화자 검증 시스템을 활용한 종단형 화자 분할 기법

문찬영, 한민현, 김남수
서울대학교 전기정보공학부 뉴미디어통신공동연구소
{cymoon, mhhan}@hi.snu.ac.kr, nkim@snu.ac.kr

End-to-End Speaker Diarization Method Using Speaker Verification System

Chan Yeong Moon, Min Hyun Han and Nam Soo Kim
Department of Electrical and Computer Engineering and INMC, Seoul National Univ

요 약

화자 분할 시스템은 여러 사람이 섞인 음성을 입력으로 받아 각각의 화자가 언제 발화하는지 판단하는 분야이다. 따라서 각 시간마다 특징 벡터를 뽑아내는 것 또한 중요하지만 각 화자에 대한 임베딩 벡터를 잘 뽑아내는 것 또한 중요한 과제이다. 이 때, 본 논문에서는 기존에 존재하는 종단형 기법을 활용해 화자 임베딩 추출 부분에 추가적으로 화자 검증 시스템을 통해 추출한 임베딩 벡터를 가공하여 추가적인 정보로 활용함으로써 더 좋은 임베딩 벡터를 추출하는 기법을 제안한다. 실험을 통해 추가적인 화자 정보에 대한 타당성과 유효함을 검증하였다.

Ⅰ. 서 론

화자 분리는 여러 화자가 발화하는 상황에서 각각의 화자가 언제 발화하고 있는지에 대해 식별하는 분야이다. 즉, 각각의 시간 프레임에 대한 화자 존재 확률을 계산해 각각의 화자마다 발화 구간을 예측하는 분야라고 이야기할 수 있다.

초기 딥러닝을 활용한 화자 분할의 경우 화자 분할 시스템을 따로 고안하는 방식이 아닌 미리 학습된 화자 검증 시스템을 활용하여 군집화 방법을 토대로 화자 분할을 실행하곤 하였다. 화자 검증 시스템의 경우 입력 음성에 대해 길이에 상관없이 하나의 고정된 차원의 벡터를 추출하게 되는데, 입력 음성을 짧은 길이로 잘라 각각의 짧은 구간을 화자 검증 시스템의 입력으로 주게 되면 각 시간에 대한 벡터들을 얻을 수 있고, 군집화를 통해 어느 시간에 누가 말하고 있는지에 대해 식별한다. 그러나 군집화를 통한 화자 분할의 경우 목소리가 겹쳐있는 부분의 경우 오로지 한 군집에 대해 속하는 방식으로 화자 분할이 이뤄지기 때문에 중첩 구간에 대해 잘 처리하지 못한다는 한계가 존재한다.

따라서 최근 딥러닝을 활용한 화자 분할의 경우 화자 검증 시스템을 활용하는 방법이 아닌 자체적으로 종단형 화자 분할 시스템을 사용하여 중첩 구간에 대해서도 처리할 수 있도록 발전해왔다. 초기 종단형 화자 분할 시스템의 경우 encoder 뒤에 간단하게 classifier를 decoder로 달아 단순히 고정된 화자 수에 대해서 화자 분할을 하도록 발명이 되었다. 그러나 고정된 화자 수에 대해서가 아닌, 여러 화자 수에 대해서 식별하는 화자 분할의 경우 입력 음성의 화자 수에 따라 변하는 classifier를 만들어야 한다. 그러므로 encoder-decoder 구조의 EDA(encoder-decoder attractor)를 사용하여 이러한 한계를 극복할 수 있게끔 하였다. 하지만 EDA에서 decoder의 경우, RNN encoder에서 나온 시계열 정보인 hidden representation만을 정보로 사용하여 화자 수에 대해 유연한 classifier인 attractor를 추출하게 되는데, 본 논문에서는 화자 검증 시스템의 정보를 활용하는 시스템을 제안하고자 한다.

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

최근 화자 분할 분야에서 화자 검증 시스템이 아닌 자체적인 화자 분할 시스템을 적용하면서 어떤 형태의 시스템을 사용할 때 가장 적합한가에 대한 많은 연구가 진행되었다. 그중에서 현재 가장 대중적으로 사용되는 화자 분할 기법은 [1]에서 제안하는 방법인 EEND(End-to-End Neural Diarization, 종단형 화자 분할 기법)이다. [1]에서 제안하는 화자 분할의 방법의 경우, 3가지의 모듈이 결합된 형태로 이루어져 있다. 먼저, 음성을 MFCC와 같은 특징 벡터로 만들어주는 전처리 부분, 전처리 된 벡터들을 Transformer encoder를 사용하여 유의미한 정보로 가공시키는 encoding 부분, 마지막으로 간단한 Feed forward network를 활용하여 각 화자가 말하고 있는지 판별하는 즉, classifier 역할을 하는 decoder 부분으로 이루어져 있다.

