



pISSN 2005-8063
eISSN 2586-5854
2026. 9. 30.
Vol.18 No.3
pp. 129-138

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

<https://doi.org/10.13064/KSSS.2026.18.3.129>



Comparative study of single and combined acoustic features for fluent and stuttered speech classification using deep neural networks

Jin Park¹ · Chang Gyun Lee^{2,*}

¹*Department of Speech Language Rehabilitation (Speech Language Rehabilitation Center),
Catholic Kwandong University, Gangneung, Korea*

²*Department of Business Administration, Catholic Kwandong University, Gangneung, Korea*

Abstract

Stuttering is characterized by disfluencies such as repetitions, prolongations, and blocks, and accurate classification of these types is important for clinical assessment. This study compared deep neural network (DNN) models using single and combined acoustic features for automatic classification of fluent and stuttered speech. A single-feature model based on Mel-frequency cepstral coefficients (MFCCs) was compared with a combined-feature model using MFCCs, Mel-spectrogram, and Chroma features. Speech data from nine Korean adults who stutter were segmented and labeled as fluent, repetition, or block. Both models achieved an overall accuracy of 95.68%. The fluent class showed high performance in both models, with a precision of 0.99, recall of 1.00, and F1-score of 1.00. For repetition, the combined-feature model increased recall from 0.50 to 1.00 and F1-score from 0.50 to 0.67. For block, precision increased from 0.67 to 1.00, whereas recall decreased from 0.60 to 0.30, resulting in an F1-score decrease from 0.63 to 0.46. The macro-average F1-score was 0.71 for both models. These findings suggest that combining acoustic features may affect classification performance differently across disfluency types. The study highlights the importance of acoustic feature design and data balance and supports further validation using larger datasets and diverse model architectures.

Keywords: stuttering, deep neural network (DNN), acoustic features, Mel-frequency cepstral coefficient (MFCC), Mel-spectrogram, Chroma

* kdmis@cku.ac.kr, Corresponding author

Received 10 May 2026; Revised 15 September 2026; Accepted 16 September 2026

© Copyright 2026 Korean Society of Speech Sciences. This is an Open-Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. 서론

말더듬(stuttering)은 발화의 시간적 흐름이 비정상적으로 단절되거나 왜곡되는 유창성 장애로, 반복(repetition), 연장(prolongation), 막힘(block)과 같은 다양한 비유창성 유형을 특징으로 한다(van Riper, 1972). 이러한 비유창성은 단순한 발화 문제를 넘어 개인의 의사소통 효율성 저하, 대인관계 위축, 정서적 부담 등 다양한 부정적 영향을 초래할 수 있기 때문에, 이에 대한 정확한 평가와 중재는 임상적으로 매우 중요한 의미를 가진다(Bloodstein & Bernstein Ratner, 2008). 특히 말더듬 평가에서는 비유창성 유형을 정확하게 식별하고 그 빈도와 양상을 분석하는 과정이 핵심적인 절차로 간주된다. 그러나 전통적인 말더듬 평가는 주로 언어재활사의 청지각적 판단에 의존하고 있어 평가자의 경험과 숙련도에 따라 결과의 신뢰도와 일관성이 달라질 수 있으며, 분석 과정에 상당한 시간과 노력이 요구된다는 한계를 가진다(Ingham & Cordes, 1997). 또한 반복 측정이나 대규모 데이터 분석이 어려워 객관적이고 정량적인 평가 체계를 구축하는 데에도 제약이 존재한다. 이러한 한계를 보완하기 위한 대안으로, 최근에는 인공지능 및 음성 처리 기술을 활용한 말더듬 자동 분류 연구가 활발히 이루어지고 있다.

기계학습을 활용한 말더듬 자동 분류 연구는 다양한 음성 특징과 분류 알고리즘을 기반으로 발전해 왔다. Howell et al.(1997)은 말더듬 아동의 발화 데이터를 기반으로 에너지 피크 수와 지속 시간 등의 음향 특징을 활용하여 인공신경망 기반 분류를 수행하였으며, Ravikumar et al.(2009)은 MFCC(Mel-frequency cepstral coefficient)를 활용한 서포트 벡터 머신을 통해 음성 반복을 자동으로 식별하였다. 이후 Mishra et al.(2021)은 MFCC 기반 인공신경망 모델을 적용하여 유창성과 비유창성을 구분하는 분류 성능을 검증하였다. 이처럼 다양한 음향 특징과 기계학습 기법을 활용한 연구들은 말더듬 자동 분류의 가능성과 유효성을 실증적으로 보여주었다. 또한 Sheikh et al.(2022)에 따르면 MFCC는 가장 널리 활용된 특징이며, SVM, ANN뿐 아니라 랜덤 포레스트, k-NN, HMM 등 다양한 알고리즘이 적용되어 왔다. 최근에는 CNN, RNN, Transformer 기반 모델을 활용하여 음성 신호의 시간적·비선형적 특성을 반영하려는 연구도 증가하고 있다(Bhushan et al., 2021; Mendelev et al., 2020; Park & Lee, 2023). 이와 같이 모델과 알고리즘 측면에서의 발전에도 불구하고, 말더듬 발화 자체가 가지는 비유창성의 복합적 특성을 충분히 반영하는 특징 설계 문제는 여전히 중요한 과제로 남아 있다.

말더듬 발화는 반복, 연장, 막힘과 같이 서로 다른 음향적 특성을 가지는 비유창성 유형으로 구분되며, 각 유형은 시간적 패턴, 에너지 분포, 발화 지속성 등에서 상이한 양상을 보인다. 예를 들어 반복은 짧은 단위의 빠른 시간적 반복성을, 막힘은 무음 또는 저에너지 구간을, 연장은 특정 음소의 비정상적인 지속을 특징으로 한다(Yairi & Ambrose, 2013). 이러한 차이는 단일 음향 특징만으로는 충분히 포착하기 어렵고, 다양한 음향 정보를 통합적으로 반영할 필요성을 시사한다. 이와 같은 맥락에서, 서로 보완적인 음향 특징을 결합하여 음성 신호를 보다 풍부하

게 표현하려는 접근이 요구된다. MFCC는 스펙트럼 포락선(spectral envelope)을 기반으로 음성의 전반적인 음색 정보를 요약하는 특징이며(Tracey et al., 2023), 멜 스펙트로그램(Mel-spectrogram)은 시간에 따른 주파수 에너지 분포를 반영하여 반복이나 막힘과 같은 시간-주파수적 변화를 효과적으로 포착할 수 있다(O'Shaughnessy, 2000). 또한 크로마(Chroma) 특징은 주파수 성분을 피치 클래스 단위로 표현함으로써 음성의 하모닉 구조와 피치 정보를 반영한다(McFee et al., 2015). 따라서 이러한 특징들을 결합할 경우 음성 신호의 상보적인 정보를 동시에 반영할 수 있으며, 복합적인 특성을 가지는 말더듬 발화 분석에 보다 적합한 표현을 제공할 수 있다. 그럼에도 불구하고 기존 연구에서는 다양한 특징과 모델이 활용되어 왔음에도, 동일한 모델 구조를 통제된 상태에서 입력 음향 특징의 구성 방식, 즉 단일 특징과 다중 특징 결합 간의 성능 차이를 체계적으로 비교한 연구는 제한적이다. 특히 유창성과 비유창성을 구분하는 수준을 넘어, 반복, 연장, 막힘과 같은 세부 비유창성 유형을 대상으로 한 다중 클래스 분류에서 특징 결합의 효과를 실증적으로 검증한 연구는 충분히 축적되지 않았다.

