한국어 종단형 화자 분리 기법을 위한 한국어 데이터셋 구축

문찬영, 한민현, 김남수

서울대학교 전기정보공학부 뉴미디어통신공동연구소

{cymoon, mhhan}@hi.snu.ac.kr, nkim@snu.ac.kr

Development of a Korean Dataset for End-to-End Speaker Diarization Techniques

Chan Yeong Moon, Min Hyun Han and Nam Soo Kim

Department of Electrical and Computer Engineering and INMC, Seoul National Univ

요 약

화자 분리는 여러 사람이 섞인 음성을 입력으로 받아 각 화자가 언제 발화했는지 판단하는 분야로, 이 기술을 학습하기 위한 데이터셋 구축은 필수적이다. 현재 영어 기반 데이터셋을 활용한 종단형 신경망 화자 분리(EEND) 모델 연구는 활발히 진행되고 있지만, 한국어에 특화된 데이터셋은 여전히 부족하다. 이에 본 연구에서는 한국어 음성 데이터를 기반으로 다양한 소음 환경에서의 상황을 가상으로 설정하여 시뮬레이션한 데이터셋을 생성하고, 이를 활용해 한국어 음성에서도 EEND 모델이 효과적으로 학습할 수 있도록 하는 데이터셋 구축 기법을 제안한다. 또한 실험을 통해 구축된 한국어 데이터셋의 유용성과 타당성을 분석하였다.

Ⅰ. 서 론

화자 분리는 여러 화자가 발화하는 상황에서 각 화자가 어느 시점에 발화하고 있는지를 식별하는 기술로, 각 시간 프레임에서 화자의 존재 확률을 계산하여 화자별 발화 구간을 예측한다.

현재 화자 분리은 딥러닝 기반으로 크게 두 가지 접근 방식에서 연구가 활발히 진행되고 있다. 첫 번째는 음성 구간 추출, 화자 인식, 군집화 모듈 등 여러 개의 모듈을 결합하여 화자를 구분하는 군집 기반 방식이고, 두 번째는 화자 분리을 하나의 시스템으로 처리하는 종단형 방법(EEND) [1, 2]이다. 특히, 종단형 화자 분리 시스템에 대한 연구가 활발히 진행되고 있으며, 이를 학습하기 위해 두 가지 종류의 데이터셋이 사용된다.

첫 번째는 다양한 소음 환경과 서로 다른 발화자를 가정하여 구성된 시뮬레이션 데이터셋이며, 두 번째는 별도의 가공 없이 여러 화자가 실제로 대화하고 있는 데이터를 활용한 실제 데이터셋이다. 이러한 데이터 구성은 CALLHOME1,2(각각 8.5시간)[3], DIHARD Ⅲ(34시간)[4] 등 실제 음성 데이터의 학습 데이터 양이 적기 때문에, 많은 연구에서 시스템 성능을 고도화하기 위해 대화 상황을 가정한 대규모 시뮬레이션 데이터셋으로 학습하고, 실제 데이터로 적응하는 방식이 효과적임을 보여준다.

그러나 화자 분리을 위한 다양한 데이터셋이 연구되고 있음에도 불구하고, 한국어를 활용한 화자 분리 데이터셋은 여전히 부족한 실정이다. 이에 본 연구에서는 한국어 음성에 대해 강건한 성능을 발휘할 수 있도록 한국어 음성 데이터셋인 SiTEC의 낭독 문장 음성과 다양한 소음 데이터를 포함한 MUSAN[5]을 활용하여, 다양한 한국어 상황을 가정한 시뮬레이션 데이터셋을 구축하였다. 또한, 구축된 데이터셋의 유용성과 타당성을 분석하여 한국어 화자 분리 기술 발전에 기여하고자 한다.

Ⅱ.종단형 화자 분리 기법(End-to-End Neural Diarization Method)

화자 분리 연구는 최근 종단형 화자 분리 구조를 중심으로 발전하고 있으며, 그중 가장 널리 사용되는 방법은 EEND(End-to-End Neural Diarization) [1] 이다. EEND는 음성 데이터를 입력으로 받아 화자별 발화 구간을 추정하는 종단형 학습 시스템으로, 세 가지 주요 구성 요소로 이루어져 있다. 첫 번째는 음성을 MFCC와 같은 특징 벡터로 변환하는 전처리 모듈, 두 번째는 Transformer Encoder를 활용해 이 특징 벡터를 유의미한 정보로 가공하는 인코더 모듈, 마지막으로 간단한 Feed Forward Network를 통해 각 화자의 발화를 분류하는 디코더 모듈으로 구성되어 있다.

