한국음향학회지 제42권 제6호 pp. 536~543 (2023)
The Journal of the Acoustical Society of Korea Vol.42, No.6 (2023)
https://doi.org/10.7776/ASK.2023.42.6.536

pISSN : 1225-4428
eISSN : 2287-3775

단일 레이블 분류를 이용한 종단 간 화자 분할 시스템 성능 향상에 관한 연구

A study on end-to-end speaker diarization system using single-label classification

정재희,[1] 김우일[1†]

(Jaehee Jung[1] and Wooil Kim[1†])

[1]인천대학교 컴퓨터공학부

(Received August 8, 2023; accepted September 6, 2023)

초 록: 다수의 화자가 존재하는 음성에서 "누가 언제 발화했는가?"에 대해 레이블링하는 화자 분할은 발화 중첩 구간에 대한 레이블링과 화자 분할 모델의 최적화를 위해 심층 신경망 기반의 종단 간 방법에 대해 연구되었다. 대부분 심층 신경망 기반의 종단 간 화자 분할 시스템은 음성의 각 프레임에서 발화한 모든 화자의 레이블들을 추정하는 다중 레이블 분류 문제로 분할을 수행한다. 다중 레이블 기반의 화자 분할 시스템은 임계값을 어떤 값으로 설정하는지에 따라 모델의 성능이 많이 달라진다. 본 논문에서는 임계값 없이 화자 분할을 수행할 수 있도록 단일 레이블 분류를 이용한 화자 분할 시스템에 대해 연구하였다. 제안하는 화자 분할 시스템은 기존의 화자 레이블을 단일 레이블 형태로 변환하여 모델의 출력으로부터 레이블을 바로 추정한다. 훈련에서는 화자 레이블 순열을 고려하기 위해 Permutation Invariant Training(PIT) 손실함수와 교차 엔트로피 손실함수를 조합하여 사용하였다. 또한 심층 구조를 갖는 모델의 효과적인 학습을 위해 화자 분할 모델에 잔차 연결 구조를 추가하였다. 실험은 Librispeech 데이터베이스를 이용해 화자 2명에 대한 시뮬레이션 잡음 데이터를 생성하여 사용하였다. Diarization Error Rate(DER) 성능 평가 지수를 이용해 제안한 방법과 베이스라인 모델을 비교 평가했을 때, 제안한 방법이 임계값 없이 분할이 가능하며, 약 20.7 %만큼 향상된 성능을 보였다.

핵심용어: 종단 간 화자 분할, 단일 레이블, 잔차 연결 구조, 손실함수 조합

ABSTRACT: Speaker diarization, which labels for "who spoken when?" in speech with multiple speakers, has been studied on a deep neural network-based end-to-end method for labeling on speech overlap and optimization of speaker diarization models. Most deep neural network-based end-to-end speaker diarization systems perform multi-label classification problem that predicts the labels of all speakers spoken in each frame of speech. However, the performance of the multi-label-based model varies greatly depending on what the threshold is set to. In this paper, it is studied a speaker diarization system using single-label classification so that speaker diarization can be performed without thresholds. The proposed model estimate labels from the output of the model by converting speaker labels into a single label. To consider speaker label permutations in the training, the proposed model is used a combination of Permutation Invariant Training (PIT) loss and cross-entropy loss. In addition, how to add the residual connection structures to model is studied for effective learning of speaker diarization models with deep structures. The experiment used the Librispech database to generate and use simulated noise data for two speakers. When compared with the proposed method and baseline model using the Diarization Error Rate (DER) performance the proposed method can be labeling without threshold, and it has improved performance by about 20.7 %.