이에 본 연구에서는 MFCC 단일 특징 기반 심층신경망(deep neural network, DNN) 모델과 MFCC, Mel-spectrogram, Chroma를 결합한 다중 특징 기반 DNN 모델을 동일한 조건에서 구축하고, 유창성과 말더듬 비유창성 유형 분류 성능을 비교·분석하고자 한다. 이를 통해 음향 특징 결합 방식이 분류 성능에 미치는 영향을 검증하고, 효과적인 특징 구성 전략을 제시하는 것을 목적으로 한다. 이를 위해 다음과 같은 연구 문제를 설정하였다.

첫째, MFCC 단일 특징 기반 DNN 모델은 유창, 반복 및 막힘을 분류하는 데 있어 어느 정도의 성능(정확도, 정밀도, 재현율, F1-score)을 보이는가?

둘째, MFCC, Mel-spectrogram, Chroma를 결합한 다중 특징 기반 DNN 모델은 유창, 반복 및 막힘의 분류에서 어떠한 성능을 보이는가?

셋째, 두 모델 간 성능 차이는 유창, 반복 및 막힘의 각 범주에서 어떠한 양상을 보이며, 특히 반복과 막힘의 분류 성능에는 어떠한 차이가 나타나는가?

본 연구는 단일 음향 특징과 다중 음향 특징 결합 간의 성능 비교를 통해 말더듬 자동 분류 시스템 설계에서 특징 구성 방식의 중요성을 실증적으로 제시하고, 향후 인공지능 기반 언어재활 평가 도구 개발을 위한 기초 자료를 제공한다는 점에서 의의를 가진다.

2. 연구방법

2.1. 음성 데이터 수집 단계

본 연구에서는 말더듬 비유창성 자동 분류 모델 개발을 위해 말더듬 화자 9명과 정상 화자 9명, 총 18명을 대상으로 음성 데이터를 수집하였다. 말더듬 화자는 모두 남성으로 평균 연령은

34세(표준편차=5.32세)이며, 중증도는 고도(severe) 5명, 중도(moderate) 2명, 경도(mild) 2명으로 구성되었다. 정상 화자 역시 모두 남성으로 평균 연령은 37세(표준편차=7.43세)이며, 말더듬이나 기타 언어 및 의사소통 장애가 없는 성인으로 구성하였다. 음성 데이터는 방음 처리된 실험실 환경에서 수집되었으며, 모든 참여자는 동일한 읽기 과제를 수행하였다. 읽기 자료는 총 143어절(404음절)로 구성된 고등학교 수준의 구절 자료를 활용하였다(Park & Lee, 2023). 음성 신호는 16 kHz의 샘플링 주파수로 녹음되었으며, WAV 형식으로 저장되었다.

2.2. 음성 데이터 전처리 단계

수집된 WAV 형식의 음성 데이터는 어절 단위로 분절하기 위해 Google Cloud 기반 STT(speech-to-text) 시스템을 활용하였다. Google Cloud STT의 어절 단위 타임스탬프를 기준으로 자동 분절을 수행하는 과정에서, 막힘 구간의 무음-저에너지 신호가 인식되지 않거나 반복된 음절이 하나의 어절로 병합되어 인식되는 오류의 경우 연구자가 음성 파형 및 스펙트로그램을 육안으로 대조하여 어절 경계를 수동으로 재설정하였으며, 반복 음절이 누락되거나 병합된 경우 해당 구간을 별도로 분리하여 재라벨링하였다. 단어별 음성 인식 과정에서 제공되는 타임스탬프 정보를 기준으로 전체 음성을 자동 분절하였으며, 인식 오류가 발생한 구간은 수작업으로 보정하였다. 어절 분절 결과, 정상 화자는 평균 143어절(표준편차=1.73), 말더듬 화자는 평균 133어절(표준편차=12.31)로 나타났다. 이는 말더듬 화자의 경우 발화의 비유창성으로 인해 자동 분절의 일관성이 상대적으로 낮을 수 있음을 시사한다.

이후 분절된 음성 데이터에 대해 라벨링을 수행하였다. 라벨링은 말더듬 평가 및 중재 경험이 10년 이상인 1급 언어재활사 3인이 참여하여 진행하였으며, 각 샘플을 유창(fluent), 반복(repetition), 막힘(block), 연장(prolongation)의 네 가지 범주로 분류하였다. 라벨링의 신뢰도를 확보하기 위해 평가자 간 합의를 통해 최종 라벨을 결정하였다. 라벨링 과정에서 3인 평가자 간 합의를 통해 최종 라벨을 결정하였으며, 합의 도출 과정에서 다음 세 가지 기준에 해당하는 어절은 분석 대상에서 제외하였다. 첫째, 3인 중 2인 이상이 동일 라벨로 분류하지 못한 어절, 둘째, 읽기 과제와 무관한 구간(발화 재시작, 머뭇거림 등)으로 판단된 어절, 셋째, STT 자동 분절 오류로 인해 경계가 불명확한 어절이 제외되었다. 이러한 과정을 거쳐 말더듬 화자 9명의 전체 발화(약 1,197어절) 중 최종적으로 545개의 음성 샘플이 확보되었으며, 유창 477건, 반복 37건, 막힘 26건, 연장 5건으로 나타났다. 정상 화자 9명의 발화는 라벨링 기준 수립을 위한 참조 목적으로 수집되었으며, 모델 학습 및 평가에는 포함되지 않았다. 그러나 연장 유형은 전체 데이터에서 차지하는 비율이 매우 낮아 안정적인 학습이 어렵다고 판단되어 해당 5건을 제외하였으며, 최종적으로 540건을 대상으로 모델 학습 및 성능평가를 수행하였고 표 1은 전체 음성 데이터셋의 구성을 제시한다.

