pISSN 2005-8063
eISSN 2586-5854
2025. 12. 31.
Vol.17 No.4
pp. 67-76

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2025.17.4.067

AI-based automatic detection of repetitions and prolongations: A performance comparison between deep neural network (DNN) and convolutional neural network (CNN)

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 compares the learning characteristics and classification performance of deep neural network (DNN) and convolutional neural network (CNN) models for automatic stuttering detection using identical acoustic features, namely, Mel-frequency Cepstral Coefficients (MFCCs). To this end, the LibriStutter synthetic speech dataset containing five speech types - fluent, sound-repetition, word-repetition, phrase-repetition, and prolongation - was used. Both models were implemented in Python using the Keras library: the DNN adopted fully connected dense layers, while the CNN utilized convolutional layers designed to capture localized temporal-spectral patterns in the MFCCs. Performance was evaluated using accuracy, precision, recall, and F1-score. The CNN classifier outperformed the DNN across all evaluation metrics, showing substantial F1-score improvements in word repetition (143%) and prolongation (81%). The CNN demonstrated lower overfitting and superior generalization due to the ability of its convolutional structure to extract temporal-spectral patterns effectively. CNNs showed structural advantages in MFCC-based stuttering detection and provided a promising foundation for developing automated fluency-assessment systems. This study contributes to understanding how neural architectures affect stuttering-recognition performance and supports the advancement of AI-assisted clinical evaluation.

Keywords: stuttering, artificial intelligence (AI), convolutional neural network (CNN), deep neural network (DNN), mel-frequency cepstral coefficients (MFCCs)

1. 서론

말더듬(stuttering)은 반복, 연장, 막힘 등과 같은 비유창성(disfluency)이 불수의적으로 나타나는 대표적인 유창성장애이다(van Riper, 1972). 이러한 비유창성은 화자의 의도와 상관없이 비정상적인 발화 흐름을 초래하여, 의사소통 효율성과 사회적 참여에 부정적 영향을 미친다(Gabel et al., 2004; Park, 2021). 전통적으로 임상에서는 언어재활사가 말더듬의 유형과 빈도를

* kdmis@cku.ac.kr, Corresponding author
Received 22 October, 2025; Revised 26 November, 2025; Accepted 9 December, 2025

지각적으로 판단하여 말더듬 백분율(percentage of syllables stuttered, %SS)이나 유형별 빈도를 산출함으로써 평가가 이루어진다. 그러나 이러한 지각적 평가는 숙련도에 따라 편차가 크며, 평가자 간 신뢰도(inter-rater reliability)와 재현성(reproducibility)이 낮다는 한계가 있다(Kully & Boberg, 1988; Yaruss, 1997). 또한 장시간의 녹음자료를 수동으로 분석해야 하는 비효율성은 임상 적용의 현실적 제약으로 작용한다.

이러한 한계에 따라 최근에는 인공지능(artificial intelligence, AI)을 활용한 자동 말더듬 검출(automatic stuttering detection, ASD) 기술이 주목받고 있다. 특히, 심층 신경망(deep neural network, DNN)을 중심으로 한 딥러닝(deep learning) 접근은 방대한 음성 데이터로부터 비선형적 음향 특징을 자동으로 학습함으로써, 기존의 통계 기반 모델(e.g., support vector machine, SVM; hidden markov model, HMM)보다 높은 인식 성능을 보이는 것으로 보고되고 있다(Alnashwan et al., 2023). 최근에는 DNN 외에도 합성곱 신경망(convolutional neural network, CNN), 순환 신경망(recurrent neural network, RNN) 등 다양한 구조가 음성 인식, 감정 분류, 화자 인식 등 여러 음성 처리 분야에 적용되고 있으며, 말더듬 인식 연구에서도 이러한 구조적 접근이 점차 확대되는 추세이다.

특히 DNN과 CNN은 구조적 차이에 따라 음향 데이터를 처리하는 방식에서 뚜렷한 차이를 보인다. DNN은 완전 연결층(fully connected layer)으로 구성된 다층 퍼셉트론(multilayer perceptron, MLP)의 확장형으로, 입력 데이터의 모든 요소를 은닉층 노드와 연결하여 전역적(global) 통계 패턴을 학습한다(Hinton et al., 2012). 이로 인해 제한된 데이터에서도 안정적 학습이 가능하지만, 시간적 연속성(temporal continuity)이나 주파수 간 상호작용(spectral correlation)을 명시적으로 반영하기 어렵다. 반면 CNN은 합성곱(convolution) 연산을 통해 입력 특징의 인접 영역(local receptive field)을 중심으로 지역적(local) 패턴을 추출하고, 풀링(pooling)을 통해 잡음을 제거하며 위치 불변성(invariance)을 확보한다(Krizhevsky et al., 2012; LeCun et al., 1998). MFCCs(mel-frequency cepstral coefficients) 특징을 시간 × 계수의 2차원 행렬 형태로 변환하면, CNN은 이를 이미지처럼 처리하여 시간-주파수 평면상에서 반복되는 음소나 연장된 모음의 지속적 에너지 패턴을 자동으로 학습할 수 있다. 이러한 구조적 특성은 반복(repetition)과 연장(prolongation)과 같이 시간적 패턴이 뚜렷한 비유창성 유형을 인식하는 데 특히 유리할 것으로 예측된다.

이처럼 DNN은 전역적 통계 학습(global statistical learning)에, CNN은 지역적 패턴 인식(local pattern learning)에 각각 강점을 지닌다. 따라서 두 모델 간 구조적 차이가 말더듬의 시간적 패턴 인식에 어떤 영향을 미치는지를 규명하는 것은 학문적으로나 임상적으로 중요한 의미를 가진다. 또한 모델 구조의 차이는 성능 지표(정확도, 정밀도, 재현율, F1-score)는 물론, 오분류(confusion) 양상에도 영향을 미칠 수 있다. 반복과 연장은 음성 신호에서 길이와 에너지 변화의 국소적 차이가 크기 때문에, CNN이 이러한 변동을 더 세밀하게 포착할 가능성이 있는 반면 DNN은 전역적 통계 패턴에 기반하여 장기적 평균 특성에 민감할 수 있다. 그럼에도 불구하고, 동일한 음향 특징(MFCCs)을 입력으로 하여 DNN과 CNN의 학습 특성과 성능 차이를 직접 비교한 연구는 매우 제한적이다. 말더듬 자동 평가 기술의 임상적 활용성을 높이기 위해서는 모델 구조별 학습 특성과 인식 성능의 차이를 체계적으로 규명할 필요가 있다.

