

마비말장애 중증도 자동 분류를 위한 퓨샷 러닝

여은정, 정민화*
서울대학교

ej.yeo@snu.ac.kr, *mchung@snu.ac.kr

Few-shot learning for automatic severity classification of dysarthric speech

Yeo Eun Jung, Chung Minhwa*
Seoul National Univ.

요 약

마비말장애 중증도 자동 분류 기술은 객관적이고 일관적인 결과를 산출함으로써 언어재활사의 평가를 보조할 수 있다. 그러나 마비말장애 음성은 대량의 균형 코퍼스 수집이 어렵기 때문에 딥러닝 모델 사용에 제한이 있다. 따라서 본 논문에서는 다양한 소규모 이미지 데이터 분류 과제에서 좋은 성능을 보이고 있는 퓨샷 러닝 방법론을 소규모 음성 데이터 분류 과제인 마비말장애 중증도 자동 분류에 적용할 것을 제안한다. 우선, 본 논문은 음성을 시각화한 스펙트로그램을 사용하여 장애 중증도 분류를 시도한다. 더 나아가, 마비말장애 중증도 분류에 도움이 되는 것으로 알려진 음질 특징과 운율 특징을 함께 사용할 것을 제안한다. 실험 결과, 퓨샷 러닝을 사용한 모델이 베이스 모델인 Resnet 또는 파인 튜닝한 모델보다 더 높은 분류 정확도를 보였다. 또한 스펙트로그램을 단독으로 사용했을 때보다 스펙트로그램과 음질 특징, 운율 특징을 함께 사용했을 때 더 높은 성능이 나타났다. 본 연구는 마비말장애 중증도 자동 분류를 위한 퓨샷 러닝의 초기 연구로서, 장애 음성 분류에 대한 퓨샷 러닝 방법론의 적용 가능성과 방향성을 탐구한 것에 그 의의가 있다.

I. 서 론

마비말장애는 신경계 손상으로 말소리 산출과 관련된 근육이 마비되거나 약해지면서 발생하는 말운동장애이다. 마비말장애는 선천적 뇌손상, 후천적 뇌손상, 노화에 따른 퇴행성 신경질환 등 넓은 범위에서 나타나며, 호흡, 발성, 공명, 조음, 운율 등 다양한 말 기능에 영향을 미친다. 언어재활사는 마비말장애 중증도 평가를 통해 화자가 의사소통 상황에서 겪는 어려움의 정도를 확인하고, 화자의 상황에 알맞은 중재 방안을 모색한다. 그러나 임상 현장에서 주로 사용되는 말 명료도 평가는 청지각적 평가로, 주관적일 뿐만 아니라 많은 시간과 노력이 소요된다. 따라서 숙련된 언어재활사의 평가 결과와 유사한 결과를 산출하는 마비말장애 중증도 자동 분류 기술은 객관적이고 일관적인 결과를 바탕으로 언어재활사의 평가에 유용하게 사용될 수 있다.

최근 딥러닝 기술이 발전함에 따라 마비말장애 중증도 자동 분류 연구에서도 딥러닝 방법론이 시도되고 있다. 그러나 장애 발화는 대규모 코퍼스 수집이 어렵기 때문에 딥러닝 기술 적용이 쉽지 않은 실정이다. 소규모 데이터를 딥러닝 모델 학습에 사용할 경우, 과적합(overfitting)이 발생하기 쉽기 때문이다. 따라서 딥러닝 모델을 사용한 대부분의 선행연구들은 대개 비장애 음성과 장애 음성만을 구분하는 장애 음성 검출 연구에 그치고 있다.[1][2]

소규모 데이터로 인한 과적합 문제에 대한 해결 방안으로는 우선 사전 학습(pre-trained)된 모델의 일부를 훈련하는 파인 튜닝(fine-tuning)이 있다. 훈련하는 모델 층의 개수는 훈련 데이터의 양과 종류에 따라 다르게 설정할 수 있는데, 장애 발화에 대한 파인 튜닝은 과적합 문제에서 자유롭지 못하다. 대부분의 장애

발화 실험의 경우 학습 데이터가 사전 훈련에 사용된 데이터와 속성이 다르면서 분량 역시 매우 적기 때문이다.