그러나 이러한 EEND 기법의 경우 decoder의 형태가 고정된 벡터로 존재하기 때문에, 고정된 화자 수에 대해서만 화자 분할을 할 수 있다. 따라서 화자가 몇 명이 말하는지 알 수 없는 경우에 대해서 화자 분할을 할 수 없는 문제가 발생하였다. 이러한 문제를 해결하고자 나온 모델이 바로 [2]에서 제안하는 EEND-EDA (End-to-End Neural Diarization - Encoder decoder attractor)이다. EEND-EDA의 경우 EEND의 encoder 부분 뒤에 추가적으로 RNN(Recurrent Neural Network)을 encoder-decoder를 접합하여 화자 수만큼 classifier인 attractor를 뽑아낼 수 있다.

Ⅲ. 화자 검증 임베딩을 활용한 종단형 화자 분할 기법(End-to-End Speaker Diarization Method Using Speaker Verification System)

EEND-EDA에서 EDA 부분을 살펴보면 RNN encoder의 경우 입력으로 가공된 특징 벡터를 사용하고 이를 통해 hidden representation를 뽑아내고, 다시 RNN decoder에 hidden representation으로 들어간다. 이 때, RNN decoder에 입력으로 오로지 hidden representation의 정보만 들어가고 실제 input으로는 영벡터들이 들어간다. 따라서 본 논문에서는 영벡터를 대신해서 추가적인 입력으로 화자 검증 임베딩을 활용하는 방법을 제안하고자 한다.

화자 분할 분야의 경우, 입력 음성에 대해 여러 화자가 섞여 있기 때문에 여러 화자에 대한 정보를 넣어줄 수 있어야 한다. 따라서 미리 학습된 화자 검증 시스템을 활용하여 화자가 섞이기 전의 음성을 활용해 각 화자에 대한 임베딩을 추출한 뒤, 모든 화자의 임베딩을 합해 모든 화자에 대한 정보를 넣어주고자 하였다. 따라서 입력 음성에 학습 가능한 토큰을 붙여 EEND encoder에 넣어 그 결과가 모든 화자 임베딩의 합이 되도록 loss 함수를 추가하였다. 또한 EDA encoder의 입력 부분에서는 encoder

을 통과한 토큰이 RNN에 들어가지 않도록 다시 떼어내 주었고, 대신 화자 수만큼 이를 복제하여 RNN decoder의 입력으로 사용하게 하였다.

이에 따라 loss 함수의 구성을 보게 되면 EEND에서 사용하는, diarization에 대한 loss인 PIT(Permutation Invariant Training) loss와 본 논문에서 제안하는 토큰 임베딩이 화자 검증 임베딩의 합의 값이 될 수 있도록 하는 distillation loss 2가지를 사용하게 되는데, PIT loss의 경우 이를 수식으로 표현하면 다음과 같다.

$$\Phi^{*} = \underset{\Phi \in perm(S)}{\arg\min} \sum_{s=1}^{S} \sum_{t=1}^{T} BCE(z_{s,t}, y_{\Phi_s,t}), \quad (1)$$

$$L_{PIT} = \sum_{s=1}^{S} \sum_{t=1}^{T} BCE(z_{s,t}, y_{\Phi_s^{*},t}), \quad (2)$$

이 때, $perm(S)$는 S명의 화자에 대해 가능한 정답 라벨의 집합을 의미하고, S는 입력 음성에 들어있는 총 화자의 수를, T는 시간 축으로의 프레임 수, $z_{s,t}$는 transformer encoder layer를 통과한 결과값 벡터와 attractor 벡터간의 벡터곱을 통해 나온 $S \times T$(화자 수 X 전체 시간) 형태의 벡터값, $y_{\Phi_s,t}$, $y_{\Phi_s^{*},t}$는 각각 화자에 대한 정답 라벨 조합과 최적의 정답 라벨을 의미한다. BCE 함수의 경우 Binary Cross-Entropy의 약자로 그 값이 정답인지에 대한 확률값을 나타낸다. 나머지 distillation loss의 경우 아래와 같은 수식으로 나타낼 수 있다.