Keywords: End-to-End speaker diarization, Single label, Residual connection, Loss combination

PACS numbers: 43.72.Bs, 43.72.Ne

†**Corresponding author:** Wooil Kim (wikim@inu.ac.kr)
Department of Computer Science and Engineering, Incheon National University, 119 Academy-ro, Yeonsu-gu, Incheon 22012, Republic of Korea
(Tel: 82-32-835-8459, Fax: 82-32-835-0780)

I. 서 론

화자 분할은 다수의 화자가 존재하는 음성에서 누가 언제 발화했는지를 레이블링하는 기술로 다수의 화자가 존재하는 전화나 미팅 분석 등 다양한 음성 응용 애플리케이션에 활용된다.

전통적인 화자 분할 시스템은 클러스터링을 기반으로 수행되며, 음성 검출기와 화자 특징 추출기를 포함한 여러 모듈로 구성되어 있다.[1-3] 먼저, 묵음 구간을 제거하기 위해 음성 검출을 수행한다. 묵음이 제거된 음성은 일정 길이 또는 화자 변화 감지 등을 통해 얻은 구간들로 구간화를 수행한다. 각 세그먼트들은 화자 특징 추출기를 통해 x-vector, d-vector 등의 화자 특징을 추출한 뒤 Agglomerative Hierarchical Clustering(AHC), spectral clustering 등의 클러스터링 과정을 수행하여 화자 레이블을 얻는다. 이와 같은 클러스터링 기반의 화자 분할 방법은 중첩 구간에 대해 여러 화자 레이블을 예측할 수 없고, 여러 모듈을 한 번에 최적화할 수 없다는 단점을 가진다.

클러스터링 기반 기법의 단점을 보완하기 위해 심층 신경망을 기반으로 하는 종단 간 화자 분할 시스템이 제안되었다.[4-8] 종단 간 화자 분할 시스템은 화자를 하나의 레이블로 보고 다중 레이블 분류 문제로 분할을 수행한다. 다중 레이블을 갖도록 화자 분할을 수행할 수 있으므로 중첩된 구간에 대해서도 정확한 레이블을 얻을 수 있다. 종단 간 화자 분할 모델에서는 프레임별 화자 레이블의 사후 확률을 추정한다. 추정된 확률과 임의로 설정한 임계값과의 비교를 통해 각 프레임에서 발화한 화자들의 최종 레이블을 얻을 수 있다. 그러나 이러한 화자 분할 모델의 성능은 임계값에 의존된다. 임의로 설정한 임계값에 따라 모델의 성능이 많이 달라질 뿐만 아니라, 음성 종류나 상태에 따라서도 분할 성능이 매우 다를 수 있다.

이런 임계값 의존 문제를 해결하기 위해 단일 레이블 분류 문제로 변환하여 화자 분할을 수행하는 Speaker Embedding-aware Neural Diarization(SEND)가 제안되었다.[6] 제안된 기법은 다중 레이블 분류 기반의 화자 분할을 수행하지만, 모델을 통해 추정된 출력을 단일 레이블로 변환한 뒤에 화자 레이블을 추정한다. 또한 화자 특징 학습을 위한 추가 화자 특징 네트워크를 같이 학습한다. 제안된 기법에서는 모델을 통해 단일 레이블을 바로 추정하지는 않는다.

본 논문에서는 임계값 의존 문제를 해결하기 위해 단일 레이블 분류 기반의 화자 분할 모델에 대해 연구를 진행하였다. SEND[6]와 달리 모델의 출력으로 바로 단일 레이블을 얻을 수 있도록 하였고, 화자 레이블의 순열을 고려하여 모델을 학습하기 위해 기존에 사용되었던 Permutation Invariant Training(PIT) 손실함수[7,8]와 교차 엔트로피 손실함수를 조합하여 사용하였다. 심층 신경망 구조를 갖는 모델은 결과를 추정하는데 상대적으로 낮은 계층의 영향을 덜 받게 된다.[9] 이에 따라 본 논문에서는 낮은 계층 학습에 도움이 되고 결과 추정 시에도 출력 계층에만 의존하지 않도록 잔차 연결 구조를 추가하는 방법에 대해서도 연구를 진행하였다. 종단 간 화자 분할을 위해 Self-Attentive End-to-End Neural Diarization(SA-EEND)[8] 모델을 베이스라인 시스템으로 하여 연구를 수행하였다.