하지만 EEND는 디코더 구조가 고정된 형태로 설계되어 있어, 대화에서 화자 수가 미리 정해지지 않은 상황에서는 사용이 어렵다는 한계가 있다. 그러므로 이를 해결하기 위해 제안된 방법이 바로 EEND-EDA(End-to-End Neural Diarization - Encoder-Decoder Attractor)[2] 이다. EEND-EDA는 기존 EEND의 구조를 기반으로, 디코더 모듈을 단순한 Feed Forwad Network 대신 RNN Encoder-Decoder의 형태의 EDA(Encoder-Decoder Attractor) 모듈을 통해 화자의 수에 맞는 attractor를 생성한다. EDA는 입력 음성을 Transformer Encoder를 통해 처리한 후, 문맥 정보를 담은 벡터를 RNN Encoder로 생성하고, 이를 조건부 입력으로 활용하여 attractor를 생성하는 방식으로 동작한다.

이러한 연구가 활발하게 진행되고 있음에도 현재 한국어로 구성된 화자 분리 데이터셋에 대한 연구가 진행되고 있지 않거나 그 수준이 미비하여 한국어 음성을 활용하여 진행한 화자 분리의 연구가 많이 이루어지고 있지 않다. 따라서 본 논문에서는 한국어 화자 분리를 위한 데이터셋을 구축하고 이에 따른 유용성을 확인하고자 한다.

Ⅲ. 한국어 화자 분리 기법 학습을 위한 시뮬레이션 데이터셋 구축 기법 (Simulation Dataset for Training Korean Speaker Diarization Techniques)

한국어 화자 분리 시뮬레이션 데이터셋을 구축하기 위해서는 먼저, 각각의 음성은 단일 화자 음성으로 구성되어 있어야 하며, 둘째, 화자 정보가 명확히 표시된 음성 데이터셋이어야 한다. 그리고 마지막으로는 Voice activity detection 모델을 활용하여 어느 구간이 화자 발화 구간인지에 대해 가공하여 화자 발화 구간에 대한 라벨을 만들어 실험을 진행하기 위해서는 조용한 상태에서 녹음된 음성이어야 한다. 이에 우리는 한국어 음성 데이터셋으로 SiTEC 데이터셋이 내 Dictation용 낭독 문장 음성 데이터셋이 적합하다 판단하여 이를 활용하였다.

Dictation용 낭독 문장 음성 데이터셋은 16k sampling rate의 총 약 20,800개 문장으로 구성되어 있으며, 400명의 서로 다른 화자(남성 200명, 여성 200명)가 각각 약 100개의 문장을 발화한 데이터를 포함하고 있다. 이러한 구성은 각 화자별로 단일 문장이 아닌 다양한 문장 발화 상황을 상정할 수 있도록 해준다. 또한, 남성 간, 여성 간, 남성과 여성 간의 대화를 포함한 다양한 다화자 시나리오를 효과적으로 구성할 수 있다. 이를 활용하여 음성을 겹침으로써 다화자 발화 상황을 구성하였습니다.

다양한 소음 상황을 가정하여 실험을 진행하기 위해, augmentation에서 자주 활용되었으며 다화자 음성 데이터셋인 wsj0-mix[7]에서 활용된 MUSAN 데이터셋을 활용하였다. MUSAN은 음악, 음성, 그리고 다양한 소음 데이터를 포함하고 있다. 본 연구에서는 다화자 음성 데이터셋에 소음을 추가하여 다양한 소음 환경을 시뮬레이션하는 형식으로 MUSAN[5] 데이터셋의 Music과 Noise 데이터를 활용하였다. MUSAN의 Speech Noise 데이터는 다화자 발화 환경의 소음으로 적합하지 않아 사용하지 않았다. Music 데이터는 클래식, 재즈, 팝 등 다양한 장르의 음악으로 구성되어 있으며, Noise 데이터는 거리 소음, 사무실 소음, 자연 소음 등 현실적인 환경에서 발생할 수 있는 다양한 소음 유형을 포함하고 있다. 이러한 소음 데이터를 활용함으로써 실제 환경에서의 소음 상황을 보다 정밀하게 모사하고, 다화자 화자 분리 모델의 성능을 평가할 수 있는 실험 환경을 구축하였다.

방법론적으로, 본 연구에서는 LibriMix[6]와 wsj0-mix[7] 논문을 참고하여 데이터셋을 구성하였다. 먼저, 서로 다른 두 화자의 음성을 랜덤하게 선택하되, 동일한 화자가 선택될 경우 이를 예외로 처리하여 새로운 음성 파일을 다시 선택하도록 설계하였습니다. 선택된 음성 파일에 대해서는 SpeechBrain의 VAD(Voice Activity Detection) 모듈을 활용하여 각 음성 파일의 발화 구간을 추출하였다. Kaldi simulation document에서 각각의 음성이 소음이 없는 환경에서 녹음되었을 때, 간단한 VAD를 활용하여 발화 구간을 식별하는 방법이 충분히 효과적인 방법임이 검증되었다. 이를 바탕으로, 본 연구에서도 VAD를 활용해 시뮬레이션 데이터셋의 화자 발화 구간 라벨을 구성하였다. 그러나 딥러닝 기반 VAD 라벨링은 간혹 잘못된 판단을 내릴 수 있다. 대표적인 사례로, 발화가 실제로 시작되지 않은 0초 지점에서 발화가 시작되었다고 표시되거나, 음성 파일의 끝까지 발화가 지속된다고 잘못 표시되는 경우가 있다. 이러한 문제를 방지하기 위해, 이러한 파일은 랜덤 샘플링 과정에서 제외되도록 설계하였다. 이후, 문제 파일을 제외한 상태에서 두 음성을 다시 랜덤하게 선택하여 데이터셋을 구성하였다.