두 번째 해결책으로는 퓨샷 러닝(few-shot learning)이 있는데, 모델에게 데이터를 구분하도록 학습시킴으로써 단 몇 개의 데이터만으로도 분류 과제를 시행할 수 있는 방법론이다. 모델의 파라미터가 아닌 데이터 간의 거리(유사도) 또는 관계를 학습하기 때문에 과적합의 우려가 다소 적다. 최근 들어 데이터 라벨링이 어려운 소규모 데이터 분류 과제를 위해 다양한 퓨샷 러닝 접근법이 제안되고 있지만 대개 이미지 분류에 치우쳐 보고되었다. 본 연구에서는 퓨샷 러닝 접근법을 소량의 음성 데이터 분류 과제인 마비말장애 중증도 분류에 적용하여 그 효과를 살펴보고자 한다.

II. 본 론

i) 모델 구조

그림 1은 본 연구에서 제안하는 모델의 구조도이다. 퓨샷 러닝을 위한 다양한 모델 중 프로토타입 네트워크(Prototypical Network)[3]를 사용하였는데, 서포트 데이터의 평균 위치에 있는 프로토타입을 사용해 장애 발화처럼 불균형한 데이터셋에도 좋은 성능을 보이는 것으로 보고되기 때문이다.

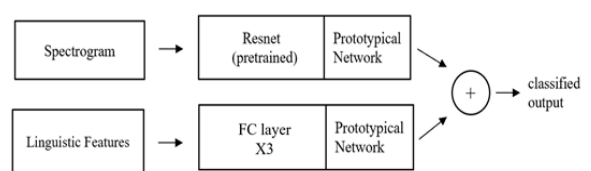


그림 1. 모델 구조도

모델은 크게 두 갈래로 구성된다. 첫 번째 갈래는 이미지를 사용한 분류로, 각 음성을 스펙트로그램으로 시각화해 입력 값으로 사용한다. 이미지 분류를 위해서는 베이스 모델(backbone model)로 ImageNet 데이터셋으로 사전 학습된 Resnet 모델을 사용하고, flatten 한 Resnet 의 마지막 층에 프로토타입 네트워크를 잇는다. 두 번째 갈래는 언어학적 특징을 이용한 분류로, 각 음성의 음질 특징, 운율 특징을 입력 값으로 사용한다. 베이스 모델로는 3 개의 완전연결층(fully-connected layer)을 쌓은 MLP 를 사용하고, 후단에 프로토타입 네트워크를 쌓는다.

각 갈래의 출력값은 프로토타입과 쿼리 데이터 간의 거리로, 두 출력값의 합이 마비말장애 중증도 분류에 사용된다. 모델은 음성을 5 분류하며, 장애 중증도는 비장애, 경도, 경도-중등도, 중등도-중도, 중도로 구분된다.

ii) 사용 특징

본 연구에서 사용된 특징은 스펙트로그램과 언어학적 특징으로 구분된다. 우선 스펙트로그램은 시간에 따른 주파수와 진폭에 대한 변화를 시각화한 그림으로, 음성의 발음 정보를 담고 있다. librosa 0.8.를 통해 멜 필터뱅크(mel filterbanks) 개수 400 개, 최고 주파수 8000Hz 로 추출하였다.

마비말장애 중증도 분류에는 발음 정보뿐만 아니라 음질 정도와 운율 정보도 함께 고려되어야 한다. 음질 특징은 음성 장애를 진단할 때 사용되는 jitter, shimmer, HNR (Harmonics to Noise Ratio)와 화자의 유성음 유지 능력을 확인할 수 있는 voice breaks 의 개수와 정도를 포함한다. 모든 음질 특징은 Praat 6.1.16 을 사용하여 추출하였다. 운율 특징은 음높이 특징과 발화 속도 특징으로 구분된다. 음높이 특징은 F0 의 평균값, 표준편차, 최솟값, 최댓값, 중앙값, 25 분위수, 75 분위수를, 발화 속도 특징은 전체 길이, 말 길이, 말 속도, 조음 속도를 포함한다. 특징들은 각각 Praat 6.1.16 과 Parselmouth 0.4.0 를 사용하여 추출하였다.