따라서 본 연구의 목적은 반복과 연장 유형의 말더듬 발화를 자동으로 식별하는 과정에서, DNN과 CNN이 동일한 MFCCs 특징을 어떻게 다르게 학습하고 분류하는지를 비교・분석하는 것이다. 구체적인 연구 질문은 다음과 같다.

첫째, 반복과 연장 유형을 분류할 때 DNN과 CNN 중 어느 모델이 더 높은 정확도, 정밀도, 재현율, F1-score를 보이는가?

둘째, 두 모델의 구조적 차이(전역적 학습 vs. 국소적 학습)는 말더듬 유형별 성능 차이에 어떤 영향을 미치는가?

마지막으로 반복과 연장 유형에서 DNN과 CNN의 오분류(confusion) 양상은 어떻게 다른가?

2. 연구 방법

2.1. 음성데이터 개요

본 연구에서 활용한 음성 데이터는 Queen's University에서 공개한 LibriStutter 데이터셋(Kourkounakis et al., 2021)이며 말뭉치를 기반으로 구축된 대규모 말더듬 음성 데이터셋으로, 총 50명의 말더듬 화자(남성 23명, 여성 27명)로부터 수집된 약 20시간 길이의 합성 음성 샘플로 구성되어 있다. LibriStutter 데이터셋은 유창함(clean)을 포함해 다섯 가지 유형으로 라벨링되어 있다. 유창함은 말더듬이 없는 정상적인 발화를 의미하며, 135,092개의 샘플로 전체 데이터의 대부분을 차지한다. 음소 반복(sound repetition)은 “th-th-this”와 같이 음소가 반복되는 현상으로 1,606개의 샘플이 있다. 단어 반복(word repetition)은 “why why”와 같이 단어 전체가 반복되는 현상으로 1,597개의 샘플이 있다. 구 반복(phrase repetition)은 “know I know that”와 같이 어구가 반복되는 현상으로 1,537개의 샘플이 있다. 연장(prolongation)은 “whoooooo is there”와 같이 특정 음소가 비정상적으로 길게 이어지는 현상으로 1,564개의 샘플이 있다. LibriStutter 데이터셋은 분류 레이블이 포함된 메타데이터 파일, 각 분류별 음성 파일, 시간에 따라 정렬된 전사본을 포함하고 있으며 음성 데이터는 FLAC(free lossless audio codec) 형식으로 저장되어 있으며, 샘플링 레이트는 22 kHz로 설정되어 있다.

2.2. 음성데이터 전처리

음성 데이터는 본질적으로 진폭(amplitude)과 시간(time)으로 구성된 파형 형태의 연속형 데이터로 기록된다. 이러한 원시 음성 신호는 고차원의 복잡한 주파수 성분을 포함하고 있어, 기계학습 모델에 직접 입력하기에는 적합하지 않기 때문에 음성 신호로부터 의미있는 특징을 추출하여 모델이 학습하기 용이한 형태로 변환하는 전처리 과정이 필수적이다. 본 연구에서는 MFCCs를 음향 특징 추출 방법으로 사용하였고, MFCCs는 음성

인식 및 화자 인식 분야에서 가장 널리 사용되는 특징으로, 인간의 청각 특성을 반영한 멜 스케일(Mel scale)을 기반으로 음성 신호의 주파수 특성을 효과적으로 표현한다(Davis & Mermelstein, 1980; Rabiner & Schafer, 2010).

MFCCs 추출을 위해 Python의 Librosa 라이브러리를 사용하였으며, 다음과 같이 파라미터를 설정하였다. n_mfcc는 추출할 MFCC 계수의 개수를 결정하는 파라미터로, 본 연구에서는 100으로 설정하여 다양한 음향 특징을 충분히 포착할 수 있도록 하였고, 샘플링 레이트(sr)는 원본 데이터와 동일하게 22 kHz로 설정하였다. n_fft[fast Fourier transform (FFT) window size]는 시간 영역 신호를 주파수 영역으로 변환하기 위한 푸리에 변환 윈도우 크기로, 자연어 처리 분야에서 일반적으로 사용되는 25 ms 프레임 길이를 반영하여 550(22,000×0.025)으로 설정하였다. hop_length는 연속된 FFT 윈도우 간의 이동 거리를 나타내는 파라미터로, 표준적인 10 ms 간격을 반영하여 220(22,000×0.01)으로 설정하였다. 각 음성 샘플의 길이가 다르므로, 일정한 길이로 패딩(padding) 또는 트리밍(trimming)을 수행하여 모든 샘플이 동일한 차원의 입력 벡터를 가지도록 하였다. 이러한 파라미터 설정을 통해 각 음성 샘플로부터 100차원의 MFCC 특징 벡터를 추출하였다. CNN 모델의 경우, 추출된 MFCCs를 2차원 이미지 형태로 시각화하였고, MFCCs의 각 계수는 시간에 따라 변화하므로, 이를 시간-주파수 평면상의 이미지로 표현하였다. 가로축은 시간, 세로축은 MFCC 계수 번호를 나타내며, 각 픽셀의 색상 강도는 해당 시점의 계수 값을 나타낸다. 이렇게 생성된 MFCCs 이미지는 음성 신호의 시간적 변화 패턴을 시각적으로 표현하며, CNN이 합성곱 연산을 통해 지역적 특징을 효과적으로 학습할 수 있도록 하였다.

반면 DNN 모델의 경우, 동일한 MFCCs 특징을 추출한 후 이를 1차원 벡터로 평탄화(flatten)하여 입력 데이터로 사용하였다. 유창, 음소 반복, 단어 반복, 구 반복, 연장의 다섯 가지 유형에 대해 전처리를 수행한 결과, 각 유형별로 고유한 시각적 패턴을 관찰할 수 있었다. 유창 발화는 연속적이고 규칙적인 패턴을 보이는 반면, 음소 반복은 짧은 구간이 반복되는 패턴, 단어 및 구 반복은 보다 긴 구간이 반복되는 패턴, 연장은 특정 주파수 대역이 길게 지속되는 패턴을 보였다. 이러한 시각적 차이는 CNN이 각 유형을 구별하는 중요한 특징으로 판단된다. 그림 1은 각 유형의 MFCCs 시각화 예시이다.

