Random Forest와 DSP 전처리 기법을 적용한 웹 기반 실시간 지문자 및 음성 인식 시스템 구현

이유준 · 강진성 · 최규상*

영남대학교

qpworms@naver.com, k75114765@gmail.com, castchoi@yu.ac.kr

Implementation of Web-based Real-time Fingerspelling and Speech Recognition System using Random Forest and DSP Pre-processing Techniques

Lee Yoo Joon · Jin Seong Kang · Kyu Sang Choi*

요 약

본 논문에서는 청각 장애인과 비장애인의 원활한 의사소통을 지원하기 위해 웹 브라우저 환경에서 동작하는 양방향 번역 시스템을 제안한다. 제안하는 시스템은 별도의 프로그램 설치 없이 웹캠과 마이크를 통해 입력된 신호를 실시간으로 처리한다. 지문자 인식부는 MediaPipe를 통해 손의 관절 좌표를 추출한 후, 벡터 내적 기반의 각도 특징(Angle Feature)을 산출하여 Random Forest 모델로 분류한다. 음성 인식부는 자체 구현된 디지털 신호 처리(DSP) 모듈을 통해 Pre-emphasis 및 VAD(Voice Activity Detection) 전처리를 수행한 후 Whisper 모델을 통해 텍스트로 변환한다. 실험 결과, 제안하는 지문자 인식 모델은 92%의 정확도와 1.2ms의 추론 속도를 보였으며, 웹 환경에서의 실시간 인식 가능성을 입증하였다.

Ⅰ. 서 론

청각 장애인은 수어를 사용하지 않는 비장애인과의 소통 과정에서 필연적인 어려움을 겪는다. 이를 해결하기 위해 딥러닝 모델들이 도입되었으나, 과도한 GPU 사용 및 실시간 소통에 딜레이가 발생해 실시간 추론에는 제약이 발생한다. 반면 단순 거리 기반 알고리즘은 연산에 의한 딜레이는 발생하지 않으나 외부 노이즈 및 기본적인 성능이 부족하여 실제 환경에서의 정확성을 보장하기에는 큰 어려움이 있다.

그러나 기존 연구들은 청각장애인이 비장애인에게 의사를 전달하는 단방향 소통에만 초점을 맞추고 있다. 실제 대화 상황에서는 비장애인의 발화를 청각장애인이 이해할 수 있도록 실시간 음성 인식이 함께 필요하며, 긴급 상황이나 반복적인 의사 표현에서는 미리 등록된 문구를 즉시 호출할 수 있는 보조 수단이 유용하다.

이에 본 연구는 웹 표준 기술인 WebSocket과 Random Forest 알고리즘을 결합하여 속도와 정확도 간의 상충 관계(Trade-off)를 해결하고, 지문자 인식, 실시간 음성 인식, 프리셋 문구 데이터베이스를 통합한 양방향 의사소통 시스템을 설계하였다. Random Forest는 다수의 결정 트리를 앙상블하는 방식을 통해 딥러닝 모델 대비 연산 속도를 획기적으로 향상시키면서도 단순 알고리즘 기법보다는 우수한 일반화 성능을 보장한다. 결론적으로 본 논문은 이러한 경량화된 모델과 신호처리 기술을 결합해, 고성능 장비가 부재한 저사양 CPU에서도 딜레이 없는 양방향 통신 환경을 구축하는 데 그 목적이 있다. 본 연구에서는 수어 중 자모 단위로 표현하는 지문자(Fingerspelling)를 대상으로 한다.