iii) 코퍼스

본 연구에서는 뇌성마비 마비말장애인 대상 음성인식기 개발을 위해 구축된 한국어 코퍼스인 QoLT 를 사용하였다. 코퍼스는 고립단어와 문장으로 구성되지만, 본 연구에서는 음질 특징, 운율 특징, 발음 특징 모두를 종합적으로 평가에 반영하기 위해 문장만을 사용하였다. 화자는 비장애인 화자 10 명(남 5 명, 여 5 명)과 마비말장애 화자 70 명(남 45 명, 여 25 명)으로 구성되며, 모든 화자는 5 개의 문장을 2 회 반복 발화하였다. 화자의 장애 중증도는 2 년 이상의 임상 경험이 있는 언어재활사 5 명이 모든 문장을 듣고 판단한 것으로, 각 장애 중증도 별 화자 수는 경도, 경도-중등도, 중등도-중도, 중도 순으로 각각 14 명, 14 명, 9 명, 4 명이다. 분류 실험에는 훈련 화자, 검증 화자, 테스트 화자를 중증도 집단 별 6:2:2 로 구성하였으며, 각 과정에서 화자는 겹치지 않았다.

III. 실험 및 실험 결과

본 모델은 5-way 5-shot 7-query 태스크로 훈련되었다. 훈련과 검증은 각 에피소드마다 서포트 데이터와 쿼리 데이터를 각 중증도 별로 랜덤하게 선택하였으며, 이 과정을 훈련과 검증 단계에서 각 5000 회를 반복하였다. 옵티마이저(optimizer)로는 Adam 을 사용하였으며, learning rate 는 0.01 이었다.

실험 결과, 퓨샷 러닝을 적용했을 때 분류 정확도 67.42%로 베이스라인인 사전 학습된 Resnet 의 33.52% 또는 과인 튜닝의 52.94%보다 더 높은 분류 정확도를 보였다. 더 나아가, 스펙트로그램만을 입력값으로 사용했을 때와 스펙트로그램과 음질, 운율 특징을 함께 사용했을 때의 분류 정확도를 비교했을 때, 각각 62.39%, 67.42%로 이미지와 음향학적 특징을 함께 사용했을 때 8.06%의 상대적 증가율을 보였다.

VI. 결론

본 연구는 마비말장애 중증도 자동 분류를 위해 스펙트로그램과 음향학적 특징(음질 특징, 운율 특징)을 입력값으로 사용하는 퓨샷 러닝 모델을 제안하였다. 본 연구의 최고 성능은 67.42%로, QoLT 코퍼스를 사용한 선행연구[4]의 SVM 분류기 최고 성능인 80.15% 또는 MLP 분류기의 최고 성능인 78.09%보다 낮은 성능이었다. 이는 본 연구에서 사용하지 않은 음소 단위의 발음 특징을 두 실험에서는 포함했기 때문이라고 추측된다. 향후 연구에서는 음소 단위의 발음 특징도 함께 특징으로 포함하여 그 효과를 살펴보아야 할 것이다. 반면, 본 연구와 동일한 범주인 프레임 단위의 특징(MFCCs 또는 스펙트로그램), 음질 특징, 운율 특징만을 사용한 연구 중에서는 본 논문이 가장 높은 성능을 보였으며, 이는 본 선행연구[5]의 MLP 성능보다 8.61% 상대적 증가한 값이다.

본 연구는 기존 소규모 이미지 데이터 분류 과제에서 좋은 성능을 보이는 퓨샷 러닝 방법론을 소규모 음성 데이터 분류 과제인 마비말장애 중증도 자동 분류 과제에 적용한 초기 연구로서 소규모 장애 음성 분류에 대한 퓨샷 러닝 방법론의 적용 가능성과 방향성을 탐구한 것에 그 의미가 있다.

ACKNOWLEDGMENT

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터육성지원사업의 연구결과로 수행되었음" (IITP-2021-2016-0-00464)

참 고 문 헌

- [1] Bhat, C., & Strik, H. (2020). Automatic assessment of sentence-level dysarthria intelligibility using BLSTM. *IEEE Journal of Selected Topics in Signal Processing*, 14(2), 322-330.
- [2] Millet, J., & Zeghidour, N. (2019, May). Learning to detect dysarthria from raw speech. " *IEEE International Conference on Acoustics, Speech and Signal Processing*, 5831-5835.
- [3] Snell, J., Swersky, K., & Zemel, R. (2017, December). Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4080-4090.
- [4] Yeo, E. J., Kim, S., & Chung, M. (2021). Automatic severity classification of dysarthria using voice quality, prosody, and pronunciation features. *Phonetics and Speech Sciences*, 13(2), 57-66.
- [5] Hernandez, A., Kim, S., & Chung, M. (2020). Prosody-based measures for automatic severity assessment of dysarthric speech. *Applied Sciences*, 10(19), 6999.