2.3. DNN(deep neural network) 및 CNN(convolutional neural network) 식별기 모델

본 연구에서는 다섯 가지 유형(유창, 음소 반복, 단어 반복, 구 반복, 연장)을 자동으로 분류하기 위해 DNN과 CNN 기반의 식별기 모델을 각각 구축하였다. 두 모델은 Python의 Keras 라이브러리를 사용하여 구현되었으며, 각 알고리즘의 특성에 맞는 구조로 설계되었다. DNN 모델은 완전 연결층(Dense layer)을 기반

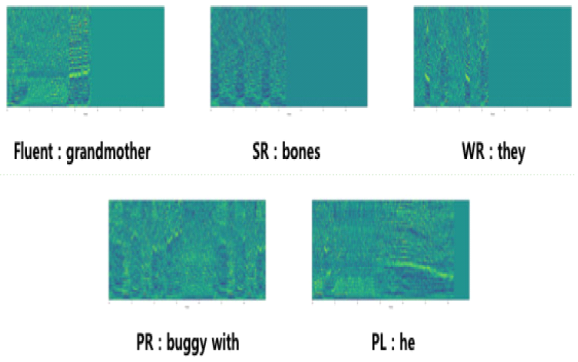

그림 1. 오디오 데이터 MFCCs 시각화(SR, 음소 반복; WR, 단어 반복; PR, 구절 반복; PL, 연장)
Figure 1. The visualized data (MFCCs) for the audio sample. (SR, sound repetition; WR, word repetition; PR, phrase repetition; PL, prolongation)

으로 하는 전통적인 심층 신경망 구조로서 입력층은 평탄화된 MFCCs 특징 벡터를 받아들인다. 그림 2는 MFCCs와 같은 음향 특징을 추출한 후, 이를 1차원 벡터로 평탄화하여 DNN의 입력으로 사용하는 일반적인 처리 과정을 도식화한 것이다. 이와 같은 전처리 과정을 통해 각 음성 샘플의 길이를 일정하게 표준화함으로써, 모델이 일관된 입력 차원에서 효율적으로 학습할 수 있도록 하였다. DNN 모델의 구체적 구조는 그림 3과 같다. 모델은 2–3개의 이상의 은닉층(Dense layer)으로 구성되며, 각 은닉층 다음에는 dropout layer를 배치하여 과적합을 방지하였다. 은닉층의 활성화 함수로는 ReLU(rectified linear unit)를 사용하였으며, 이는 음의 값을 0으로 만들고 양의 값은 그대로 전달하여 기울기 소실(vanishing gradient) 문제를 완화하는 역할을 한다. 출력층은 5개의 뉴런을 가진 은닉층으로 구성되며, Softmax 활성화 함수를 적용하여 각 클래스에 대한 확률 분포를 출력하였다.

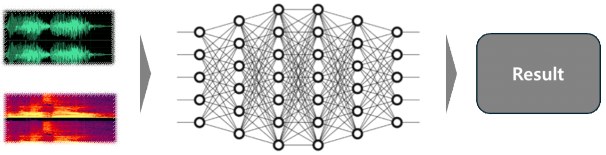

그림 2. DNN(deep neural network) 알고리즘
Figure 2. Deep neural network (DNN) algorithm

Layer (type)	Output Shape	Param #
flatten_108 (Flatten)	(None, 30000)	0
dense_378 (Dense)	(None, 128)	3840128
dropout_270 (Dropout)	(None, 128)	0
dense_379 (Dense)	(None, 64)	8256
dropout_271 (Dropout)	(None, 64)	0
dense_380 (Dense)	(None, 32)	2080
dropout_272 (Dropout)	(None, 32)	0
dense_381 (Dense)	(None, 5)	165

Total params: 3,850,629
Trainable params: 3,850,629
Non-trainable params: 0

그림 3. DNN(deep neural network) 식별기 모델
Figure 3. Deep neural network (DNN) classifier model

한편 CNN 모델은 합성곱층과 풀링층을 기반으로 하며 MFCCs 이미지의 시각적 특징을 효과적으로 추출하도록 하였고, 입력층은 2차원 MFCCs 이미지 데이터로 하고 각 이미지의 크기를 일정하게 조정하여 표준화 하였다. 그림 4는 MFCCs와 같은 음향특징을 시각화하여 CNN의 입력으로 사용하는 전처리 및 학습 개념을 도식화한 것이다. 이와 같이 MFCCs를 이미지 형태로 변환함으로써, CNN은 합성곱 연산을 통해 시간-주파수 영역에서의 지역적 특징(local features)을 학습하고, 풀링 연산을 통해 특징의 차원을 축소하면서 중요한 정보를 보존할 수 있다. 특히 반복과 연장은 CNN 구조에서 단계적으로 구분되어 학습된다. 첫 번째 합성곱층은 짧은 시간 단위의 국소 패턴을 포착하여 음소 반복과 같은 주기적 변동(프레임 간 변화)을 추출하고, 두 번째 합성곱층은 더 긴 시간축을 고려하여 단어 · 구 반복의 연속성과 패턴을 검출한다. 반면 연장은 특정 주파수 대역에서의 지속적인 에너지 분포와 프레임 간 상관성을 통해 특징 맵에 잔류하게 되며, 이는 이후 Flatten–Dense 층에서 반복 특성과 결합되어 유형별 분류에 활용된다.

CNN 구체적 구조는 그림 5와 같다. 첫 번째 합성곱층(Conv2D)은 여러 개의 필터를 사용하여 입력 이미지로부터 지역적 특징을 추출하였고, 각 필터는 작은 윈도우를 이미지 위에서 슬라이딩하며 합성곱 연산을 수행하고, 활성화 함수를 통과시켜 특징 맵(feature map)을 생성하게 하였다. 첫 번째 합성곱층 다음에는 MaxPooling2D layer가 위치하며, 이는 특징 맵의 크기를 축소하면서 가장 중요한 특징을 보존하도록 하였다. 다음으로 MaxPooling은 지정된 윈도우 내에서 최댓값을 선택하는 방식으로, 공간적 불변성(spatial invariance)을 제공하고 연산량을 줄이도록 하였으며 첫 번째 합성곱-풀링 블록 다음에는 dropout layer를 배치하여 과적합을 방지하도록 하였다. 두 번째 합성곱층은 첫 번째 층에서 추출된 특징 맵을 입력으로 받아 더 고수준의 특징을 추출하도록 하며 두 번째 합성곱층 역시 MaxPooling2D와 Dropout을 거쳐 특징의 차원을 축소하고 일반화 능력을 향상시키도록 하였다. 합성곱-풀링 블록을 거친 후에는 Flatten layer를 통해 다차원 특징 맵을 1차원 벡터로 평탄화하여 완전 연결층에 입력하였다. 평탄화된 특징 벡터는 2개의 Dense layer를 거치며, 각 층은 고차원 특징을 압축하고 최종 분류하도록 하였다. 출력층은 DNN과 동일하게 5개의 뉴런과 Softmax 활성화 함수로 사용하였다.