다음 2장에서는 종단 간 화자 분할 수행 과정과 베이스라인으로 사용한 화자 분할 모델 구조를 설명하고, 3장에서는 제안하는 단일 레이블 분류 기반의 화자 분할 모델과 잔차 연결 구조를 추가한 모델에 대해 설명한다. 4장에서는 사용한 데이터 및 실험과 실험 결과에 대해 논의하고 5장에서는 결론을 맺는다.

II. 종단 간 화자 분할 시스템

2.1 종단 간 화자 분할[4-8]

종단 간 화자 분할 모델은 다수의 화자가 포함된 음성으로부터 추출된 특징을 입력으로 사용하여 프레임별로 화자들의 레이블에 대한 사후 확률을 예측한다. 현재 대부분의 화자 분할은 각 프레임에 포함되는 화자들의 레이블을 추정하는 다중 레이블 분류 문제로 설정하고 모델 학습을 수행한다.

화자 분할 모델의 입력으로 사용되는 음성 특징으로 로그-멜 스펙트럼 또는 Mel-Frequency Cepstral Coefficient(MFCC)를 많이 사용한다. 본 논문에서는 2명 화자가 혼합된 음성의 로그-멜 스펙트럼을 이용

하여 화자 분할을 수행하였다. 입력 스펙트럼이 X일 때, 모델을 통해 추정된 프레임별 화자 레이블의 사후 확률 Z는 다음과 같다.

$$X = [x_1, ..., x_t, ..., x_T],\ x_t \in \mathrm{R}^D. \tag{1}$$

$$Z = [z_1, ..., z_t, ..., z_T],\ z_t \in [0,1]^C. \tag{2}$$

화자 분할 모델의 입력 X는 실수값을 갖는 D차원의 T개 프레임으로 이루어져 있다. Z는 0에서 1 사이의 확률값을 갖는 C차원의 T개의 프레임으로 구성되어 있다. 여기서, C는 음성에 존재하는 화자의 최대 수이다. 테스트 단계에서 모델을 통해 추정된 프레임별 화자 레이블의 사후 확률은 다음과 같이 임계값을 이용해 각 프레임에서 발화한 화자 레이블을 결정할 수 있다.

$$\hat{g}_{t,c} = \begin{cases} 1 & if\ z_{t,c} \geq Threshold \\ 0 & else \end{cases}. \tag{3}$$

$$\hat{G} = [\hat{g}_1, ..., \hat{g}_t, ..., \hat{g}_T],\ \hat{g}_t \in \{0,1\}^C. \tag{4}$$

Eq. (3)에서 $\hat{g}_{t,c}$는 t번째 프레임에서 c번째 화자 레이블을 나타내며, Z의 t번째 프레임, c번째 화자 레이블에 대한 사후 확률 $z_{t,c}$과 설정한 임계값인 $Threshold$와의 비교를 통해 값이 결정된다. $\hat{g}_{t,c}$ 값이 1인 경우, t번째 프레임에서 c번째 화자가 발화했음을 나타낸다. $\hat{G}$는 추정한 프레임별 발화한 화자 레이블을 나타내며, 0 또는 1의 값을 갖는 C차원의 T개의 프레임으로 구성되어 있다.

2.2 화자 분할 모델

종단 간 화자 분할 모델은 SA-EEND[8]를 베이스라인 시스템으로 사용하였다. Fig. 1은 SA-EEND 모델 구조를 나타낸다. 모델은 선형 계층과 p개의 인코더 블록으로 구성되어 있으며 다음과 같이 수행된다.