표 1. 말더듬 화자 음성 데이터셋 구성

Table 1. Frequency and its overall rate of fluent and stuttered disfluencies in the dataset

Types of speech sample	Frequency (N)	Overall rate (%)
Fluent (F)	477	88.3
Repetition (R)	37	6.9
Block (B)	26	4.8
Total	540	100.0

2.3. 음성 데이터 특징 추출 단계

특징 추출 단계에서는 분절된 음성 데이터를 기계학습 모델의 입력으로 활용하기 위해 음향 특징을 추출하였다. 본 연구에서는 음성 신호의 다양한 음향학적 특성을 반영하기 위해 MFCC, Mel-spectrogram, Chroma의 세 가지 특징을 사용하였다. 특징 추출 이후에는 세 가지 특징을 결합하여 하나의 통합 입력 데이터로 구성하였다. 특징 추출에는 librosa 0.9.2를 사용하였다. MFCC와 Chroma 특징은 librosa 기본 스케일을 그대로 사용하였으며, Mel-spectrogram은 librosa.power_to_db()를 적용하여 power spectrum을 dB 스케일로 변환하였다. Chroma 특징은 단시간 푸리에 변환(short-time Fourier transform, STFT) 기반의 librosa.feature.chroma_stft()를 사용하여 산출하였다. 세 가지 특징을 결합하기 전, 특징 벡터 간 스케일 차이를 제거하기 위해 StandardScaler를 적용하였다. 구체적으로 전체 데이터를 학습(70%)·평가(30%) 세트로 분할한 후, 학습 데이터에 대해서만 StandardScaler를 적합(fit)하고, 평가 데이터에는 학습 데이터의 평균과 표준편차를 적용하여 변환(transform)하였다. 이를 통해 평가 데이터의 통계값이 학습 전처리에 영향을 미치는 정보 누출을 방지하였다. 이후 NumPy의 hstack 함수를 활용하여 각 특징 벡터를 수평 방향으로 결합하였다. 구체적으로 각 음향 특징(MFCC, Chroma, Mel-spectrogram)은 시간 프레임에 걸쳐 평균값으로 집계하여 고정 차원 벡터로 변환한 후, Python NumPy의 hstack 함수를 이용해 하나의 240차원 벡터로 수평 결합(concatenation)하였다. 이 과정에서 개별 시간 프레임의 순서 정보는 평균화로 인해 소실된다. 이를 통해 MFCC, Mel-spectrogram, Chroma가 포함된 다중 특징 벡터를 생성하였다(그림 1). 이러한 특징 결합 방식은 서로 다른 음향 정보를 동시에 반영함으로써 단일 특징 기반 접근에 비해 보다 풍부한 음성 표현을 가능하게 한다.

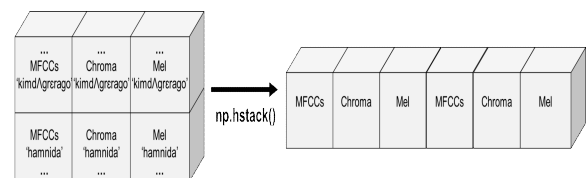


그림 1. MFCCs(Mel-frequency cepstral coefficients)-Mel spectrogram-Chromas 수평 방향 행렬 결합

Figure 1. Mel-frequency cepstral coefficient (MFCC)-Mel spectrogram-Chromas horizontal matrix concatenation

음성데이터 전처리에서는 유창, 반복 및 막힘 유형 분류를 위해 음성 데이터로부터 세 가지 음향 특징을 추출하였다. 공통 파라미터 설정으로, 시간 영역의 신호를 주파수 영역으로 변환하기 위한 샘플링 레이트(sr)는 원본 음성과 동일한 16,000 Hz(16 kHz)로 설정하였다. 푸리에 변환 창 크기를 결정하는 n_fft 는 550 samples(약 34.4 ms), hop_length 는 220 samples(약 13.75 ms)로 설정하였으며, 이는 동일 연구진의 선행연구(Park & Lee, 2023)에서 한국어 말더듬 음성 분석에 적용한 파라미터를 준용하였다. 선행연구에서 해당 값은 16 kHz 한국어 말더듬 음성의 주파수 에너지 분포를 안정적으로 추출하기 위해 경험적으로 설정된 값으로, 본 연구에서도 동일한 음성 데이터 조건에서의 일관성 확보 및 결과 비교 가능성을 위해 동일하게 적용하였다. 분석 창 크기(win_length)는 별도로 지정하지 않았으며, 이 경우 librosa 기본값에 따라 $win_length = n_fft = 550$ samples로 자동 적용된다. 이는 분석 창 전체에 FFT를 적용하는 방식으로, 제로패딩 없이 실제 신호만을 기반으로 주파수 성분을 추출하는 일반적인 설정으로 실시하였다. 첫째, MFCC는 인간의 청각 인지 특성을 모사하는 특징으로, 말더듬의 반복 및 막힘 구간에서 나타나는 발성 패턴의 차이를 효과적으로 포착하고자 하였다. 추출되는 MFCC 계수의 수를 결정하는 파라미터인 n_mfcc 는 음성의 세부적인 조음 정보를 충분히 반영하기 위해 100으로 설정하였으며, 각 계수는 시간 프레임에 걸쳐 평균값으로 집계되어 100 차원의 특징 벡터로 표현하고자 하였다. n_mfcc 는 100으로 설정하였으며, 이는 동일 연구진의 선행연구(Park & Lee, 2023)에서 한국어 말더듬 음성의 세부적인 조음 패턴 분석을 위해 경험적으로 설정한 값을 준용하였다. 일반적으로 음성 분석에서는 12, 13, 20, 40개의 MFCC 계수를 사용하는 경우가 많으나, 선행연구에서는 말더듬 발화의 반복 및 막힘 구간에서 나타나는 미세한 조음 변화를 포착하기 위해 보다 높은 차원의 계수를 적용하였으며, 본 연구도 동일한 조건에서 이를 준용하였다. 다만 $n_mfcc=100$ 은 소규모 데이터셋 환경에서 과적합 위험이 수반될 수 있는 설정값임을 인지하고 있으며, 이를 부분적으로 완화하기 위해 각 은닉층 이후 드롭아웃(dropout=0.2)을 적용하였다. 그러나 드롭아웃 적용만으로 과적합이 완전히 해소되었다고 단정하기 어려우며, 이는 본 연구의 제한점 중 하나로, 향후 더 많은 데이터 확보 및 차원 비교 실험을 통한 검증이 필요하다. 둘째, Chroma 특징은 12개의 음계(pitch class)에 대한 에너지 분포를 나타내며, 말더듬 환자의 음성에서 나타나는 음정(pitch)의 불규칙한 변동 및 반복 구간의 음계 패턴 변화를 분석하는 데 활용하고자 하였다. Chroma 특징의 음계 분류 수를 결정하는 n_chroma 는 반음계(chromatic scale) 구성에 따라 기본값인 12로 설정하였으며, STFT 결과를 기반으로 추출하여 12차원의 특징 벡터로 표현하고자 하였다. 셋째, Mel spectrogram은 Mel 스케일로 변환된 주파수 에너지 분포를 나타내며, 막힘 구간에서 발생하는 특정 주파수 대역의 에너지 집중 현상 및 반복 구간의 주기적 에너지 패턴을 포착하는 데 효과적이다. Mel 필터뱅크의 수를 결정하는 n_mels 는 음성 신호의 주파수 해상도와 계산 효율성 간의 균형을 고려하여 128로 설정하였으며, 시간 프레임에

걸친 평균값으로 집계되어 128차원의 특징 벡터로 표현하고자 하였다. 최종적으로 세 가지 특징을 연결(concatenate)하여 총 240차원(MFCC 100+Chroma 12+Mel 128)의 특징 벡터를 구성하였으며, 이를 말더듬 유형 분류 모델의 입력 데이터로 전처리하여 사용하였다.