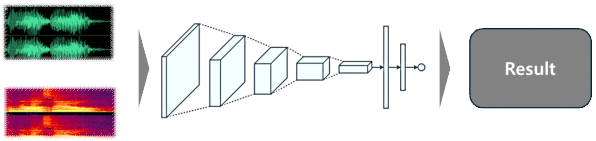

그림 4. CNN(convolutional neural network) 알고리즘
Figure 4. Convolutional neural network (CNN) algorithm

Layer (type)	Output Shape	Param #
conv2d_96 (Conv2D)	(None, 98, 298, 32)	320
max_pooling2d_96 (MaxPooling	(None, 32, 99, 32)	0
dropout_192 (Dropout)	(None, 32, 99, 32)	0
conv2d_97 (Conv2D)	(None, 30, 97, 32)	9248
max_pooling2d_97 (MaxPooling	(None, 10, 32, 32)	0
dropout_193 (Dropout)	(None, 10, 32, 32)	0
flatten_48 (Flatten)	(None, 10240)	0
dense_144 (Dense)	(None, 128)	1310848
dropout_194 (Dropout)	(None, 128)	0
dense_145 (Dense)	(None, 128)	16512
dropout_195 (Dropout)	(None, 128)	0
dense_146 (Dense)	(None, 5)	645

Total params: 1,337,573
Trainable params: 1,337,573
Non-trainable params: 0

그림 5. CNN(convolutional neural network) 식별기 모델
Figure 5. Convolutional neural network (CNN) classifier model

2.4. 식별기 모델별 하이퍼파라미터 설계

딥러닝 모델의 성능은 하이퍼파라미터 설정에 크게 의존하므로, 최적의 하이퍼파라미터를 찾는 과정이 필수적이다. 본 연구에서는 grid search 방식을 채택하여 DNN과 CNN 각각에 대해 체계적인 하이퍼파라미터 최적화를 수행하였다. Grid search는 지정된 하이퍼파라미터 값들의 모든 가능한 조합을 시도하고, 각 조합에 대해 모델을 학습시킨 후 검증 데이터셋에서의 성능을 평가하여 가장 우수한 조합을 선택하는 방법이다.

DNN 모델의 경우 하이퍼파라미터 그리드의 설정을 (1)과 같이 하였다. 은닉층의 개수(layers_n)는 2개 또는 3개로 설정하여, 모델의 깊이가 성능에 미치는 영향을 평가하도록 하였다. 각 은닉층의 뉴런 수는 32, 64, 128 중에서 선택하도록 하였고,

Dropout rate는 0.25 또는 0.5로 설정하여, 과적합 방지 강도를 조정하도록 하였다. 이러한 그리드 설정에 따라 총 2×3×3×3×2=108개의 조합이 생성되었으며, 각 조합에 대해 10회 epoch 동안 학습을 수행하도록 하였고 검증 데이터셋에서의 정확도를 기준으로 최적 조합을 선택하도록 하였다.

CNN 모델은 (2)와 같이 합성곱층과 관련된 하이퍼파라미터를 포함하였다. 필터의 개수(filters)는 32 또는 64로 설정하였으며, 이는 추출되는 특징 맵의 개수를 결정한다. 커널 크기(kernel_size)는 (3, 3) 또는 (5, 5)로 설정하여, 지역적 특징을 추출하는 윈도우의 크기를 조정하였고, 작은 커널은 세밀한 특징을, 큰 커널은 보다 넓은 영역의 특징을 포착할 수 있다. 풀링 윈도우 크기(pool_size)는 (2, 2) 또는 (3, 3)으로 설정하여, 특징 맵의 축소 정도를 조정하였다. 완전 연결층의 뉴런 수(dense_units)는 32, 64, 128 중에서 선택하였으며, dropout rate는 DNN과 동일하게 0.25 또는 0.5로 설정하였다. 이러한 그리드 설정에 따라 총 2×2×2×3×2=48개의 조합이 생성되었으며, 각 조합에 대해 10회 epoch 동안 학습을 수행하도록 하였고 검증 데이터셋에서의 정확도를 기준으로 최적 조합을 선택하도록 하였다.

(1) DNN 하이퍼파라미터 그리드 설정
layers_n = [2, 3]
dense1_units = [32, 64, 128]
dense2_units = [32, 64, 128]
dense3_units = [32, 64, 128]
dropout_rates = [0.25, 0.5]

(2) CNN 하이퍼파라미터 그리드 설정
filters = [32, 64]
kernel_sizes = [(3, 3), (5, 5)]
pool_sizes = [(2, 2), (3, 3)]
dense_units = [32, 64, 128]
dropout_rates = [0.25, 0.5]

2.5. 성능지표 정의

DNN과 CNN의 말더듬 분류 성능을 비교하기 위해 정확도(accuracy), 정밀도(precision), 재현율(recall), F1-score를 기준으로 두 모델의 성능을 정량적으로 비교하고 혼동행렬(confusion matrix)을 통해 분류 유형별 패턴을 질적으로 분석하였다. 정확도는 전체 샘플에서 정확히 분류한 비율(수식 (1))을 의미하고, 정밀도는 모델이 참으로 예측한 샘플 중 실제로 참인 비율(수식 (2))을 의미한다. 재현율은 실제 참인 샘플 중 모델이 올바르게 참으로 예측한 비율(수식 (3))을 의미하고 F1-score는 정밀도와 재현율의 조화평균(수식 (4))으로서 두 지표를 균형있게 고려한 종합적 분류 성능을 의미한다.

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

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

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

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

3. 연구 결과

3.1. DNN(deep neural network)식별기 성능평가 결과

DNN 기반 말더듬 식별기의 성능을 평가하기 위해 100회의 epoch 동안 학습을 수행하였으며, 학습 과정에서의 정확도와 손실 변화를 추적하였다. 최적 모델은 (3)과 같이 3개의 은닉층(128-64-32 뉴런)과 dropout(0.25)을 포함한 구조로 설정되었다.

