음성 기반 설명가능한 마비말장애 판단 인공지능 기술 동향*

정원[1] 김장연[1] 박형민[2]
[1]서강대학교 인공지능학과
[2]서강대학교 전자공학과
wjeong@sogang.ac.kr, jykim97@sogang.ac.kr, hpark@sogang.ac.kr

Advances in Explainable AI for Speech-Based Dysarthria Detection: A Review of Technology Trends

Won Jeong[1] Jangyeon Kim[1] Hyung-Min Park[2]
[1]Dept. of Artificial Intelligence, Sogang University
[2]Dept. of Electronic Engineering, Sogang University

요 약

마비말장애는 의사소통과 삶의 질에 큰 영향을 미치는 운동성 언어장애이다. 주관적인 평가에 의존한 현재의 마비말장애 진단 방법 대신 객관적이고 신뢰할 수 있는 평가 시스템의 필요성이 제기되고 있다. 설명가능한 AI(XAI) 기술은 이러한 객관성과 신뢰성을 확보할 수 있는 방법으로서 중요성을 함유하고 있다.

본 논문은 XAI 기법에 초점을 맞춰 AI를 기반으로 한 마비말장애 평가의 연구 동향을 분석한다. 연구는 크게 기존 청지각적 평가 방법을 보조하는 음성 인식률 기반 연구 방법과 구조화된 데이터를 사용하는 특징 기반 방법, 그리고 음성 데이터를 end-to-end 방식으로 사용하는 딥러닝 모델 기반 연구 방법으로 논의된다. 이러한 방법에 XAI 기법을 통합함으로써 의료 전문가와 환자들을 위한 결과 해석과 신뢰성의 향상으로 마비말장애의 효과적인 진단과 치료를 용이하게 할 수 있다.

1. 서 론

마비말장애(dysarthria)는 뇌 및 말초신경 이상으로 인해 호흡・발성・공명・조음・운율 등의 각 과정에서 말 조절 근육의 마비나 약화, 불협이 나타나는 운동언어장애이다. 마비말장애 발생의 일반적인 원인은 파킨슨병, 알츠하이버병과 같은 신경계 퇴행성 질환과 안면 마비, 혀 또는 목 근육 약화를 일으키는 뇌졸중, 뇌손상, 다발성 경화증 등이 있다. 마비말장애는 말 명료도(speech intelligibility)의 저하, 반복적인 소통의 실패로 인해 궁극적으로 삶의 질 하락을 야기한다.

현재 임상에서는 마비말장애 진단을 위해 훈련된 의료진 혹은 언어재활사에 의한 Mayo Clinic Rating System이나 Frenchay Dysarthria Assessment, 혹은 말 명료도 평가와 같은 청지각적 평가 방법이 사용되고 있다[1]. 그러나 이러한 방법은 많은 노동력과 시간을 필요로 할 뿐만 아니라 평가하는 사람의 경험과 주관에 의존한다는 한계가 있다. 따라서 환자의 상태 추적과 치료를 효과적으로 수행하기 위한 객관적이고 정확한, 신뢰할 수 있는 마비말장애 판단 시스템이 필요하다. 이러한 시스템은 신경학적 장애의 조기 징후를 탐지할 뿐 아니라 환자 상태에 대한 장기적인 추적 또는 약리학적 치료나 재활의 이점을 평가하는 데에도 사용될 수 있다.

최근 머신러닝, 딥러닝 기술이 발전함에 따라 마비말장애 연구에도 마비말장애 탐지부터 심각도 평가에 이르기까지 인공지능 기술이 적용된 다양한 방법론이 적용되고 있다. 대부분의 연구들은 모델의 성능 향상 혹은 어떤 입력 특징이 모델의 결정에 영향을 미치는지에 대한 파악에 중점을 두고 있다. 하지만 생명과 직결되는 의료 도메인의 특성상 모델이 왜 그런 결과를 내었는지에 대한 설명가능한 AI, 즉 XAI(Explainable AI) 기술의 사용이 매우 중요하다. 모델이 제공하는 정보나 추천 사항이 어떤 이유로 결정되었는지를 이해하고 설명할 수 있을 때, 의료진 혹은 언어재활사가 AI의 추천을 신뢰하고 환자에게 적용할 수 있다. 또한 환자 역시 자신의 치료에 대한 결정을 내리기 위해 필요한 정보를 이해할 필요가 있다.