2.4. 모델 학습 단계

모델 학습 단계에서는 추출된 음향 특징을 입력으로 하여 유창, 반복 및 막힘을 분류하는 DNN 모델을 구축하였다. 본 연구에서는 입력 특징에 따른 성능 차이를 비교하기 위해 동일한 은닉층 구조를 유지한 상태에서 입력 데이터만을 달리하는 실험을 설계하였다. 단, 두 모델은 은닉층 구조는 동일하나 입력 차원이 다르므로 학습 파라미터 수에 차이가 있다. MFCC 단일 특징 모델(입력 100차원)의 총 파라미터 수는 약 67,203개이며, 다중 특징 결합 모델(입력 240차원)의 총 파라미터 수는 약 103,043개로 단일 특징 모델 대비 약 53% 많다. 따라서 두 모델 간 성능 차이는 음향 특징의 정보량 차이뿐 아니라 모델 용량 차이에도 기인할 수 있어, 특징 정보량의 효과와 모델 용량의 효과를 완전히 분리하기 어렵다는 점을 결과 해석 시 고려할 필요가 있다.

모델은 입력층, 3개의 은닉층, 출력층으로 구성된 완전 연결 신경망 구조를 사용하였다. 각 은닉층의 노드 수는 각각 256, 128, 64로 설정하였으며, 활성화 함수로 ReLU를 적용하였다. 과적합 방지를 위해 각 은닉층 이후에 드롭아웃(dropout)을 적용하였으며, dropout 비율은 0.2로 설정하였다. 출력층에서는 Softmax 함수를 사용하여 각 클래스(유창, 반복, 막힘)에 대한 확률값을 산출하도록 하였다. 모델 학습을 위한 손실 함수는 categorical cross-entropy를 사용하였으며, 최적화 알고리즘으로는 Adam optimizer를 적용하였다(Goodfellow et al., 2016). 학습은 총 100 epoch 동안 수행하였고, 배치 크기는 32로 설정하였다. 전체 540개 샘플은 화자 구분 없이 어절 단위로 무작위 70:30 비율로 학습 데이터와 평가 데이터로 분할하였다. 이에 따라 동일 화자의 발화가 학습 데이터와 평가 데이터에 동시에 포함될 수 있으며, 이러한 분할 방식은 화자 독립적인 일반화 성능을 평가하는 데 한계가 있다. 또한 본 연구에서는 소규모 데이터셋의 특성상 별도의 검증 데이터셋을 구성하지 않고, 학습 과정에서 모델의 수렴 추이를 확인하기 위해 평가 데이터를 검증 목적으로 활용하였다. 따라서 본 연구에서의 평가 데이터는 모델 학습 과정에서의 검증에도 사용되었으므로, 최종 모델의 성능을 완전히 독립적인 데이터로 평가한 것으로 보기 어렵다는 제한점이 있다. 이러한 점을 고려하여 본 연구의 결과는 탐색적 연구 결과로 해석할 필요가 있다. 한편, 입력 특징에 따른 성능 차이를 비교하기 위해 동일한 데이터 분할 조건과 동일한 딥러닝 모델 구조 및 학습 조건에서 MFCC 단일 특징 데이터와 MFCC, Mel-spectrogram 및 Chroma를 결합한 다중 특징 데이터를 각각 학습하고 성능을 비교하였다. 그림 2는 MFCC 단일 특징 데이터와 다중 특징 결합 데이터에 대한 동일한 모델 구조에 대한 구체적 형태를 보여주고 있다.

Model: "sequential_13"		
Layer (type)	Output Shape	Param #
module_wrapper_89 (ModuleWrapper) (None, 256)		61696
module_wrapper_90 (ModuleWrapper) (None, 256)		0
module_wrapper_91 (ModuleWrapper) (None, 128)		32896
module_wrapper_92 (ModuleWrapper) (None, 128)		0
module_wrapper_93 (ModuleWrapper) (None, 64)		8256
module_wrapper_94 (ModuleWrapper) (None, 64)		0
module_wrapper_95 (ModuleWrapper) (None, 3)		195
Total params: 103,043		
Trainable params: 103,043		
Non-trainable params: 0		

그림 2. 단일 특징과 다중 특징 데이터에 사용한 동일 DNN(deep neural network) 모델

Figure 2. The same deep neural network (DNN) model used for both single-feature (MFCC) and combined-feature (MFCC-Mel spectrogram-Chroma) data

2.5. 출력 및 평가 단계

출력 단계에서는 학습된 DNN 모델을 활용하여 각 음성 샘플의 유창, 반복 및 막힘 가운데 어느 범주에 해당하는지를 예측하였다. 모델은 입력된 음성 특징 벡터에 대해 각 클래스에 대한 확률값을 출력하며, 가장 높은 확률을 갖는 클래스를 최종 예측 결과로 선택하였다.

모델의 분류 성능 평가는 정확도(accuracy), 정밀도(precision), 재현율(recall), F1-score를 기반으로 수행하였다. 정확도는 전체 샘플 중 올바르게 분류된 비율을 의미하며, 정밀도는 특정 클래스로 예측된 샘플 중 실제 해당 클래스에 속하는 비율을 나타낸다. 재현율은 실제 특정 클래스에 속하는 샘플 중 모델이 올바르게 분류한 비율을 의미한다. F1-score는 정밀도와 재현율의 조화 평균으로, 불균형 데이터 환경에서 모델 성능을 종합적으로 평가하는 지표로 활용된다.

본 데이터는 클래스 간 불균형이 존재하므로 정확도뿐만 아니라 정밀도, 재현율 및 F1-score를 함께 고려하여 모델 성능을 평가하였다. 또한 혼동행렬(confusion matrix)을 활용하여 클래스 간 오분류 패턴을 분석하였다. 이를 통해 MFCC 단일 특징 모델과 다중 특징 결합 모델 간의 분류 성능 차이를 비교·검증하였다.

3. 연구결과

3.1. 단일 특징(MFCC) DNN 모델 성능 결과

MFCC 단일 특징을 입력으로 사용한 DNN 모델의 성능을 평가한 결과, 전체 정확도(accuracy)는 95.68%, 손실값(loss)은 0.0995로 나타났다. 그러나 이러한 정확도는 전체 데이터에서 유창 발화가 높은 비율을 차지하는 데이터 분포의 영향을 크게 반영한 결과로, 모델의 실제 분류 성능을 충분히 설명하는 데에는 한계가 있다.