(3) Best Hyperparameters:
{'layern': 3, 'units1': 128, 'units2': 64, 'units3': 32, 'dropout_rate': 0.25}

그림 6의 학습 곡선에 따르면 DNN 모델은 초기 20 epoch에서 급격한 성능 향상을 보였으며 이후 안정적으로 수렴하였다. 학습 정확도는 0.98, 검증 정확도는 0.97로 나타나 전반적으로 높은 수준이지만, 중반 이후 검증 손실이 소폭 증가하여 약한 과적합 경향을 보였다.

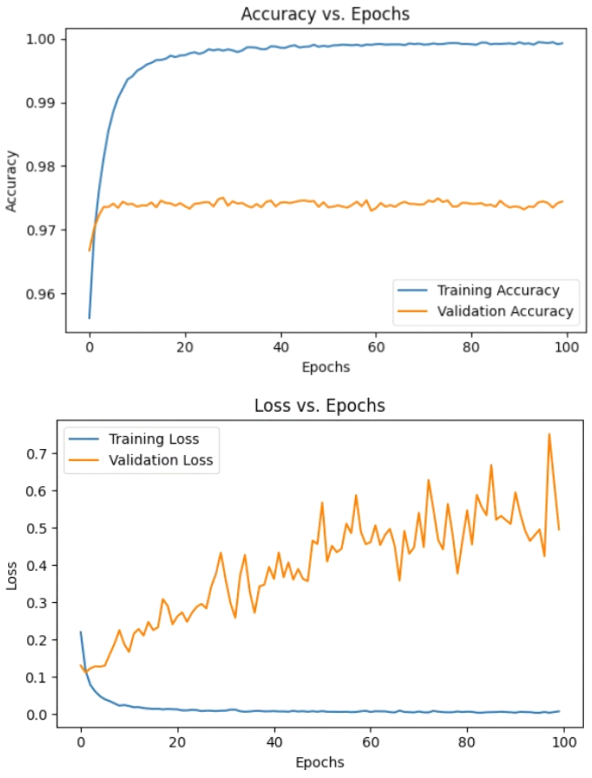

그림 6. DNN(deep neural network) 식별기 검증 손실 및 정확도 추이

Figure 6. Verification loss and accuracy trends by deep neural network (DNN) classifier

검증 데이터셋에 대한 최종 성능은 전체 정확도 0.97, macro 평균 F1-score는 0.66으로 나타났다. 유형별 F1-score는 유창 0.99, 음소 반복 0.84, 단어 반복 0.30, 구 반복 0.65, 연장 0.53이었다. 특히 단어 반복과 연장에서 낮은 재현율을 보여 시간적 패턴 인식의 한계를 드러냈다. 유창 발화는 데이터 비중이 크고 음향적으로 명확하여 가장 높은 성능을 보였으며, 음소 반복은 비교적 안정적으로 분류되었다. 반면 단어 반복과 구 반복은 서로 혼동되는 경향을 보였다. 전체 macro 평균은 정밀도 0.82, 재현율 0.59, F1-score 0.66으로, 클래스 불균형의 영향을 받았다. weighted 평균은 정밀도・재현율・F1-score 모두 0.97로 높게 나타났으나, 이는 유창 샘플의 과다 비중이 반영된 결과이다. 그림 7의 혼동행렬 분석 결과, 단어 반복은 유창 또는 음소 반복으로, 연장은 유창으로 오분류되는 경향이 두드러졌다. 이는 DNN이 평탄화된 1차원 특징 벡터를 사용함으로써 시간적 연속성과 반복 패턴의 변화를 충분히 학습하지 못했음을 시사한다.

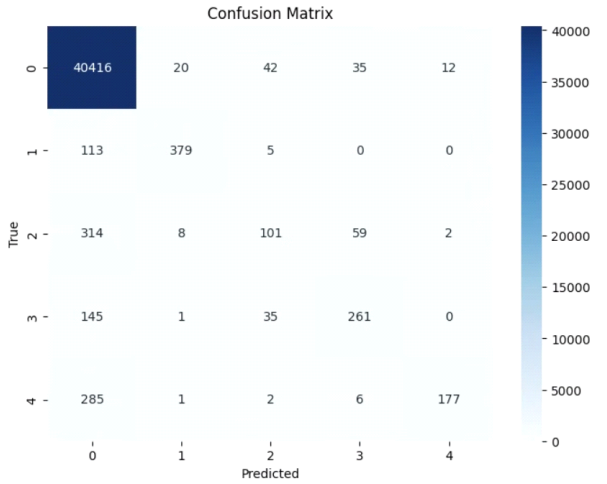

그림 7. DNN(deep neural network) 식별기 혼동행렬(0, 유창; 1, 음소 반복; 2, 단어 반복; 3, 구 반복; 4, 연장)
Figure 7. Deep neural network (DNN) Classifier's confusion matrix (0, fluent; 1, sound repetition; 2, word repetition; 3, phase repetition; 4, prolongation)

3.2. CNN(convolutional neural network)식별기 성능평가 결과

CNN 기반 말더듬 식별기의 성능을 평가하기 위해 DNN과 동일한 조건(100 epoch)에서 학습을 수행하고, 학습 및 검증 과정의 정확도와 손실 변화를 추적하였다. 그리드 서치 결과, 최적 하이퍼파라미터 조합은 (4)와 같으며, 32개의 필터, (3×3) 커널, (3×3) 풀링 윈도우, 128개의 뉴런과 0.25의 dropout rate를 포함하였다. 작은 커널과 제한된 필터 수가 선택된 것은 MFCCs 이미지의 미세한 지역적 패턴을 효과적으로 포착하기 위함으로 해석된다. 최적 조합에서 (3×3) 커널이 선택된 것은 음소 반복이나 연장과 같이 짧은 시간 구간에서 발생하는 국소적 음향 변동을 세밀하게 포착하기에 적합하기 때문으로 해석된다. 또한 32개의 필터 수는 반복의 주기성, 연장의 지속 에너지, 유창 발화의 연속성 등 비유창성 유형별 핵심 특징을 다양하게 학습하면서도 과적합을 방지하는 균형점으로 작용한 것으로 판단된다.

