자동 말투(Speech Style) 인식: 다자간 대화 상황에서의 화자인식 기술 개발

강가람, Jin Guangxun*, 권오병**
경희대학교, *경희대학교, **경희대학교

1st9aram@khu.ac.kr, *jinguangxun0407@khu.ac.kr, **obkwon@khu.ac.kr

Automatic Speech Style Recognition: Development of Speaker Recognition Technology in Multilateral Conversation

Kang Ga Ram, Jin Guangxun*, Kwon Oh Byung**
Kyung Hee Univ., * Kyung Hee Univ., ** Kyung Hee Univ.

요 약

종결어미 중에는 존어체와 경어체가 있어, 화자의 높임법을 추측하게 하여 화자와 청자 사이의 우열을 파악하는데 유용하기도 하다. 또한 종결어미는 시간의 흐름에 따라 새로운 종결어미가 등장하기도 하고 사라지기도 하며, 그 의미가 변천하기도 하는 매우 역동적인 양상을 보인다. 이에 본 연구의 목적은 화자인식의 정확도를 개선하기 위해 화자가 발화한 문장에 등장하는 종결어미의 등장 특성을 바탕으로 화자인식 하는 방법을 제안하는 것이다. 이를 위해 '응답하라 1994'라는 K-Drama 자막 데이터를 학습데이터로 하여 종결어미에 대한 문장 시퀀싱에 의한 w-vector 를 기반으로 화자인식을 수행하였다. 그 결과 음성 정보에만 의존하여 i-vector, x-vector, 딥러닝 등의 방법을 혼합하여 화자인식 하는 방법을 보완하려고 했다. 본 연구는 종결어미에 기반한 문장 시퀀싱이라는 방법을 제안한 최초의 연구이며, 향후 실존하는 음성인식 시스템과 함께 활용되어 화자인식에 의한 지능형 대화 시스템과 이를 기반한 전자거래 및 각종 음성 기반 서비스에 활용될 것을 기대한다.

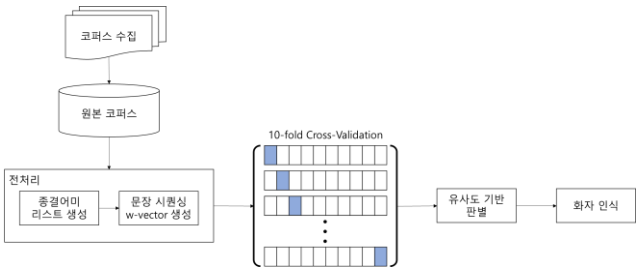
Ⅰ. 서 론

화자인식(Speaker Recognition)이란 특정인의 음성 샘플로부터 그가 누구인지를 자동으로 구별하는 기술이다[1]. 이를 위해 화자인식은 음성 샘플을 주요 입력 데이터로 필요로 한다[2]. 휴대용 기기의 발전과 음성 기술, 오디오 콘텐츠 분야 등이 계속해서 확장됨에 따라, 화자인식 기술의 중요성은 더욱 부각되고 있으며, 일반적인 거래나, 법의학 분야에서도 중요한 수단으로 사용되고 있다[3]. 그동안 음성 파일을 기반으로 하여 그 음성의 화자가 누구인지를 자동으로 판정하려는 화자인식 연구는 화자 판정의 정확도를 올리는 목표를 가지고 진행되어왔다. 음원에서 개선된 특징을 추출하는 접근법이 많이 제안되는데[4], 예를 들어 음성 신호를 네트워크의 입력으로 넣어 주기 전에, VAD(Voice activity detection) 및 feature extraction 과정을 거쳐 잡음을 제거하려는 접근법이 있다[5]. 최근에는 화자인식 성능을 제고하기 위해 CNN 과 같은 딥러닝 모델을 활용하기도 한다[6]. 학습된 CNN 의 feature vector 와 화자와의 유클리드 거리(Euclidean distance)를 기반으로 화자인식 하는 접근은 그 한 예이다[7]. 또한 음성 정보에는 감정 정보가 포함되어 있어, 이를 활용하여 화자인식의 정확도를 높이는 접근도 가능하다[8]. 또한 음성 정보를 낮은 차원의 정보로 변경한 i-vector[9]나 x-vector[10]를 기반으로 일련의 판별자(예: Probabilistic Linear Discriminant Analysis(PLDA))[11]를 활용하여 i-vector 또는 x-vector 에 등장하는 화자가 누구인지를 인식하는 방법을 제안하기도 한다[12]. 또한 GAN 기술을 활용하여 화자인식의 성능을 제고하는 노력도 최근 소개되고 있다[13].