클래스별 성능을 살펴보면, 유창(class 0)의 경우 정밀도 0.99, 재현율 1.00, F1-score 1.00으로 매우 높은 성능을 나타냈다. 반면 비유창성 유형에서는 전반적으로 낮은 성능이 나타났다. 반복(class 1)의 경우 정밀도 0.50, 재현율 0.50, F1-score 0.50로 나타나 절반 수준의 분류 성능을 보였으며, 이는 반복의 일부가 막힘으로 오분류되었음을 의미한다. 막힘(class 2)의 경우 정밀도 0.67, 재현율 0.60, F1-score 0.63으로 나타나 반복에 비해 상대적으로 높은 성능을 보였으나, 일부 샘플이 반복과 유창으로 오분류되는 경향이 확인되었다.

Macro average 기준에서는 정밀도 0.72, 재현율 0.70, F1-score 0.71로 나타났다. 이러한 결과는 유창에서는 매우 높은 분류 성능을 보인 반면, 반복과 막힘에서는 상대적으로 낮은 성능이 나타났음을 보여준다. 혼동 행렬에서도 반복과 막힘 간의 오분류와 함께 일부 비유창성 발화가 유창으로 분류되는 양상이 확인되었다. 그림 3은 단일 특징 DNN 모델의 검증 손실 및 정확도 추이를, 그림 4는 혼동 행렬을 제시한다.

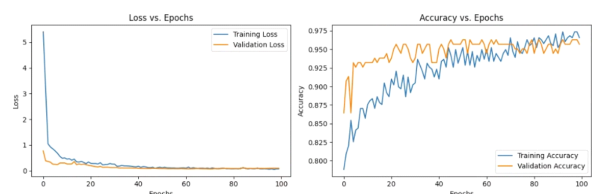


그림 3. 단일 특징 DNN(deep neural network) 모델의 검증 손실 및 정확도 추이. 손실값(loss) = 0.0995, 정확도(accuracy) = 0.9568.

Figure 3. Validation loss and accuracy trends of the single-feature (MFCC). Loss=0.0995, accuracy=0.9568.

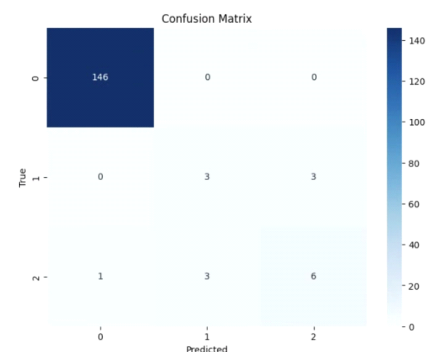


그림 4. 단일 특징 DNN(deep neural network) 모델의 혼동 행렬
Figure 4. Confusion matrix of the single feature (MFCC) deep neural network (DNN) model

3.2. 다중 특징 결합(MFCC-Mel spectrogram-Chroma)

DNN 모델 성능 결과

MFCC, Mel-spectrogram, Chroma를 결합한 다중 특징 DNN 모델의 성능을 평가한 결과, 전체 정확도는 95.68%, 손실값은 0.0902로 나타나 단일 특징 모델과 유사한 결과를 보였다. 단일 특징 모델과 동일하게 해당 정확도 역시 유창의 높은 비율이 반영된 값이라는 점에서 해석에 주의가 필요하다.

클래스별 성능을 살펴보면, 유창(class 0)은 정밀도 0.99, 재현

율 1.00, F1-score 1.00으로 매우 높은 성능을 유지하였다. 반복(class 1)의 경우 정밀도 0.50, 재현율 1.00, F1-score 0.67로 나타나, 단일 특징 모델과 비교 시 실제 반복 발화를 모두 반복으로 탐지하여 재현율 1.00을 달성하였으나, 그러나 일부 막힘 역시 반복으로 오분류되어 정밀도는 0.50에 그쳤다. 막힘(class 2)의 경우 정밀도 1.00, 재현율 0.30, F1-score 0.46으로 나타났다. 막힘으로 예측한 샘플은 모두 실제 막힘이었으나, 실제 막힘 중 일부만 올바르게 분류되고 나머지는 주로 반복 또는 유창으로 오분류되어 재현율이 0.30으로 낮게 나타났다.

Macro average 기준에서는 정밀도 0.83, 재현율 0.77, F1-score 0.71로 나타났다. 즉, 다중 특징 결합 모델에서는 반복의 재현율과 F1-score가 높아진 반면, 막힘의 재현율과 F1-score는 낮아져 특징 결합에 따른 변화가 유형별로 서로 다르게 나타났다. 혼동 행렬 분석에서도 다중 특징 결합 모델은 단일 특징 모델에 비해 반복의 탐지율이 높아진 반면, 막힘이 반복으로 오분류되는 사례가 증가하는 양상이 확인되었다. 그림 5는 다중 특징 결합 DNN 모델의 검증 손실 및 정확도 추이를, 그림 6은 혼동 행렬을 보여주고 있다.

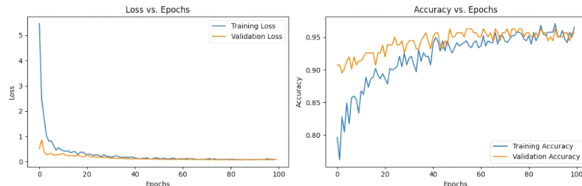


그림 5. 다중 특징 DNN(deep neural network) 모델의 검증 손실 및 정확도 추이. 손실값(loss)=0.0902, 정확도(accuracy)=0.9568.

Figure 5. Validation loss and accuracy trends of the combined feature (MFCC-Mel spectrogram-Chroma) deep neural network (DNN) model. Loss=0.0902, accuracy=0.9568.

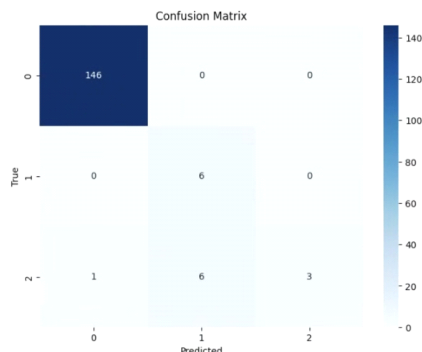


그림 6. 복합 특징 결합 DNN(deep neural network) 모델의 혼동 행렬

Figure 6. Confusion matrix of the combined feature (MFCC-Mel spectrogram-Chroma) deep neural network (DNN) model

3.3. 두 모델의 성능 비교 결과

단일 특징(MFCC) 모델과 다중 특징 결합 모델 간의 성능을 비교한 결과, 두 모델의 전체 정확도는 모두 95.68%로 동일하였다(표 2). 그러나 이 수치는 유창의 높은 비율에 영향을 받은 결과로, 보다 의미 있는 성능 차이는 비유창성 유형에서 나타났다. 소표본 조건 예비적 비교이므로 결과 해석의 유의가 필요하다. 구체적으로 반복(class 1)의 경우 단일 특징 모델 대비 다중

특징 결합 모델에서 F1-score가 0.50에서 0.67로 높아지는 경향이 관찰되었으며, 막힘(class 2)의 경우 F1-score가 0.63에서 0.46으로(재현율 0.60→0.30) 낮아지는 경향이 관찰되었다. 다만 시험표본이 반복 6건, 막힘 10건에 불과하여 이러한 변화는 소수 사례의 분류 결과 변동에 따른 것일 수 있으며, 현재 표본 크기로는 통계적으로 유의미한 향상 또는 저하로 단정하기 어렵다. 이에 화자 단위 반복 교차검증 및 신뢰구간 추정을 통해 해당 경향의 재현성을 추가로 검증할 필요가 있다.