Ⅱ. 본론

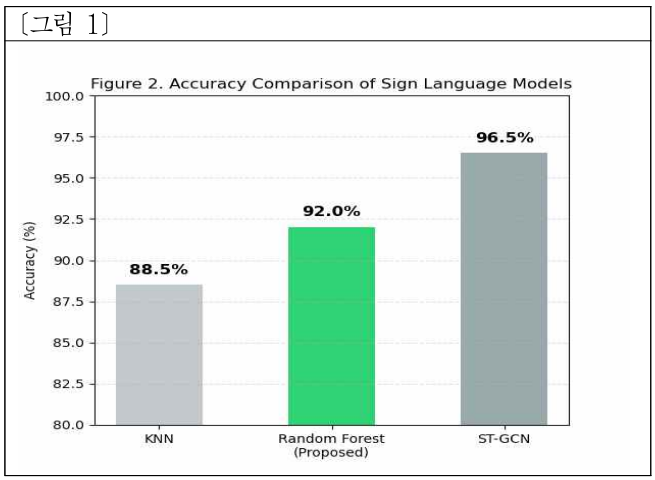
2.1 시스템 구조 및 설계

본 연구는 사용자가 별다른 제약 없이 쉽게 접근할 수 있도록 접근성 극대화를 최우선 가치로 삼아, 클라이언트 - 서버 구조를 기반으로 시스템을 설계하였다. 백엔드는 Python FastAPI를 중심으로 구성해 서버 측 연산 지연을 최소화했으며, 프론트엔드와의 데이터 흐름은 WebSocket을 사용해 16kHz PCM 오디오와 비디오 스트림을 실시간 스트리밍 방식으로 지연 없이 주고받을 수 있도록 구현했다. 또한 MySQL 데이터베이스와 연동되는 하이브리드 인터페이스를 도입한 점이 특징이다. 이는 지문자나 발화가 어려운 긴급 상황을 고려한 보완 장치로, DB에 미리 등록된 프리셋 문구를 '원클릭' 트리거로 즉시 호출해 텍스트와 음성 형태로 곧바로 송출함으로써, 사용자 경험(UX)에서 발생할 수 있는 공백을 줄였다.

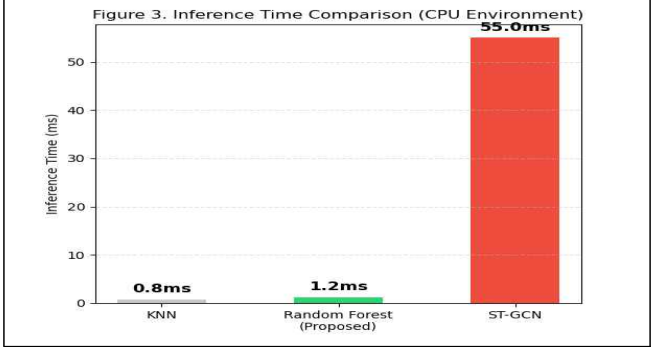
2.2 지문자 인식 모델 비교 및 선정

웹 브라우저 환경은 CPU 자원이 제한적이기 때문에, 본 연구는 실시간성과 정확도를 동시에 만족시키기 위해 서로 다른 방식의 세 모델을 정량적으로 비교했다. 비교 대상은 ST-GCN(딥러닝 기반), KNN(기초 머신러닝), Random Forest(앙상블)이며, 총 3,150개 샘플(한글 자모 31종: 자음 14개, 모음 17개)을 8:2 비율로 학습/테스트 세트를 분리하여 실험을 진행했다. 실험 결과 [그림 1]과 같이 ST-GCN은 시계열 특성을 잘 반영해 96.5%의 높은 정확도를 보였지만, 그래프 기반 연산의 복잡도로 인해 프레임당 추론 시간이 55ms 이상으로 증가하여 웹 서비스에서 요구되는 실시간성을 떨어뜨리는 병목이 확인되었다. 반면 KNN은 0.8ms 수준의 매우 빠른 추론이 가능했으나, 노이즈에 민감해 정확도가 88.5%로 상대적으로 낮았다. 종합적으로, 본 연구는 두 모델 사이의 균형점을 제공하는 Random Forest를 최종 모델로 선정하였다. Random Forest는 앙상블 구조를 통해 92.0%의 안정적인 분류 정확도를 확보하면서도, 1.2ms의 경량 추론 속도를 달성해 브라우저 기반 실시간 서비스 요구사항에 가장 적합한 선택임을 확인했다.