Ⅱ. 본론

본 연구의 목적은 화자인식의 정확도를 개선하기 위해 화자가 발화한 문장에 등장하는 종결어미의 등장 특성을 바탕으로 화자인식 하는 방법을 제안하는 것으로 이를 위해 <그림 1>과 같이 진행하고자 한다. 또한 종결어미의 등장 특성을 바탕으로 화자인식 하는 방법을 제안하는 것이므로 음성 데이터를 텍스트 데이터로 변환하는 STT(Speech-to-Text) 단계는 생략한다. 화자인식을 위한 학습 데이터는 청각장애를 앓고 있는 분들을 위한 한글 자막 제작 모임인 '공이자막(https://blog.naver.com/dr_kisabi)'에서 제공하는 '응답하라 1994'라는 K-Drama 자막 데이터를 다운로드 받아서 사용하였다. 이후 전처리를 위해 다운로드 받은 '.smi' 파일을 '.txt' 파일 형태로 변환하고 먼저 이진 분류(Binary Classification)를 통한 화자인식을 실험하기 위해 남자 주연 1 명, 여자 주연 1 명으로 데이터셋을 구성한다. 생성된 데이터셋을 통해 종결어미 리스트를 생성하기 위해 형태소 분석기 RHINO 3.7 을 사용하였다. 해당 형태소 분석기를 통해 종결어미를 추출하고 중복 값이 존재하지 않게 종결어미 리스트를 생성한다. 문장 시퀀싱(sentence sequencing)이란 학습할 세션(session)에 등장하는 종결어미 빈도와 감성분석 결과를 통해 w-vector 를

생성하는 과정이다. 앞서 생성된 종결어미 리스트를 활용하여 각 화자별로 구성된 세션의 w-vector 를 계산하여 생성한다. 이후 10-Fold Cross-Validation 을 통해 각 세션에 생성된 w-vector 값을 활용하여 Cosine Similarity 계산 후 유사도 기반 판별을 통해 화자를 인식한다.



<그림 1> 전체적인 프로세스

Ⅲ. 결론

본 연구에서 우리는 종결어미에 대한 문장 시퀀싱에 의한 w-vector 를 기반으로 화자 인식하는 방법을 제안했다. 이를 통해 기존의 전통적인 음성인식 후 잡음 요소 제거 등을 수행하고 난 후에 음성 정보에만 의존하여 i-vector, x-vector, 딥러닝 등의 방법을 혼합하여 화자인식 하는 방법을 보완하려고 했다. 이를 위해 실제 존재하는 K-Drama 대본을 가지고 코퍼스를 구축하여 실험하였으며, 그 결과로 문장 시퀀싱 방법은 기존의 화자인식의 정확도를 더욱 개선한바 유용한 딥러닝 전처리 방법이 될 수 있음을 보였다. 본 연구는 종결어미에 기반한 문장 시퀀싱이라는 방법을 제안한 최초의 연구이며 화자인식의 정확도 제고에 유용함을 보인 것이다. 향후 실존하는 음성인식 시스템과 함께 활용되어 화자인식에 의한 지능형 대화 시스템과 이를 기반한 전자거래 및 각종 음성 기반 서비스에 활용될 것을 기대한다.

ACKNOWLEDGMENT

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea(NRF-2018S1A5A2A03036394).

참 고 문 헌

[1] Wang, N., Ching, P. C., Zheng, N., & Lee, T. (2010). Robust speaker recognition using denoised vocal source and vocal tract features. IEEE transactions on audio, speech, and language processing, 19(1), 196-205.

[2] Ramachandran, R. P., Farrell, K. R., Ramachandran, R., & Mammone, R. J. (2002). Speaker recognition—general classifier approaches and data fusion methods. Pattern recognition, 35(12), 2801-2821.

[3] 소순원. (2019). 자유 발화 데이터를 사용한 심층 인공 신경망 기반 화자 정보 분류 모델 개발 (Doctoral dissertation, 한양대학교).

[4] 강지훈, 김보람, 김규영, & 이상훈. (2020). MCE 기반의 다중 특징 파라미터 스코어의 결합을 통한 화자인식 성능 향상. 한국산학기술학회 논문지, 21(6), 679-686.

[5] 채석완. (2019). 교사-학생 학습 방법을 활용한 잡음에 강인한 화자 인식 (Doctoral dissertation, 서울대학교 대학원).

[6] Bhattacharya, G., Alam, M. J., & Kenny, P. (2019). Deep speaker recognition: Modular or monolithic?. In INTERSPEECH (pp. 1143-1147).

[7] 정희승, 윤상혁, & 박능수. (2020). 합성 샴 신경망을 이용한 화자 인식. 전기학회논문지, 69(1), 164-169.

[8] Huanjun, B., X. Mingxing, and F. Z. Thomas, “Emotion Attribute Projection for Speaker Recognition on Emotional Speech”. EUROSPEECH 2007, Antwerp, pp. 758-761, 2007.

[9] Dehak, N., P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788-798, 2011.

[10] Garcia-Romero, D., Snyder, D., Sell, G., McCree, A., Povey, D., & Khudanpur, S. (2019). x-Vector DNN Refinement with Full-Length Recordings for Speaker Recognition. In INTERSPEECH (pp. 1493-1496).

[11] Ioffe, S., “Probabilistic linear discriminant analysis,” Computer Vision-ECCV 2006, pp. 531-542, 2006.

[12] Snyder, D., Garcia-Romero, D., Sell, G., McCree, A., Povey, D., & Khudanpur, S. (2019, May). Speaker recognition for multi-speaker conversations using x-vectors. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5796-5800). IEEE.

[13] Chen, G., Chen, S., Fan, L., Du, X., Zhao, Z., Song, F., & Liu, Y. (2019). Who is real bob? adversarial attacks on speaker recognition systems. arXiv preprint arXiv:1911.01840.