$$E_0 = Norm(Linear(X)) \in \mathrm{R}^{T \times D}. \tag{5}$$

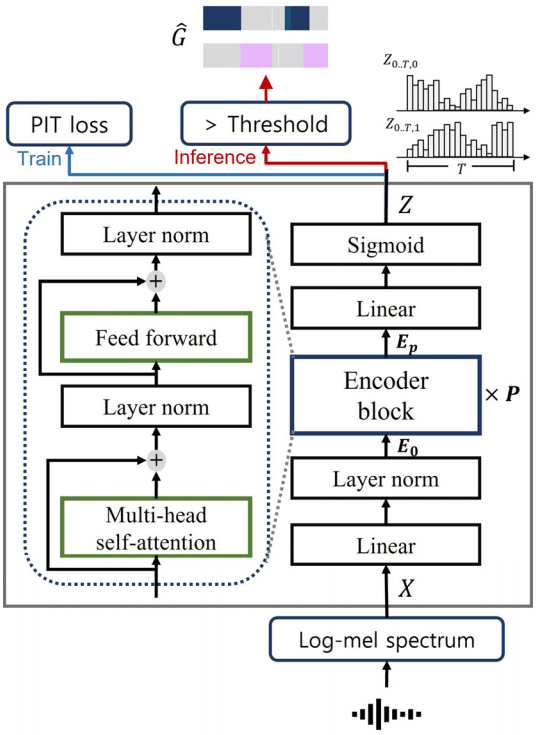

Fig. 1. (Color available online) The structure of SA-EEND model.[8]

$$E_p = Encoder_p(E_{p-1}) \in \mathrm{R}^{T \times D}. \tag{6}$$

$$Z = \sigma(Linear(E_p)) \in \mathrm{R}^{T \times C}. \tag{7}$$

$Linear$는 선형 계층을 나타내며, $Norm$은 계층 정규화를 나타낸다. p번째 인코더 블록은 $Encoder_p$로 나타내며 E_p는 인코더 블록을 수행한 결과이다. Fig. 1과 같이 인코더 블록 후의 마지막 선형 계층에는 시그모이드 함수를 적용하였고, 그 외 계층에서는 Rectified Linear Unit(ReLU) 함수를 적용하였다. Fig. 1에서 ReLU 함수는 생략하였다. Eq. (7)에서 σ는 시그모이드 활성화 함수를 나타낸다. 각 인코더 블록은 멀티-헤드 셀프-어텐션과 피드 포워드 네트워크로 구성되어 있으며 SA-EEND와 동일하게 수행하였다.

2.3 PIT 손실함수[7,8]

모델 훈련을 위해 추정된 Z와 정답 레이블을 이용해 손실함수를 계산한다. 정답 레이블 G는 다음과 같이 구성되어 있다.

$$G=[g_1,\dots,g_t,\dots,g_{\mathrm{T}}],\ g_t\in\{0,1\}^{\mathrm{C}}. \quad (8)$$

Eq. (8)에서 G는 0 또는 1의 값을 갖는 C차원의 T개 프레임으로 구성되어 있다.

화자 분할 모델은 예측된 화자들의 레이블 중 첫 번째 화자 레이블이 정답 레이블의 첫 번째 화자인지, 두 번째 화자인지 알 수 없는 화자 레이블 모호성 문제를 가지고 있다. 이를 해결하기 위해 PIT 손실함수가 제안되었다.[6] PIT 손실함수는 다음과 같이 계산한다.

$$L_{PIT}=\frac{1}{\mathrm{TC}}\min_{\phi\in perm(\mathrm{C})}\sum_{t=1}^{\mathrm{T}}H(g_t^{\phi},z_t). \quad (9)$$

Eq. (9)에서 $H(\bullet,\bullet)$는 이진 교차 엔트로피 손실함수이고, $perm(c)$는 화자들의 모든 순열에 대한 집합이다. PIT 손실함수는 정답 레이블의 모든 화자 순열을 고려하여 손실함수를 계산한다. 정답 레이블의 화자 순열은 계산한 손실함수가 최소일 때의 사용된 순열을 따른다.