본 논문에서는 음성 기반 마비말장애 판단에 대한 인공지능 기술, 특히 설명가능한 인공지능 기술에 초점을 맞추어 연구 동향을 분석하고자 한다.

2. 마비말장애 평가 모델과 XAI

먼저 자동음성인식기를 활용해 전문가의 청지각적평가

* 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00621, 대화 기반 설명가능성을 멀티모달로 제공하는 인공지능 기술 개발)

를 보조하는 연구가 있었고 이를 넘어 마비말장애 평가 자체를 자동으로 진행하려는 시도가 있었다. 자동 평가

인공지능 기술에는 직접 추출한 음성 특징을 정형 데이터로 사용해 모델을 학습하는 방법과 음성을 바로 신경망 모델의 입력으로 학습하는 두 가지 주요 접근법이 있다.

첫 번째 접근법은 마비말장애 음성의 특성을 잘 나타내는 주요 특징들을 식별하여 추출하고, 이를 기반으로 학습하는 방식이다. 정형 데이터 음성 특징 기반 접근법은 의료진이 이해하기 쉽다는 장점이 있지만, 특징을 어떻게 추출하는지에 따라 성능 편차가 심하고 장애 판단에 필수적일 수 있는 중요한 특징을 놓칠 수 있는 위험성이 존재한다. 두 번째 접근법은 신경망 모델에 음성 입력을 주고 end-to-end로 학습하는 방식이다. 이 방식은 일반적으로 충분히 학습되었을 때 첫 번째 방식과 비교하여 더 나은 성능을 보인다. 그러나 Network의 복잡한 구조로 인하여 학습된 표현과 모델의 결정 이유에 대한 해석을 어렵게 만들어 실제 의료 환경에서의 사용 가능성이 제한적이다. 최근에는 딥러닝 기술의 부상과 함께 두 접근법의 장점을 병합함으로써 해석 가능한 결과를 생성하는 딥러닝 기술에 대한 연구가 이루어지고 있다.

2.1 청지각적 평가를 보조하는 음성 인식률 기반 연구

대표 연구로서 마비말장애 말뭉치를 음성 인식 시스템에 연결하여 추출된 전사 문장을 활용하는 연구가 제안되었다[2, 3].

[2]에서는 구강암 환자 음성 데이터를 n-gram 언어모델을 포함한 자동 음성 인식 시스템으로 전사 문장을 추출한다. 정답 문장과 비교하여 단어 인식률을 계산하고, 이를 전문가가 평가한 청지각적 평가 결과와 상관관계를 분석한 결과 자동 음성 인식 시스템의 결과를 말 명료도 평가의 객관적 지표로 사용할만한 유의미한 성능을 보였고, 특히 Uni-gram 언어 모델을 사용했을 때가 가장 높은 상관관계를 보였다.

다른 연구에서는 뇌성마비 환자의 음성으로 이루어진 UA-Speech 데이터셋의 자동 음성 인식 결과로 추정한 말 명료도와 청지각적 평가로 측정된 말 명료도 간에 높은 상관관계를 보이는 단어들을 찾아낸다. 이 단어들을 청지각적 평가에 사용하여 많은 단어와 문장을 들어야하는 기존의 과정보다 빠르고 정확하게 평가할 수 있도록 했다[3].

2.2 음성 특징을 통한 정형 데이터 기반 연구

앞선 연구들은 전문가의 청지각적 평가를 돕는 것에 그쳤지만 최근에는 마비말장애 평가 자체를 자동 시스템으로 진행하는 시도가 등장했고, 좋은 성능을 보이고 있다[4, 5, 6].