Macro average 기준에서 두 모델의 F1-score는 모두 0.71로 동일하였으며, 클래스 불균형을 고려한 상황에서 전반적인 성능 차이는 나타나지 않았다. 혼동 행렬 분석에서도 반복 유형에서는 분류 성능 향상이 확인된 반면, 막힘 유형에서는 오분류가 증가하는 경향이 나타나 다중 특징 결합이 모든 비유창성 유형에 균일하게 기여하지는 않음을 시사한다.

표 2. 단일 특징(MFCC) vs. 다중 특징 결합(MFCC-Mel spectrogram-Chroma) DNN 모델의 성능 비교 결과

Table 2. Results of single feature (MFCC) vs combined feature (MFCC-Mel spectrogram-Chroma) DNN classification performance

Model	Types of fluent and stuttered speech	Precision	Recall	F1-score	Accuracy
MFCC DNN	Fluent (0)	0.99	1.00	1.00	0.96
	Repetition (1)	0.50	0.50	0.50	
	Block (2)	0.67	0.60	0.63	
	Macro Avg	0.72	0.70	0.71	
	Weighted Avg	0.95	0.96	0.96	
Combined DNN	Fluent (0)	0.99	1.00	1.00	0.96
	Repetition (1)	0.50	1.00	0.67	
	Block (2)	1.00	0.30	0.46	
	Macro Avg	0.83	0.77	0.71	
	Weighted Avg	0.98	0.96	0.95	

MFCC, Mel-frequency cepstral coefficient; DNN, deep neural network.

종합적으로, 다중 특징 결합 모델은 반복 유형에서의 분류 성능이 높아지는 경향을 보인 반면, 막힘 유형에서는 성능이 낮아지는 경향이 나타나 특징 결합의 효과가 클래스별로 상이하게 작용할 수 있음을 시사한다.

4. 논의

본 연구는 MFCC 단일 특징 기반 DNN 모델과 MFCC, Mel-spectrogram, Chroma를 결합한 다중 특징 기반 DNN 모델의 말더듬 비유창성 분류 성능을 비교하였다. 두 모델은 유창, 반복 및 막힘을 대상으로 동일한 조건에서 학습, 평가되었으며, 연구 결과를 바탕으로 다음과 같이 몇 가지 논의점을 기술하고자 한다.

4.1. 다중 음향 특징 결합의 효과

본 연구 결과, 다중 음향 특징 결합 모델은 MFCC 단일 특징 모델과 전체 정확도에서는 동일한 성능을 보였으나, 비유창성 유형별 분류 성능에서는 서로 다른 양상이 나타났다. 특히 반복 유형의 경우 단일 특징 모델에서 F1-score가 0.50이었으나 다중 특징 결합 모델에서는 0.67로 향상된 반면, 막힘 유형에서는 0.63에서 0.46으로 감소하였다. 따라서 다중 음향 특징의 결합은

전체적인 분류 정확도를 향상시키기보다는 비유창성 유형에 따라 분류 성능과 오분류 양상에 차별적인 영향을 미쳤을 가능성이 있다.

이러한 결과는 각 음향 특징이 서로 다른 정보를 제공한다는 점에서 해석할 수 있다. MFCC는 음성의 스펙트럼 포락선 특성을 반영하는 반면, Mel-spectrogram은 시간에 따른 주파수 에너지 분포를, Chroma는 음성 신호의 하모닉 구조와 관련된 정보를 반영한다. 따라서 이러한 특징을 결합할 경우 MFCC만으로 표현되는 정보에 추가적인 음향 정보가 제공될 수 있으며, 본 연구에서는 이러한 정보가 반복 유형의 분류에 상대적으로 유리하게 작용했을 가능성이 있다.

그러나 다중 특징 결합 모델에서 반복의 재현율이 0.50에서 1.00으로 증가한 동시에 막힘의 재현율은 0.60에서 0.30으로 감소하였다. 이는 반복 유형의 탐지율 향상이 막힘 유형의 일부 샘플이 반복으로 분류되는 양상과 함께 나타났음을 의미한다. 즉, 추가적인 음향 특징의 결합이 모든 비유창성 유형의 분류에 동일한 방향으로 작용한 것이 아니라, 반복과 막힘 간의 분류 경계에 영향을 미쳤을 가능성이 있다. 특히 본 연구에서는 Mel-spectrogram과 Chroma를 시간 프레임의 평균값으로 변환하여 사용하였기 때문에, 막힘에서 나타나는 일시적인 발성 중단이나 지연과 같은 시간적 특징이 충분히 반영되지 않았을 가능성도 고려할 수 있다. 반면 반복적으로 나타나는 음향적 특성은 시간 프레임의 평균을 통해서도 일정 부분 반영되어 모델의 반복 유형 분류에 활용되었을 가능성이 있다.

따라서 본 연구에서 나타난 다중 음향 특징 결합의 효과는 전체적인 분류 정확도의 향상보다는 비유창성 유형별 탐지 및 오분류 양상의 변화로 이해할 필요가 있다. 특히 반복 유형의 성능 향상과 막힘 유형의 성능 저하가 동시에 나타났다는 점은 추가적인 음향 특징이 비유창성 유형에 따라 상이하게 작용할 수 있음을 보여준다. 다만 본 연구는 반복 37개와 막힘 26개의 제한된 샘플을 대상으로 하였으며, 화자 단위의 독립적인 데이터 분할이 이루어지지 않았다는 점을 고려할 때 이러한 차이를 특정 음향 특징의 효과로 단정하기는 어렵다. 따라서 본 결과는 다중 음향 특징 결합이 비유창성 유형별 분류에 미치는 영향을 탐색적으로 보여주는 결과로 해석할 필요가 있으며, 향후 보다 많은 비유창성 자료와 화자 독립적인 검증을 통해 이를 재확인할 필요가 있다.