(4) Best Hyperparameters:
{'filters': 32, 'kernel_size': (3, 3), 'pool_size': (3, 3), 'dense_units': 128, 'dropout_rate': 0.25}

그림 8의 학습 곡선에서 CNN은 초기 15 epoch 이내에 빠르게 수렴하였으며, 학습・검증 정확도의 차이가 DNN보다 작아 과적합이 적게 발생하였다. 손실 또한 두 데이터셋에서 유사한 감소 추세를 보여 모델의 일반화 능력이 우수함을 나타냈다.

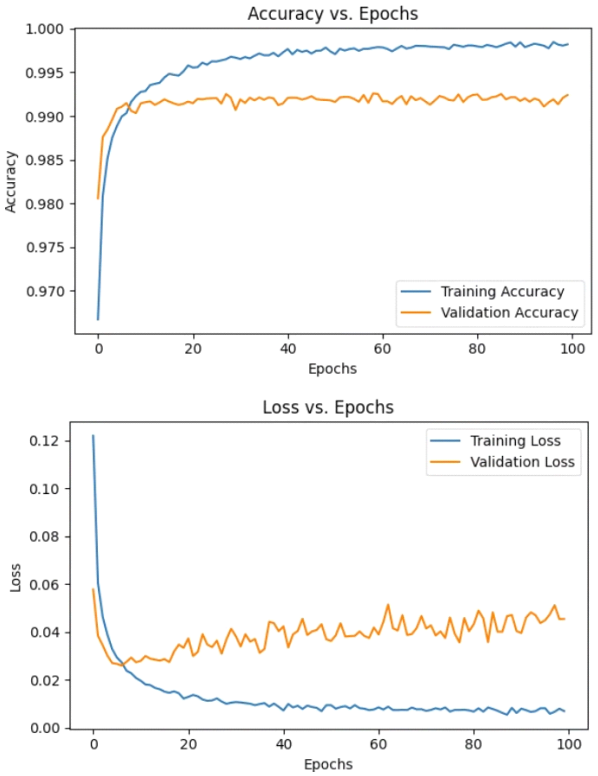

그림 8. CNN(convolutional neural network) 식별기 검증 손실 및 정확도 추이
Figure 8. Verification loss and accuracy trends by convolutional neural network (CNN) classifier

검증 데이터셋에 대한 최종 전체 정확도는 0.99로, DNN의 0.97보다 높았다. 유형별 F1-score는 유창 1.00, 음소 반복 0.92, 단어 반복 0.73, 구 반복 0.80, 연장 0.96으로 모든 항목에서 DNN보다 향상된 성능을 보였다. 특히 단어 반복(143% 향상)과 연장(81% 향상)에서 가장 큰 성능 차이를 보여, CNN이 시간-주파수 영역의 시각적 패턴을 효과적으로 학습하였음을 시사한다. macro 평균 F1-score는 0.88로, DNN(0.66)에 비해 약 33% 향상되었다. 그림 9의 혼동행렬 분석 결과, CNN은 DNN보다 각 유형을 보다 정확하게 분류하였으며, 오분류가 발생하는 경우에도 주로 음향적으로 유사한 유형 간의 혼동에 국한되었다. 단어 반복과 구 반복 간의 일부 혼동은 남았지만, DNN에서 빈번했던

연장-유창 간 오분류는 현저히 감소하였다. 이는 CNN의 합성곱 구조가 시간적 연속성과 지역적 음향 패턴을 효율적으로 학습함으로써, 비유창성 유형의 특성을 보다 정밀하게 포착한 결과로 해석된다.

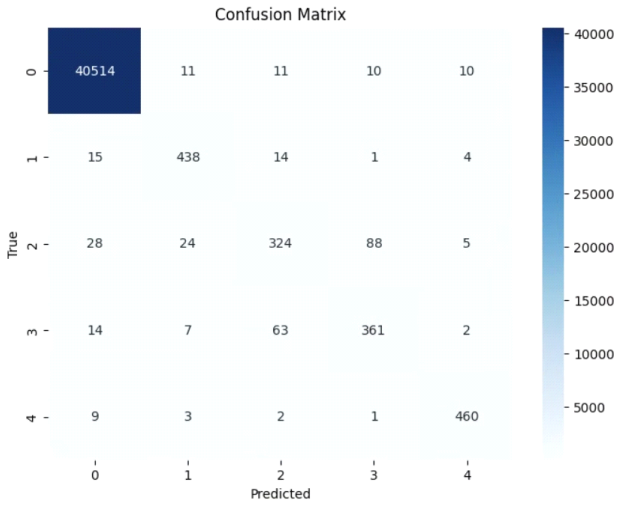

그림 9. CNN(convolutional neural network) 식별기 혼동행렬(0, 유창; 1, 음소반복; 2, 단어반복; 3, 구반복; 4, 연장)
Figure 9. Convolutional neural network (CNN) Classifier's confusion matrix (0, fluent; 1, sound repetition; 2, word repetition; 3, phase repetition; 4, prolongation)

3.3. CNN(convolutional neural network) 및 DNN(deep neural network)식별기 성능비교 결과

CNN과 DNN 식별기의 성능을 종합적으로 비교한 결과는 표 1과 같다. CNN 식별기는 모든 주요 평가 지표에서 DNN보다 높은 값을 보여, MFCCs 이미지의 시각적 특징을 학습하는 데 있어 CNN 구조의 우수성을 입증하였다. 전체 정확도는 CNN이 0.99로 DNN(0.97)보다 2%포인트 높았으며, 이는 오분류율로 환산하면 약 67% 감소에 해당한다(3%→1%). 또한 weighted average F1-score는 CNN 0.99, DNN 0.97로 비슷했지만, macro average F1-score에서 CNN(0.88)이 DNN(0.66)보다 33% 향상되어 클래스 불균형에 대한 균형적 인식 성능을 보였다.

유형별로는 CNN이 모든 비유창성 유형에서 DNN을 상회하였다. 특히 단어 반복에서 CNN의 F1-score는 0.73으로 DNN의 0.30보다 143% 높았으며, 연장에서도 0.96으로 81% 향상되었다. 구 반복은 0.80으로 DNN의 0.65보다 23% 높았고, 음소 반복 또한 0.92로 약 10% 향상되었다. 이는 CNN이 시간-주파수 평면에서 나타나는 반복과 연장의 시각적 패턴을 효과적으로 학습했음을 의미한다.

