목적으로 진행되었다. 특히, 한국어를 모국어로 하는 말더듬 화자들을 대상으로 CNN(convolutional neural network) 알고리즘을 활용한 식별기 모델을 개발하고자 하였다. 이를 위해 말더듬 성인 9명과 정상화자 9명을 대상으로 음성 데이터를 수집하고, Google Cloud STT(Speech-To-Text)를 활용하여 어절 단위로 자동 분할한 후 유창, 막힘, 연장, 반복 등의 라벨을 부여하였다. 또한 MFCCs(mel frequency cepstral coefficients)를 추출하여 CNN 알고리즘을 기반한 말더듬 자동 식별기 모델을 수립하고자 하였다. 연장의 경우 수집결과가 5건으로 나타나 식별기 모델에서 제외하였다. 검증 결과, 정확도는 0.96으로 나타났고, 분류성능인 F1-score는 '유창'은 1.00, '막힘'은 0.67, '반복'은 0.74로 나타났다. CNN 알고리즘을 기반한 말더듬 자동분류 식별기의 효과를 확인하였으나, 막힘 및 반복 유형에서는 성능이 미흡한 것으로 나타났다. 향후 말더듬의 유형별 충분한 데이터 수집을 통해 추가적인 성능 검증이 필요함을 확인하였다. 향후 말더듬 화자의 발화 빅데이터 확보를 통해 보다 신뢰성 있는 말더듬 자동 식별 기술의 개발과 함께 이를 통한 좀 더 고도화된 평가 및 중재 관련 서비스가 창출되기를 기대해 본다.

핵심어: 말더듬, 자동 말더듬 식별, 딥러닝 모델, 합성곱신경망

참고문헌

박진, 오선영, 전제표 강진석(2015). 배경소음상황에 따른 성인 말더듬화자의 발화 관련 변수 비교. *말소리와 음성과학*, *7*(1), 27-37.

심현섭, 신문자, 이은주(2010). *파라다이스-유창성검사-II*. 서울: 파라다이스복지재단.

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.

* 본 과제(결과물)는 2023년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다 (2022RIS-005).