4.2. 비유창성 유형 간 분류 난이도 및 음향적 유사성

본 연구에서는 반복과 막힘 간의 오분류가 공통적으로 나타났으며, 이는 두 비유창성 유형이 음향적으로 유사한 특성을 공유하기 때문으로 해석된다. 반복은 음절 또는 음소 단위의 빠른 재생산을 특징으로 하고, 막힘은 발화의 일시적 중단 또는 지연을 특징으로 하지만, 실제 발화에서는 두 현상이 연속적이거나 혼합된 형태(예: (막힘)가가가방, 바바...(막힘)바나나)로 나타나는 경우가 많다. 이러한 특성은 음향 신호상에서 명확한 경계로 구분하기 어렵게 만들며, 특히 MFCC와 같은 단일 특징만으로는 이러한 미세한 차이를 충분히 반영하기 어려웠던 것으로 보

인다. 실제로 단일 특징 모델의 혼동 행렬에서는 반복이 유창 또는 막힘으로 오분류되는 경향이 두드러지게 나타났다. 반면 다중 특징 결합 모델에서는 시간 프레임 평균값으로 요약된 주파수 에너지 분포 정보(Mel-spectrogram)와 하모닉 구조 정보(Chroma)가 추가됨으로써 반복 탐지율은 향상되었으나, 막힘의 반복으로의 오분류가 증가하는 부작용도 나타났다. 이는 추가된 특징이 제공하는 평균화된 주파수 에너지 분포 정보 및 하모닉 구조 정보가 반복 유형의 분류에 기여할 가능성이 있으나, 반복과 음향적으로 유사한 특성을 공유하는 막힘 유형을 반복으로 잘못 분류하게 만드는 결과로도 이어졌음을 보여준다. 이는 말더듬 비유창성 분류가 단순한 패턴 인식 문제가 아니라, 시간적·음향적 요소가 결합된 복합적 문제임을 보여주며, 향후 비유창성 유형 간 음향적 유사성을 보다 효과적으로 구분할 수 있는 특징 설계 및 모델 구조에 대한 추가 연구가 필요함을 시사한다.

4.3. 데이터 불균형과 평가 지표의 중요성

본 연구 데이터는 유창 발화가 대부분을 차지하는 불균형 구조를 가지고 있으며, 이에 따라 전체 정확도는 상대적으로 높게 나타났다. 그러나 클래스별 성능을 분석한 결과, 단일 특징 모델은 반복(F1=0.50)과 막힘(F1=0.63) 모두에서 제한적인 분류 성능을 보였다. 이는 정확도만을 기준으로 모델 성능을 평가할 경우 실제 분류 능력이 과대평가될 수 있음을 의미한다. 다중 특징 결합 모델은 반복 유형에서 F1-score가 0.50에서 0.67로 향상되었으나, 막힘 유형에서는 0.63에서 0.46으로 감소하였으며, macro average F1-score는 두 모델 모두 0.71로 동일하였다. 이는 특징 결합의 효과가 클래스별로 상이하게 작용함을 보여주며, 전반적인 성능 지표만으로는 이러한 세부적인 차이를 포착하기 어렵다는 점을 시사한다. 이러한 결과는 말더듬과 같이 클래스 간 불균형이 존재하는 문제에서는 정확도 중심 평가를 지양하고, 정밀도, 재현율, F1-score와 같은 지표를 함께 고려해야 함을 강조한다.

4.4. 한계점 및 향후 연구

본 연구는 다음과 같은 제한점을 갖는다. 첫째, 제한된 한국어 말더듬 데이터셋의 규모와 대표성이다. 본 연구는 9명의 한국어 말더듬 화자 데이터를 기반으로 수행되어 표본의 규모와 다양성 측면에서 제한이 있다. 이는 모델의 일반화 가능성에 영향을 미칠 수 있으므로, 향후에는 보다 다양한 연령, 말더듬 수준, 발화 상황 등을 포함하는 대규모 한국어 말더듬 음성 데이터셋 구축이 요구된다. 둘째, 데이터 불균형 및 특정 비유창성 유형의 데이터 부족 문제이다. 본 연구 데이터는 유창 발화가 높은 비율을 차지하고 있으며, 반복, 막힘과 같은 비유창성 유형의 데이터는 유창에 비해 매우 제한적이었다. 이러한 데이터 구조는 모델이 다수 클래스에 편향되어 과적합 학습될 가능성을 높이며, 소수 클래스의 분류 성능 저하로 이어질 수 있다. 향후 연구에서는 충분한 학습자료를 확보한 후, 데이터 증강(data augmentation), 오버샘플링(oversampling), 클래스 가중치 적용(class weight) 등의 방법을 학습자료에 한정하여 적용하고, 동일

한 독립 평가자료를 이용하여 각 방법의 효과를 비교, 검증할 필요가 있다. 셋째, 모델 구조 다양성에 대한 검증의 부족이다. 본 연구에서는 DNN 모델을 중심으로 특징 결합 효과를 검증하였으나, 다양한 모델 구조와의 비교는 이루어지지 않았다. 향후 연구에서는 CNN, RNN, Transformer 기반 모델을 적용하여 시간적, 음향적 특성을 보다 효과적으로 반영할 수 있는 모델을 탐색할 필요가 있다. 마지막으로 본 연구는 9명의 화자로 구성된 소규모 데이터셋의 특성상 학습·평가 데이터를 화자 단위로 완전히 분리하지 못하였으며, 평가 데이터를 학습 과정의 검증 데이터로 함께 활용하였다. 이로 인해 일부 화자의 음성적 특성이 학습과 평가 데이터에 중복되어 반영되었을 가능성을 배제할 수 없으며, 보고된 성능 지표는 완전히 독립적인 평가로 보기 어렵다. 따라서 본 연구의 결과는 확증적(confirmatory) 성능 비교라기보다 소표본 조건에서의 예비적(pilot)·탐색적(exploratory) 비교로 해석되어야 하며, 향후 화자 단위 분리 또는 leave-one-speaker-out 교차검증을 통한 재검증이 필요하다.

5. 결론

본 연구는 MFCC 단일 특징 기반 DNN 모델과 MFCC, Mel-spectrogram, Chroma를 결합한 다중 특징 기반 DNN 모델의 말더듬 비유창성 분류 성능을 비교하였다. 연구 결과, 전체 정확도에서는 두 모델 모두 96%로 동일하였으나, 비유창성 유형별 분류 성능에서는 의미 있는 차이가 확인되었다. 반복 유형에서는 다중 특징 결합 모델이 F1-score 0.50에서 0.67로 향상을 보인 반면, 막힘 유형에서는 F1-score가 0.63에서 0.46으로 감소하여 특징 결합의 효과가 유형별로 상이하게 나타났다.