[4]에서는 음높이 윤곽(pitch contour)으로 대표되는 운율 특성과 주파수변이(jitter), 진폭변이(shimmer), 배음 대 잡음비(harmonics-to-noise ratio) 등의 음성 품질 특성, 그리고 모음 발음시간(vowel duration), 정지시간(pause duration), Mel-Frequency Cepstral Coefficient(MFCC) 등의 발음 특성이 일반적인 음성과 마비말장애 음성을 구분하는 것에 효과가 있다는 것을 증명했다. [5]에서는 마비말장애 심각도 분류에 위상 정보를 사용하여 성능을 높일 수 있다는 것을 증명했다.

나아가 앞선 특징들에 대해 다언어 마비말장애 데이터를 분석하여 각 언어별 주요 특징을 찾아내는 연구가 진행되었다[6]. 특히 이 연구에서는 모든 특징을 분류기에 사용하는 것보다 주요 특징으로만 학습했을 때 더 좋은 분류 성능을 보였고, Extreme Gradient Boosting(XGBoost)과 같은 트리 기반 머신러닝 알고리즘을 사용하여 변수 중요도(feature importance)를 통해 인공지능 모델의 설명가능성을 제시했다.

2.3 딥러닝 기반 end-to-end 연구

컨볼루션 신경망을 포함한 딥러닝 모델이 마비말장애 분류에 높은 성능을 보인다는 연구는 존재하지만 모델의 높은 복잡도 때문에 분류 과정을 해석하기는 어려운 단점이 있다[7]. 이러한 단점을 극복하고 고차원의 입력 음향 특징 공간으로부터 직접적으로 심각도 분류를 수행하는 대신, Bottleneck 특징 추출기로서 동작하는 중간 계층을 더하여 마비말장애 음성의 심각도와 이를 설명하는 특징들을 출력하는 모델이 제안되었다[8, 9]. 모델이 특징들과 분류 레이블을 결합하여 학습하도록 멀티태스크 훈련 방식을 사용하였고 중간 계층은 임상의들이 진단에 사용하는 청지각적 특징인 비정상적인 비음(nasality), 음성 품질(vocal quality), 조음 정확도(articulatory precision), 그리고 운율(prosody)을 정답 레이블로 학습하여 모델이 음성-언어 병리학자들의 주관적인 평가와 높은 연관성을 보이는 설명가능한 결과를 출력하도록 한다[8]. 이를 확장하여, 앞선 연구와 달리 주관적 레이블을 배제하고 음성에서 계산될 수 있는 주요 음성 특징(principal acoustic feature)에 집중한 연구가 있었다[9].

중간층은 모음-자음 전이 정확도(consonant-vowel transition precision), 과비성성(hypernasality), 조음정확도, 그리고 음성 품질의 네 가지 특징들을 학습하도록 훈련

표 1. 마비말장애 판단 AI 모델 요약

S. No.	References	Task	Databases used	Learning approach used	Contribution	Explainability
1	A. Maier et al. (2009) [2]	Dysarthria feature selection	Oral cancer data	ASR	n-gram 언어 모델	Correlation between the class and feature
2	A. Tripathi, S. Bhosale, and S. K. Kopparapu (2020) [3]	Dysarthria feature selection	UA-Speech database [14]	ASR	ASR을 통한 말 명료도 판단에 효율적인 단어 분석	Correlation between the class and feature
3	J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan (2015) [4]	Binary intelligibility classification	NKI CCRT Speech Corpus and TORGO database [15]	KNN, SVM, and LDA	마비말장애 판단 음성 특징 분석 - 운율, 음성 품질, 발음	.
4	K. Gurugubelli and A. K. Vuppala (2020) [5]	Dysarthric speech detection and intelligibility assessment	UA-Speech database	i-vector/PLDA	마비말장애 판단 음성 특징 분석 - 위상	.
5	E. J. Yeo, K. Choi, S. Kim, and M. Chung (2022) [6]	Dysarthria severity classification	TORGO, QoLT, and SSNCE database [16,17]	XGBoost	다국가 언어별 마비말장애 판단에 유리한 특징 비교분석	Feature importance of the model
6	E. J. Yeo, K. Choi, S. Kim, and M. Chung (2022) [7]	Dysarthria severity classification	QoLT database	wav2vec, multi-task learning	사전학습된 대용량 딥러닝 모델을 사용, 마비말장애 심각도 분류 성능 향상	.
7	Tu, M., Berisha, V., and Liss (2017) [8]	Dysarthria severity classification	Database from Motor Speech Disorders Lab at ASU	DNN	DNN 중간의 Bottleneck 특징 추출기로 딥러닝 모델에서의 설명가능성 제안	Output of bottleneck layer (DAB layer)
8	Xu, Lingfeng, Julie Liss, and Visar Berisha (2023) [9]	Dysarthria severity classification	Database from Motor Speech Disorders Lab at ASU	CNN	SHAP value 시각화를 통한 설명가능성 제안	SHAP value
9	Pan, Yilin, et al. (2020) [11]	Speech disorder classfication	IVA database	SincNet, CNN	SincNet을 통한 raw waveform에서의 음성 특징 추출	.
10	Hung, Chao-Hsiang, et al. (2022) [12]	Speech disorder classfication	FEMH Speech Disorders database	SincNet, CNN	Sinc함수를 학습가능한 변수로 설정해 효율적인 필터 확보	Learned cutoff frequency of SincNet