III. 단일 레이블 분류 기반 화자 분할

3.1 단일 레이블 변환[6,7]

현재 연구되는 대부분의 종단 간 화자 분할 시스템은 2.1에서 설명하는 과정과 같이 모델을 통해 프레임별로 발화하는 화자 레이블 사후 확률을 추정하고 임계값보다 클 때 해당 화자가 발화했음을 결정한다. 이러한 화자 분할 시스템의 성능은 임계값에 상당히 의존적이다. 임계값을 어떤 값으로 결정하느냐에 따라 화자 분할의 성능이 크게 달라지며, 모델의 최적 임계값을 설정했을 때 특정 음성에는 정확한 화자 분할 결과를 얻을 수 있지만 다른 음성 종류나 상태에서는 정확하지 않은 분할 결과를 얻을 수 있다.

본 논문에서는 화자 분할 시스템의 임계값 의존 문제를 해결하고 모델을 통해 분할 결과를 바로 얻을 수 있도록 단일 레이블 분류를 이용한 모델에 대해 연구를 수행하였다. 기존 화자 분할 모델은 화자 분할을 다중 레이블 분류 문제로 설정하게 되어 임계값이 필요하다. 다중 레이블 분류를 단일 레이블 분류 문제로 변환한다면 임계값 의존을 해결할 수 있다. 다중 레이블을 단일 레이블로 변환하기 위해 다음과 같이 멱집합을 이용한다. 본 논문에서 화자 수는 2명으로 설정하여 실험을 진행하였다.

$$S=\{spk_1,\ spk_2\}. \quad (10)$$

$$\mathscr{P}(S)=\{\{\varnothing\},\{spk_1\},\{spk_2\},\{spk_1,spk_2\}\}. \quad (11)$$

Eqs. (10)과 (11)에서 S는 화자들의 집합이며, $\mathscr{P}$(S)는 멱집합을 나타낸다. 이를 이용해 단일 레이블로 변환할 수 있다.[6,7] 다중 레이블 분류를 기반으로 화자 분할을 수행하는 SA-EEND 모델은 (화자 1, 화자 2) 레이블을 갖지만, 단일 레이블로 변환하면서 (묵음, 화자 1만 발화, 화자 2만 발화, 동시 발화) 레이블을 갖는다. spk_1과 spk_2는 각각 화자 1과 화자 2일 때, $\varnothing$는 어떤 화자도 포함되지 않은 묵음 집합을 나타낸다.

모델 훈련을 위해 정답 레이블 G는 Eqs. (12)와 (13)과 같이 계산하여 예측된 단일 레이블에 대한 정답으로 변환할 수 있다.

$$G'=[g_1',\dots,g_t',\dots,g_{\mathrm{T}}'],\ g_t'\in\{0,1\}^{2^{\mathrm{C}}}. \quad (12)$$

$$g_t'=onehot(\Sigma_{c=1}^{(C)}(g_{t,c}\bullet 2^{(c-1)})). \quad (13)$$

G'은 기존 정답 레이블을 단일 레이블로 변환한 것으로 기존 C차원에서 2^{C}차원으로 레이블 길이가 확장된다. $onehot$은 원-핫 인코딩을 수행하는 함수이며 g_t'는 t번째 프레임의 레이블로 첫 번째 화자만 발화한 $g_t=[1,0]$인 경우, [0, 1, 0, 0]으로 표현된다. 해당 과정은 모델 훈련 중 손실함수를 계산하는 과정에서 수행되며, Fig. 2의 'Multi to Single'에서 수행된다.

화자 분할 모델에서는 변환된 단일 레이블을 예측하기 위해 마지막 선형 계층의 차원을 수정하였고, 시그모이드 활성화 함수를 소프트맥스 활성화 함수로 변경하였다.