$$L_{distill} = MSE(y_{token}, \sum_{s=1}^{S} e_s), \quad (3)$$

이 때, y_{token}은 토큰 임베딩이 Transformer encoder를 통과한 결과값을 의미하고, e_s는 화자 검증 시스템을 통과해 나온 각 화자에 대한 임베딩을, MSE함수는 Mean Square Error를 의미한다. 따라서 전체 loss 함수의 경우 이 두 함수를 더한 값이 되므로 아래와 같이 표현할 수 있다.

$$L_{total} = L_{PIT} + L_{distill}. \quad (4)$$

IV. 실험

입력 음성을 위한 acoustic feature의 경우 32ms 길이 및 8ms 간격의 hanning window를 통한 80차원의 Log Mel Filter Bank를 추출하여 사용하였으며, 이에 따라 토큰 임베딩의 차원 또한 80차원으로 구성되어 있다. acoustic feature를 다시 한번 더 가공하는 단계를 거치게 되는데, 이 때 크기 15를 갖는 1D convolution을 통해 192차원의 임베딩을 얻어낸다. 이 프레임 단위 입력을 처리하는 네트워크로는 [1]에서 사용한 것과 같이 4 head, 4 layer transformer encoder를 사용한다. EDA의 경우 각각 1 layer RNN의 encoder-decoder 구조로 이루어져 있으며 RNN encoder의 경우 input으로 transformer encoder의 192차원의 결과값을 사용하고, 이때 나오는 context vector를 RNN decoder에 context로 활용한다. 또한, RNN decoder의 input을 넣을 때 사용되는 화자 검증 시스템에서 사용되는 최신 모델인 SKA-TDNN[3]을 사용해 화자 192차원 임베딩을 추출한다.

실험에 사용된 데이터는 LibriMix[4] dataset을 사용하였다. Librimix dataset은 Librispeech dataset에서 train-clean100/dev-clean/test-clean과 noise 소리인 WHAM!을 섞어 만들었으며, 이 중에서 우리는 2명의 화자가 섞인 음성에 대한 결과를 확인하기 위해 Libri2Mix를 사용하였고, 16kHz sampling rate와 max mode로 설정하여 실험을 진행하였다.

실험 결과는 총 2가지 모델에 대해 비교하였다. 첫 번째 모델은 기존에 널리 사용되는 모델인 EEND-EDA를 사용한 결과이며, 두 번째 모델은 화자 검증 시스템을 활용한 EEND 모델이다. 실험에서 사용한 성능 지표의 경우 DER(Diarization Error Rate)를 사용하였다. DER은 화자 분할에서 사용하는 성능 지표이며, 여러 화자가 섞인 음성에 대해 각 화자 발화 구간을 나눌 때 발생할 수 있는 모든 error 프레임을 전체 프레임 수로 나눈 값이다. 즉, 3가지 error의 합으로 구성되어 있으며, 각각 화자에 대해 햇갈린 경우 Speaker Confusion, 화자가 발화하지 않았는데 발화했다고 말하는 경우 False Alarm, 화자가 발화하였는데 발화하지 않았다고 말하는 경우 MISS로 부르며 이를 모두 더한 값이 최종 성능 지표인 DER이 된다.

Models	DER[%]	Speaker Confusion[%]	False Alarm[%]	MISS[%]
EEND-EDA	4.79	0.27	2.16	2.36
Proposed	**4.52**	**0.26**	**2.34**	**1.92**

〔표 1〕 실험 결과

V. 결론

본 논문에서는 화자 검증 시스템을 활용한 종단간 화자 분할 기법을 제안한다. 본 기법은 여러 화자가 혼합된 경우를 가정해 각 화자에 대한 임베딩을 화자 검증 시스템인 SKA-TDNN 추출한 뒤 더한 값이 attractor 추출 부분인 RNN decoder의 input으로 사용하여 attractor가 화자 분할을 위한 임베딩을 추출할 때 추가적인 정보를 활용할 수 있게끔 만든다. 실험을 통해 화자 검증 시스템이 화자 분할 기법에 효과적인 정보를 제공하는 것을 검증하였다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부-경찰청)의 재원으로 과학기술사업화진흥원(과학치안 공공연구성과 실용화 촉진 시범사업)의 지원을 받아 수행된 연구임(No. RS-2023-00235082)

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] S. H. Mun, J.-w. Jung, M. H. Han, and N. S. Kim, “Frequency and multi-scale selective kernel attention for speaker verification,” in *Proc. 2022 IEEE Spoken Language Technology Workshop (SLT).* IEEE, pp. 548-554, 2023.

[4]J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, E. Vincent “LibriMix: An Open-Source Dataset for Generalizable Speech Separation,” “LibriMix: An opensource dataset for generalizable speech separation,” *arXiv preprint arXiv:2005.11262,* 2020.