또한, 소음 음성은 다양한 소음 환경을 시뮬레이션하기 위해 SNR(Signal-to-Noise Ratio) 값을 10, 15, 20 dB로 설정하였으며, 이를 균등 분포(Uniform Distribution)를 기반으로 랜덤하게 샘플링하였다. 이를 통해, 소음 크기에 따른 다양한 시나리오를 포함하는 데이터셋을 구축하였다.

IV. 실험

데이터셋의 유용성을 검증하기 위해, 현재 화자 분리 분야에서 가장 널리 사용되는 EEND-EDA[2] 모델을 활용하였다. 한국어 시뮬레이션 음성 데이터셋 SiTEC2Mix의 성능을 평가하고 비교하기 위해, 영어 음성 데이터셋 Libri2Mix 데이터셋을 사용하였다.

Libri2Mix[6] 데이터셋은 총 13,900개의 음성 파일로 구성되어 있으며, 약 58시간 분량의 데이터이며, 성능 평가를 위해 별도로 구성된 test 데이터셋은 3,000개 파일로 약 11시간 분량으로 구성된다.

우리가 제안하는 SiTEC2Mix 데이터셋은 EEND-EDA [2] 모델의 성능을 유지하면서도 한국어 음성에 적합하도록 설계된 데이터셋이다. 따라서 학습용 데이터셋은 1,500개의 파일로 약 2시간 반 분량이며, 검증을 위한 test 데이터셋은 300개의 파일로 약 30분 분량으로 구성하였다.

실험은 크게 서로 다른 두 가지의 방식을 비교하여 진행했다. training 방식의 경우, 학습 단계에서 Libri2Mix와 SiTEC2Mix 데이터셋을 결합하여 학습 데이터셋을 구성하였고, 이를 기반으로 모델을 학습시켰다. fine-tuning 방법의 경우, 상대적으로 더 큰 Libri2Mix 데이터셋으로 사전 학습된 (pretrained) 모델을 활용하여 SiTEC2Mix 데이터를 추가적으로 사용해 모델에 적응시켰다.

Training Dataset	ALL	SiTEC2Mix (Korean)	Libri2Mix (English)
Only Libri2Mix (Baseline)	7.32	28.33	5.23
Libri2Mix + SiTEC2Mix (training)	**6.22**	**11.78**	**5.67**
Libri2Mix + SiTEC2Mix (fine-tuning)	**6.42**	**16.25**	**5.45**

〔표 1〕 실험 결과 (DER〔%〕)

결과적으로, 한국어와 영어가 혼합된 데이터셋에 대해 Libri2Mix 데이터만으로 학습한 경우에 비해 각각 15.02%와 12.29%의 상대 성능 향상을 보였다. 세부적으로는, 한국어 데이터셋인 SiTEC2Mix에 대해 16.55%와 12.08%의 절대 성능 향상이 있었으나, Libri2Mix에 대해서는 각각 0.44%와 0.22%의 절대 성능 감소가 나타났으며, 이는 영어 음성에 대한 성능 저하가 미미한 수준임을 의미한다.

V. 결론

본 연구에서는 한국어 화자 분리를 위한 시뮬레이션 데이터셋 SiTEC2Mix를 구축하고, EEND-EDA 모델을 활용하여 성능을 평가하였다. 제안된 데이터셋은 Libri2Mix와 결합하여 학습 및 미세 조정 과정에서 기존 방법 대비 유의미한 성능 향상을 보여주었으며, 특히 한국어 음성 환경에서의 강건한 성능을 확인할 수 있었다. 이는 한국어 화자 분리 기술 발전에 있어 실질적인 기여를 하여 향후 다양한 다화자 음성 환경 및 실제 데이터 적용 가능성을 높이는 데 기여할 것으로 기대된다.

ACKNOWLEDGMENT

이 논문은 2024년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의해 지원되었음.

참 고 문 헌

[1] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe “End-to-end neural speaker diarization with self-attention,” in *Proc. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 296-303. 2019.

[2] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu, “End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors,” *in Proc. Interspeech,* pp. 269 - 273, 2020.

[3] A. F Martin and M. A Przybocki, “2000 NIST speaker recognition evaluation,” *in Philadelphia: Linguistic Data Consortium*, 2001.

[4] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, “The third dihard diarization challenge,” arXiv preprint arXiv:2012.01477, 2020.

[5] David Snyder, Guoguo Chen, and Daniel Povey, “MUSAN: A music, speech, and noise corpus,”arXiv:1510.08484, 2015.

[6] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent, “LibriMix: An opensource dataset for generalizable speech separation,” arXiv preprint arXiv:2005.11262, 2020.

[7] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: discriminative embeddings for segmentation and separation,” in ICASSP, 2016, pp. 31 - 35.