되며, Bottleneck 특징 추출기로서 솔루션 공간을 제한하고 모델의 정확도를 향상시키는 역할을 한다. 제안된 방법은 SHapley Additive exPlanations (SHAP)을 도입하여 최종 예측 결과에 네 가지 특징이 각각 얼마나 기여했는지 분석할 수 있도록 했다[10].

스펙트로그램을 모델의 입력으로 사용하는 앞선 연구와 달리 첫 계층에 sinc 함수로 만들어진 컨볼루션 필터를 도입해 파라미터의 수를 크게 줄이면서도 오디오 신호의 특징을 잘 추출하는 SincNet을 활용한 연구들이 있다[11, 12]. SincNet의 저자는 신경망에서 원시 파형으로부터 음성을 직접 처리하는 첫 번째 계층의 추출 능력이 중요함을 주장하였다[13]. 이러한 SincNet을 앞단에 배치한 특징 추출기를 이용하여 건강한 화자와 신경퇴행성 질환 및 인지 장애를 가진 화자를 구별하는데 필요한 특징을 추출하는 방법이 제안되었다[11]. 저자는 SincNet 계층의 누적 주파수 응답을 분석함으로써, 건강한 화자와 장애를 가진 화자를 구별함에 있어 저주파 정보를 분석하는 것이 중요하다는 근거를 보였다. 다른 연구는 Sinc 필터의 컷오프 주파수를 학습 가능한 인자로 하여, 병리적인 음성을 구별하는 데 필요한 음향 특징을 효과적으로 추출하여 명확한 물리적 의미를 갖는 필터 뱅크 대역폭 도출 방법을 제안하였다[12]. CNN 필터와 Sinc 필터의 전력 스펙트럼 밀도를 비교하여 Sinc 필터로 처리된 음성 신호가 첫 번째 공명주파수(F0)를 더 잘 보존함을 보였다.

3. 결론 및 향후 연구

본 논문에서는 국내외 마비말장애 탐지 및 심각도 분류에 대한 연구 동향을 살펴보고 방법에 대한 전반적인 내용을 설명하였다. 사람이 진행하는 청지각적 평가에 도움을 주는 연구부터 머신러닝 알고리즘을 사용한 연구와 최근에는 딥러닝 모델에 마비말장애 데이터를 사용하여 분류 성능을 끌어올린 연구까지 살펴보았다. 음성 특징을 정형데이터로 사용한 머신러닝 알고리즘의 경우 변수중요도를 통해 환자들에게 분류 결과를 설명할 수 있고, Bottleneck 특징 추출기를 통해 딥러닝 모델에서도 설명가능성이 있다는 것을 제시한 연구도 찾아볼 수 있었다.