〔그림 1〕



〔그림 2〕



2.3 벡터 특징 추출 및 한글 오토마타

지문자 인식 파이프라인은 MediaPipe Hands를 이용해 손의 21개 관절(landmark) 좌표를 추출하는 단계에서 시작된다. 다만 이때 얻어지는 원시 좌표값(x, y, z)을 그대로 모델 입력으로 사용하면, 촬영 거리나 손 크기 차이처럼 피사체의 심도(Depth)와 스케일 변화에 따라 값이 쉽게 흔들리면서 모델의 일반화 성능이 떨어질 수 있다. 이에 본 연구는 이러한 스케일 변동성(Scale Variance) 문제를 줄이기 위해, 단순 좌표 대신 관절 간 벡터 관계에 주목했다. 구체적으로는 식 (1)과 같이 인접 관절로부터 벡터를 구성한 뒤, 벡터 간 내적(dot product)을 이용해 각도(Angle) 정보를 계산하여 15차원 특징 벡터를 구성하였다. 각도 기반 특징은 거리나 크기 자체에 덜 민감하기 때문에, 사용자나 촬영 환경이 달라져도 비교적 일관된 패턴을 제공하며 실시간 웹 환경에서도 안정적인 인식 성능을 기대할 수 있다. 인식 안정성을 위해 15프레임 동안의 예측 결과를 다수결 투표(60% 이상)로 확정하였으며, 같은 글자의 연속 입력 방지를 위해 1.2초의 쿨다운을 적용하였다.

$$\theta = \arccos\left(\frac{\vec{v_1} \cdot \vec{v_2}}{\|\vec{v_1}\| \, \|\vec{v_2}\|}\right) \quad -(1)$$

정제된 특징 벡터는 Random Forest 분류기를 통해 먼저 지문자 단위로 분류된다. 다만 이 결과는 어디까지나 자소가 순서대로 나열된 형태라서, 그대로는 사람이 읽기 편한 완성형 글자가 되기 어렵다. 그래서 본 시스템은 후처리 단계에 자체 설계한 한글 오토마타(Hangul Automaton) 엔진을 탑재했다. 이 엔진은 4개의 상태(EMPTY, CHO, CHO_JUNG, CHO_JUNG_JONG)를 가지는 유한 상태 기계로 설계되었으며, 실시간으로 들어오는 초성·중성·종성을 한글 결합 규칙에 따라 즉시 조합해(예: ㅇ + ㅏ + ㄴ → 안) 사용자가 바로 이해할 수 있는 완성형 텍스트로 변환해 출력한다. 결과적으로 인식 결과가 단순 "분류 출력"에서 끝나는 게 아니라, 실제 대화 상황에서 바로 쓸 수 있는 형태로 정리된다.

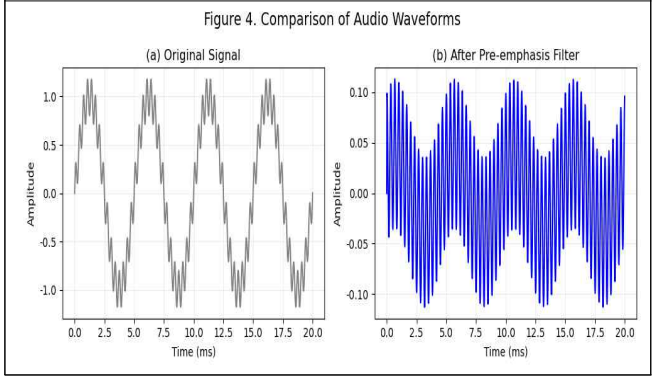
2.4 DSP 기반 음성 신호 전처리

본 시스템은 음성 인식 모델(Whisper)의 성능을 최대한 끌어올리기 위해, 원시 오디오가 모델에 입력되기 전에 자체 설계한 DSP(Digital Signal Processing) 전처리 파이프라인을 거치도록 구성했다. 첫 단계로는 자연 발화 과정에서 흔히 나타나는 고주파 대역 에너지 감소를 보정하기 위해, 식 (2)에 기반한 프리 엠퍼시스(Pre-emphasis) 필터를 적용했다. 이 과정은 신호의 스펙트럼 평탄도(spectral flatness)를 개선해 자음·모음 등 음소 간 구분을 더 또렷하게 만들고, 결과적으로 모델이 발화를 더 안정적으로 인식하도록 돕는다.