표 1. DNN & CNN 식별기 분류 성능 결과
Table 1. Results of DNN & CNN classification performance

Types of fluent and stuttered disfluencies		Precision	Recall	F1-score	Accuracy
DNN	sound repetition	0.93	0.76	0.84	0.97
	word repetition	0.55	0.21	0.30	
	phase repetition	0.72	0.59	0.65	
	prolongation	0.93	0.38	0.53	
CNN	sound repetition	0.91	0.93	0.92	0.99
	word repetition	0.78	0.69	0.73	
	phase repetition	0.78	0.81	0.80	
	prolongation	0.96	0.97	0.96	

DNN, deep neural network; CNN, convolutional neural network.

4. 논의 및 결론

본 연구는 동일한 음향특징(MFCCs)을 입력으로 하여, 반복과 연장 유형의 말더듬 발화를 자동으로 식별하는 과정에서 DNN과 CNN이 어떤 학습적 특성과 분류 성능의 차이를 보이는지를 비교·분석하는 것을 목적으로 하였다. 분석 결과, CNN은 DNN보다 전반적으로 높은 정확도, 정밀도, 재현율, F1-score를 보였으며, 특히 단어 반복과 연장 유형에서 현저한 성능 향상이 나타났다. 또한, CNN은 오분류(confusion) 양상에서도 DNN보다 명확한 구분 성향을 보여, 음향적으로 유사한 유형 간 혼동이 감소하였다. 이는 CNN의 합성곱 구조가 시간-주파수 평면에서의 지역적(local) 패턴을 효과적으로 포착하여, 각 비유창성 유형의 음향적 특성을 보다 정밀하고 안정적으로 반영할 수 있음을 실증적으로 보여주는 결과라 할 수 있다.

본 연구결과를 바탕으로 몇 가지 논의점을 기술하자면 다음과 같다. 첫째, CNN이 DNN보다 높은 인식 성능을 보인 이유는 말더듬의 핵심적 음향 단서가 시간적 연속성과 주파수 간 상관성에 의해 형성된다는 점과 밀접히 관련된다. 반복과 연장은 각각 특정 구간에서 소리가 반복되거나 길게 지속되는 등 시간적으로 연장된 특성을 지니는데, 이러한 패턴은 단순한 정적 스펙트럼 정보만으로는 충분히 포착되지 않는다. CNN은 합성곱 연산을 통해 입력 스펙트로그램 상의 인접 프레임 간 관계를 유지하면서 국소 영역의 특징을 학습하므로, 반복이나 연장처럼 시간적 패턴을 지닌 음성 변동을 세밀하게 인식할 수 있었다. 반면, DNN은 평탄화(flatten) 과정에서 시간적·공간적 구조가 소실되어 발화 내의 리듬적 변화나 지속 구간의 미세한 변동을 충분히 반영하지 못했다. 이러한 차이는 CNN의 지역적 필터링이 비유창성 탐지에 구조적으로 적합함을 시사한다.

둘째, CNN은 학습 안정성과 일반화 성능 측면에서도 우수하였다. 본 연구에서 CNN은 학습 및 검증 정확도 간의 차이가

DNN보다 현저히 작았으며, 손실 함수의 변동 폭도 적어 과적합(overfitting) 가능성이 낮았다. 이는 CNN의 가중치 공유(weight sharing)와 지역 연결(local connectivity) 구조가 학습 파라미터 수를 줄이고, 제한된 데이터에서도 안정적인 수렴을 유도했기 때문으로 해석된다. 이러한 구조적 특성은 데이터 다양성이 큰 임상 음성 환경에서도 모델의 일관성과 신뢰성을 유지할 수 있게 하며, 자동 말더듬 탐지 시스템의 실용화를 위한 중요한 기반이 된다.

셋째, DNN은 단어 반복 유형에서 낮은 재현율을 보였는데, 이는 DNN이 시간적 의존성(temporal dependency)을 명시적으로 학습하지 못하는 구조적 한계에서 비롯된 것으로 해석된다. 단어 반복은 음절 경계와 단어 길이의 리듬적 차이를 구분해야 하는데, DNN은 모든 입력을 독립적 벡터로 처리하기 때문에 발화 흐름 내 시간적 순서를 고려하지 못한다. 반면, CNN은 커널이 시간축 방향으로 이동하며 인접 구간의 변화를 포착하므로, 반복되는 단어의 주기성과 구간적 연속성을 보다 잘 반영한다. 이러한 구조적 차이는 향후 말더듬 자동 인식 연구에서 입력 신호의 시간적 연속성을 어떻게 보존하느냐가 핵심적 설계 요인임을 시사한다.

넷째, CNN의 구조적 장점은 단순히 성능 향상에 그치지 않고, 임상적 활용 가능성 측면에서도 의미가 크다. CNN은 음성의 짧은 구간 단위에서 특징을 자동으로 추출할 수 있어, 별도의 전문가 수동 분석 과정 없이 실시간 평가 시스템에 적용할 수 있다. 이는 언어재활사가 장시간의 녹음 자료를 수작업으로 분석해야 하는 기존 평가 방식의 비효율성을 해소하고, 보다 객관적이고 일관된 평가 체계를 구축하는 데 기여할 수 있다. 나아가 CNN의 지역적 특징 학습 메커니즘은 말더듬뿐 아니라 연축성 발성장애, 근긴장성 음성장애 등 다른 발화 장애 탐지에도 확장 가능성이 크다(Jo et al., 2022).

마지막으로 본 연구 결과는 데이터 효율성과 학습 자원 측면에서도 CNN의 실용적 강점을 보여준다. CNN은 동일한 입력 데이터에서 파라미터 공유를 통해 DNN보다 적은 학습 자원으로도 높은 성능을 달성할 수 있었으며, 이는 임상 현장에서 데이터가 제한적인 상황에서도 충분히 적용 가능함을 의미한다. 특히 언어재활 분야는 대규모 레이블링된 음성데이터 확보가 어려운 영역으로, CNN의 구조적 효율성은 이러한 현실적 제약을 극복할 수 있는 실질적 대안이 된다. 따라서 CNN 기반 자동 식별기는 소규모 임상데이터 환경에서도 성능 저하 없이 적용 가능한 경량화 모델(lightweight model) 개발의 기반을 제공한다는 점에서 그 의의가 크다.