다중 특징 결합 모델은 반복 유형의 F1-score를 높인 반면, 이는 막힘 발화 일부를 반복으로 오분류함으로써 막힘 유형의 재현율과 F1-score가 낮아지는 결과와 맞물려 있었다. 따라서 다중 특징 결합을 모든 비유창성 유형에 대한 전반적인 성능 개선으로 단정하기보다, 클래스 간 오분류 양상이 재배분된 결과로 해석하는 것이 타당하다. 특히 단일 특징 모델에서 F1-score 0.50 수준에 그쳤던 반복 유형이 다중 특징 결합 모델에서 0.67로 향상되었다는 점은, 다중 음향 특징의 결합이 어절 수준의 평균화된 스펙트럼 정보를 보완적으로 반영함으로써 특정 비유창성 유형의 분류에 영향을 미칠 수 있음을 보여주는 결과이다. 다만, 막힘 유형에서의 성능 저하는 특징 결합이 모든 비유창성 유형에 균일하게 작용하지 않을 수 있음을 보여주며, 향후 비유창성 유형별 특성을 고려한 특징 설계와 모델 구조의 최적화가 필요함을 시사한다.

본 연구는 동일한 모델 구조를 유지한 상태에서 입력 특징 구성만을 달리하여 성능 차이를 비교함으로써, 음향 특징 구성에 따른 말더듬 비유창성 분류 성능의 차이를 실증적으로 확인하였다는 점에서 의의를 가진다. 특히 다중 음향 특징 결합에 따라 전체적인 분류 정확도에는 차이가 없었으나, 비유창성 유형별 분류 성능과 오분류 양상에서는 차이가 나타났다는 점을 확인하였다. 다만 제한된 표본 수와 클래스별 자료의 불균형, 화자

단위의 독립적인 데이터 분할이 이루어지지 않은 점 등을 고려할 때, 본 연구의 결과는 음향 특징 구성의 효과를 탐색적으로 확인한 결과로 해석할 필요가 있다. 향후 연구에서는 보다 대규모의 말더듬 음성 데이터셋을 구축하고, 화자 독립적인 검증과 다양한 모델 구조를 적용하여 성능을 비교함으로써, 음향 특징 구성에 따른 효과를 재검증하고 자동 말더듬 평가 시스템의 실용화를 위한 연구가 지속적으로 이루어질 필요가 있다.

References

- 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.
- Bloodstein, O., & Bernstein Ratner, N. (2008). *A handbook on stuttering (6th ed.)*. Clifton Park, NY: Thomson Delmar Learning.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, UK: MIT Press.
- 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.
- Ingham, R. J., & Cordes, A. K. (1997). Identifying the authoritative judgments of stuttering: comparisons of self-judgments and observer judgments. *Journal of Speech Language and Hearing Research*, 40(3), 581-94. doi: 10.1044/jslhr.4003.581. PMID: 9210116.
- McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E., & Nieto, O. (2015, July). Librosa: Audio and music signal analysis in Python. *Proceedings of the 14th Python in Science Conference*, (pp. 18-24). Austin, TX.
- Mendelev, V., Raissi, T., Camporese, G., & Giollo, M. (2020, June). Improved robustness to disfluencies in RNN-transducer based speech recognition. *Proceedings of ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON.
- Mishra, N., Gupta, A., & Vathana, D. (2021). Optimization of stammering in speech recognition applications. *International Journal of Speech Technology*, 24, 679-685.
- Park, J., & Lee, C. G. (2023). AI-based 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). An approach for objective assessment of stuttered speech using MFCC features. *Digital Signal Processing*, 9(1), 19-24.
- Sheikh, S. A., Sahidullah, M., Hirsch, F., & Ouni, S. (2022). Machine

- learning for stuttering identification: Review, challenges and future directions. *Neurocomputing*, 514, 385-402.
- O'Shaughnessy, D. (2000). *Speech communication: Human and machine* (2nd ed.). New York, NY: IEEE Press.
- Tracey, B., Volfson, D., Glass, J., Haulcy, R., Kostrzebski, M., Adams, J., ... Vogel, A. (2023). Towards interpretable speech biomarkers: Exploring MFCCs. *Scientific Reports*, 13, 22787.
- van Riper, C. (1972). *Speech correction: Principles and methods* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Yairi, E., & Ambrose, N. (2013). Epidemiology of stuttering: 21st century advances. *Journal of Fluency Disorders*, 38(2), 66-87.

• **박진 (Jin Park)**

가톨릭관동대학교 언어재활학과 교수
강원특별자치도 강릉시 범일로 579번길 24
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
Tel: 033-649-7266
Email: kdmis@cku.ac.kr
관심분야: 인공지능, 빅데이터, 사물인터넷, 데이터사이언스

심층신경망 기반 유창성과 말더듬 비유창성 분류를 위한 단일 및 다중 음향 특징 비교

박진¹·이창균²

¹가톨릭관동대학교 언어재활학과, ²가톨릭관동대학교 경영학과

국문초록

말더듬은 반복, 연장, 막힘과 같은 비유창성을 특징으로 하며, 이러한 유형을 정확하게 분류하는 것은 임상적 평가에 있어 중요하다. 본 연구는 유창성과 말더듬 비유창성의 자동 분류를 위해 단일 음향 특징과 복합 음향 특징을 활용한 심층신경망(Deep Neural Network, DNN) 모델의 성능을 비교하였다. 이를 위해 MFCCs, Mel-Frequency Cepstral Coefficients를 기반으로 한 단일 특징 모델과 MFCCs, Mel-spectrogram, Chroma를 결합한 복합 특징 모델을 비교 분석하였다. 한국어 말더듬 성인 9명의 음성 데이터를 분절한 후, 이를 유창, 반복, 막힘의 세 가지 범주로 라벨링하여 실험을 수행하였다. 연구 결과, 두 모델의 전체 정확도는 모두 95.68%로 동일하게 나타났다. 유창 클래스는 두 모델 모두 정밀도 0.99, 재현율 1.00, F1-score 1.00으로 높은 분류 성능을 보였다. 반복 클래스에서는 복합 특징 모델의 재현율이 0.50에서 1.00으로, F1-score가 0.50에서 0.67로 향상된 반면, 막힘 클래스에서는 정밀도가 0.67에서 1.00으로 증가하였으나 재현율은 0.60에서 0.30으로 감소하여 F1-score가 0.63에서 0.46으로 낮아졌다. 또한 macro average F1-score는 두 모델 모두 0.71로 동일하여 전체적인 균형 성능에서는 차이가 나타나지 않았다. 이러한 결과는 음향 특징의 결합이 유창 발화의 높은 분류 성능을 유지하면서 특정 비유창성 유형의 분류 성능에 차이를 가져올 수 있음을 보여준다. 본 연구는 자동화된 말더듬 평가 시스템 개발에 있어 음향 특징의 설계와 데이터 균형의 중요성을 강조하며, 향후 대규모 데이터셋 구축과 다양한 음향 특징 및 모델 구조를 활용한 추가 검증의 필요성을 제시한다.

핵심어: 말더듬, 심층신경망, 음향 특징, MFCC, Mel-spectrogram, Chroma

참고문헌

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