$$y[n] = x[n] - 0.97x[n-1] \quad -(2)$$

[그림 3]은 해당 필터의 적용 전후 파형을 시각화한 것으로, 적용 후 고주파 성분의 진폭이 강조되어 소음이 있는 환경에서도 강건성을 확보하였다. 이어지는 단계에서는 불필요한 연산을 줄이기 위해 25ms 프레임 단위로 단구간 에너지와 영교차율(ZCR)을 지표로 삼는 VAD 알고리즘을 가동하여 비음성 구간을 사전에 필터링한다. 단구간 에너지는 프레임 내 신호의 세기를 측정하고, ZCR은 신호가 0을 교차하는 빈도를 계산하여 유성음과 무성음을 구분하는 데 활용된다. 마지막으로 웹 환경에서 배경 잡음을 줄이기 위해 주파수 도메인 상에서 스펙트럴 차감법(Spectral Subtraction)과 위너 필터(Wiener Filter)를 적용함으로써, 신호 대 잡음비(SNR)를 개선하였다. 스펙트럴 차감법은 초기 프레임에서 추정한 잡음 스펙트럼을 전체 신호에서 차감하는 방식이며, 위너 필터는 신호와 잡음의 비율을 기반으로 최적의 필터 계수를 산출하여 잡음을 억제한다. 또한 Whisper 모델의 환각(hallucination) 출력을 방지하기 위해 뉴스 자막, 반복 단어 등을 필터링하는 후처리를 적용하였다.

〔그림 3〕



Ⅲ. 결론

본 연구는 GPU나 전용 소프트웨어의 도움 없이 웹 브라우저만으로 동작하는 양방향 통신을 구축하는 데 노력하였다. 특히 딥러닝(ST-GCN)의 고연산과 기초 머신러닝(KNN)의 낮은 정확도 사이에서 발생하는 문제점을 해결하기 위해, 본 연구에서는 Random Forest를 최적의 대안으로 채택하였다. 이를 통해 일반 CPU 환경에서도 1.2ms의 초저지연 추론과 92%의 정확도를 동시에 달성하는 성과를 거두었다. 음성처리에서는 독자 설계된 DSP 전처리 파이프라인을 활용해 웹 마이크 음질의 한계를 극복하고, 한글 오토마타를 적용하여 파편화된 지문자 인식을 넘어 완성형 한글의 실시간 변환을 구현하였다. 그러나 기하학적 각도 특징에 의존하는 현 방식은 손의 회전 정보 소실로 인해 'ㄱ'과 'ㄴ' 등 유사 형상을 구분하는 데 한계를 드러냈다. 향후 연구에서 손의 절대 방향성을 특징 벡터에 포함하고 데이터 증강 기법을 활용하여 이러한 한계를 극복할 것이며, 나아가 단순 어휘 인식을 넘어 문맥을 이해하는 연속 수어 번역(SLT) 단계로 시스템을 확장할 계획이다.

참 고 문 헌

[1] A. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. G. Yong, J. Lee, W. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A Framework for Building Perception Pipelines," arXiv preprint arXiv:1906.08172, 2019.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," arXiv preprint arXiv:2212.04356, 2022.

[3] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[4] S. I. Park, C. H. Lee, and J. H. Park, "A Study on the Korean Sign Language Recognition using Deep Learning and Hand Skeleton Features," Journal of the Korea Institute of Information and Communication Engineering, vol. 25, no. 10, pp. 1324-1331, 2021.

[5] J. G. Proakis and D. G. Manolakis, Digital Signal Processing: Principles, Algorithms, and Applications, 4th ed., Pearson, 2006.

[6] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, pp. 7444-7452, 2018.