인공지능을 통한 마비말장애 자동 평가 기술은 객관적이고 정확하며 신뢰할 수 있는 평가를 제공하는 데에 큰 잠재력을 가지고 있다. 더욱이 설명가능한 AI 기법을 통합함으로써 의료 전문가와 환자가 의사 결정 과정을 이해하고 결과를 신뢰할 수 있게 되면 늘어나는 비대면 진료 수요에 맞춰 의료 산업에도 큰 영향력을 끼칠 것이다. 하지만 의료 분야에서 상용화되려면 매우 높은 수준의 객관성을 증명해야하나 현재의 연구 결과는 인간의 청지각적 평가를 대체할 만큼의 성능을 보이지는 못한다는 한계가 있다. 많은 연구들에서 공통적으로 토로하는 대용량 데이터의 부재가 해결된다면 발전하는 AI 기술을 통해 이 한계를 극복할 수 있을 것이라 생각한다.

참 고 문 헌

[1] P. Enderby, “Frenchay dysarthria assessment,“ British Journal of Disorders of Communication, vol. 15, no. 3, pp. 165-173, 1980.

[2] A. Maier et al., “Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer,“ EURASIP Journal on Audio, Speech, and Music Processing, vol. 2010, pp. 1-7, 2009.

[3] A. Tripathi, S. Bhosale, and S. K. Kopparapu, “A novel approach for intelligibility assessment in dysarthric subjects,“ in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6779-6783, 2020.

[4] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, “Automatic intelligibility classification of sentence-level pathological speech,“ Computer speech & language, vol. 29, no. 1, pp. 132-144, 2015.

[5] K. Gurugubelli and A. K. Vuppala, “Analytic phase features for dysarthric speech detection and intelligibility assessment,“ Speech Communication, vol. 121, pp. 1-15, 2020.

[6] E. J. Yeo, K. Choi, S. Kim, and M. Chung, “Cross-lingual Dysarthria Severity Classification for English, Korean, and Tamil,“ in 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 566-574, 2022.

[7] E. J. Yeo, K. Choi, S. Kim, and M. Chung, “Automatic Severity Assessment of Dysarthric speech by using Self-supervised Model with Multi-task Learning,“ arXiv preprint arXiv:2210.15387, 2022.

[8] Tu, M., Berisha, V., and Liss, J. “Interpretable objective assessment of dysarthric speech based on deep neural networks,” in Proceedings of Interspeech, August 20-24, Stockholm, Sweden, pp. 1849-1853, 2017.

[9] Xu, Lingfeng, Julie Liss, and Visar Berisha. “Dysarthria detection based on a deep learning model with a clinically-interpretable layer.“ JASA Express Letters 3.1, 2023.

[10] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,“ Advances in neural information processing systems, vol. 30, 2017.

[11] Pan, Yilin, et al. “Acoustic feature extraction with interpretable deep neural network for neurodegenerative related disorder classification.“ Proceedings of Interspeech. International Speech Communication Association (ISCA), 2020.

[12] Hung, Chao-Hsiang, et al. “Using SincNet for learning pathological voice disorders.“ Sensors 22.17, 2022.

[13] Ravanelli, Mirco, and Yoshua Bengio. “Speaker recognition from raw waveform with sincnet.“ IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018.

[14] H. Kim et al., “Dysarthric speech database for universal access research,“ in Ninth Annual Conference of the International Speech Communication Association, 2008.

[15] F. Rudzicz, A. K. Namasivayam, and T. Wolff, “The TORGO database of acoustic and articulatory speech from speakers with dysarthria,“ Language Resources and Evaluation, vol. 46, pp. 523-541, 2012.

[16] D.-L. Choi, B.-W. Kim, Y.-W. Kim, Y.-J. Lee, Y. Um, and M. Chung, “Dysarthric Speech Database for Development of QoLT Software Technology,“ in LREC, 2012, pp. 3378-3381.

[17] M. C. TA, T. Nagarajan, and P. Vijayalakshmi, “Dysarthric speech corpus in tamil for rehabilitation research,“ in 2016 IEEE Region 10 Conference (TENCON), 2016: IEEE, pp. 2610-2613.