본 연구에는 몇 가지 한계가 존재한다. 첫째, 사용된 LibriStutter 데이터셋이 합성 음성 기반이기 때문에 실제 말더듬 화자의 자연스러운 음성 특성을 완전히 반영하지 못한다. 향후 실제 임상 음성 자료를 포함한 데이터 확장이 필요하다. 둘째, 본 연구는 연구의 목적을 위해 MFCCs 단일 특징만을 활용하였으며, 조음 위치나 발성 강도, 성대 진동 패턴 등 다른 음향・생리학적 특징을 함께 고려하지 않았다. 이러한 다중 특징 융합(multi-modal fusion)을 적용할 경우, 말더듬 탐지의 정밀도는 더욱 향상될 가능성이 있다. 셋째, 연구의 초점이 반복과 연장 유형에 한정되어 있어, 막힘(block)이나 간투사(interjection), 수정(revision) 등 다양한 비유창성 유형에 대한 모델의 인식 능력은 검증되지 않았다. 넷째, 본 연구의 딥러닝 모델은 '블랙박스(black-box)' 구조로, CNN이 어떤 음향 단서를 근거로 판단을 내렸는지 명확히 해석하기 어렵다는 한계를 지닌다. 이는 모델 내부의 복잡한 가중치 계산 과정이 인간에게 불투명하기 때문으로, 임상적 활용 시 신뢰성과 설명 가능성(explainability)을 저하시킬 수 있다 (Castelvecchi, 2016). 이러한 한계를 보완하기 위해, Grad-CAM(gradient-weighted class activation mapping)과 같은 시각화 기반 해석 기법(Selvaraju et al., 2017)을 적용하면 모델이 주목한 시간-주파수 구간을 히트맵 형태로 확인할 수 있다. 이를 통해 모델의 판단 근거를 시각적으로 검증하고, 임상적 신뢰성을 강화할 수 있을 것이다. 다섯째, 본 연구는 상대적으로 단일 화자・영어 음성 기반으로 진행되었기 때문에, 언어적・문화적 차이에 따른 일반화 가능성을 평가하기 위한 다언어(multilingual) 확장이 필요하다. 마지막으로, 말더듬은 음성 발화뿐 아니라 다양한 신체적 수반행동을 포함하는 복합적 장애이므로, 향후 연구에서는 영상 분석을 결합한 다중 양식(multi-modal) 접근을 통해 이러한 행동까지 자동 인식할 수 있는 확장 연구가 요구된다.

이러한 한계점에도 불구하고 본 연구는 동일한 조건하에서 DNN과 CNN의 구조적 차이가 말더듬 자동 검출 성능에 미치는 영향을 실험적으로 직접 비교한 연구라는 점에서 학문적 의의가 크다. 특히 CNN의 지역적 패턴 학습이 말더듬 인식의 핵심 단서로 작용함을 실증적으로 제시함으로써, 향후 Transformer 기반 모델이나 CNN-RNN 하이브리드 구조 설계의 이론적 근거(Baevski et al., 2020; Dong et al., 2018)를 제공하였다. 또한 본 연구의 결과는 언어재활 임상 현장에서 말더듬 평가의 자동화 가능성을 높이고, 평가의 객관화・표준화에 기여할 수 있다는 점에서 실용적 가치가 있다. 향후에는 실제 임상 음성 데이터를 기반으로 한 다중 특징 통합 모델 및 설명 가능한 인공지능(explainable AI, XAI)을 적용함으로써, 인간 전문가의 판단을 보조하는 신뢰성 높은 유창성 평가 시스템으로 발전시킬 수 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.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449-12460.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, *538*(7623), 20-23.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357-366.

Dong, L., Xu, S., & Xu, B. (April, 2018). Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5884-5888). Calgary, AB, Canada.

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

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A.,⋯ Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82-97.

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. (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.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097-1105.

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.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.

Park, J. (2021). A qualitative study on the experiences of adults who stutter in recruitment process and work life. *Audiology and Speech Research*, *17*(2), 229-240.

Rabiner, L. R., & Schafer, R. W. (2010). *Theory and applications of digital speech processing*. London, UK: Pearson.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (October, 2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 618-626), Venice, Italy.

van Riper, C. (1972). *The nature of stuttering*. Hoboken, NJ: Prentice-Hall.

Yaruss, J. S. (1997). Clinical measurement of stuttering behaviors. *Contemporary Issues in Communication Science and Disorders*, *24*, 27-38.

• **박진 (Jin Park)**
가톨릭관동대학교 언어재활학과 교수
강원특별자치도 강릉시 범일로 579번길 24
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
가톨릭관동대학교 경영학과
Tel: 033-649-7266
Email: kdmis@cku.ac.kr
관심분야: 인공지능, 빅데이터, 사물인터넷, 데이터사이언스

인공지능 기반 반복과 연장 자동 인식: 심층 신경망(DNN)과 합성곱 신경망(CNN) 성능 비교

박 진[1] · 이 창 균[2]

[1]가톨릭관동대학교 언어재활학과, [2]가톨릭관동대학교 경영학과

국문초록

본 연구는 동일한 음향특징(mel-frequency cepstral coefficients, MFCCs)을 입력으로 한 DNN(deep neural network)과 CNN(convolutional neural network) 기반 말더듬 자동 식별기의 학습 특성과 성능을 비교하고자 하였다. LibriStutter 합성 음성 데이터셋을 사용하여 유창, 음소 반복, 단어 반복, 구 반복, 연장 등 5가지 발화 유형을 대상으로 두 모델을 각각 Python Keras 기반으로 구현하였다. DNN은 완전연결층(dense layer), CNN은 합성곱(convolutional layer) 구조로 설계되었으며, 성능 평가는 정확도(accuracy), 정밀도(precision), 재현율(recall), F1-score로 수행하였다. CNN 식별기는 모든 주요 성능 지표에서 DNN보다 우수한 결과를 보였으며, 특히 단어 반복과 연장 유형에서 F1-score가 각각 143%, 81% 향상되었다. CNN은 시간-주파수 평면에서의 지역적 패턴을 효과적으로 학습하여 과적합이 적고 일반화 성능이 높았다. CNN은 MFCCs 기반 말더듬 인식에서 구조적 우위를 보였으며, 자동화된 유창성 평가 시스템 개발에 유용한 기반을 제공한다. 본 연구는 DNN과 CNN의 구조적 차이가 말더듬 검출 성능에 미치는 영향을 비교한 기초 연구로서, 향후 AI 기반 언어재활 평가의 객관화에 기여할 수 있을 것이다.

핵심어: 말더듬, 인공지능, 합성곱 신경망(CNN), 심층 신경망(DNN), 멜 주파수 켑스트럼 계수(MFCCs)

참고문헌

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