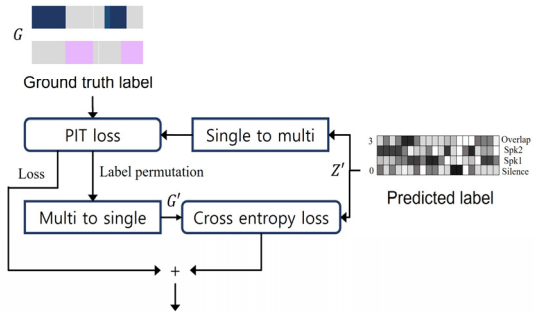

Fig. 2. (Color available online) The process of combined loss calculation.

$$Z' = Softmax(Linear(E_p)) \in [0,1]^{T \times 2^C}. \quad (14)$$

본 논문에서는 단일 레이블을 이용한 SA-EEND 모델을 Single Label(SL)-SA-EEND 이라 부르기로 한다.

3.2 잔차 연결[9]

심층 신경망 구조를 갖는 모델은 결과를 예측하는데 상대적으로 낮은 층의 영향이 적다. 마찬가지로 SA-EEND 모델은 인코더의 층이 깊어질수록 예측된 결과가 출력과 가까운 인코더 블록에 의존하는 경향을 보였다.[9] 본 논문에서는 이러한 문제점을 보완하여 낮은 층의 인코더 블록도 효과적으로 학습하기 위해 인코더 블록들에 대해 잔차 연결을 추가하여 그 성능을 관찰하였다.

p개 인코더 블록을 연결해주기 위해서 인코더 블록의 값들을 더하는 방법과 그대로 연결해주는 방법 2가지를 이용하였다. 인코더의 값들의 정보를 통합하기 위해 더해주는 방법을 사용하였고, 또한 역전파가 낮은 층의 인코더 블록까지 잘 전달될 수 있도록 인코더 블록의 값을 그대로 연결해주는 방법을 이용하였다. 잔차 연결은 다음과 같이 수행된다.

$$E = Norm(Linear(Cat(E_1,..,E_P, (E_1+..+E_P)))). \quad (15)$$

Cat은 값들을 연결해주는 함수로, 인코더 블록들을 연결한 후에는 차원을 조정해주기 위해 선형 계층을 추가하여 잔차 연결 블록을 구성하였다. 잔차 연결 블록을 처음부터 적용한다면 학습되지 않은 낮은 층의 인코더 블록에 의하여 오히려 학습이 안 될

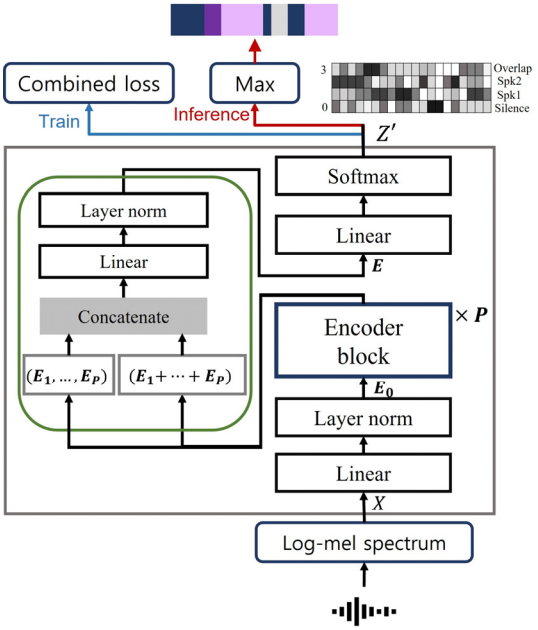

Fig. 3. (Color available online) The structure of the proposed SL-Res-SA- EEND model.

가능성을 고려하여, SL-SA-EEND 모델을 이용해 전이 학습을 수행하였다.

3.3 SL-Res-SA-EEND(SL-Residual-SA-EEND)

본 논문에서는 제안하는 잔차 연결 블록을 추가한 단일 레이블 분류 기반 화자 분할 모델인 SL-Res-SA-EEND의 구조를 Fig. 3에서 볼 수 있다. 전체적인 구조는 SA-EEND 모델과 유사하나 단일 레이블 예측을 위해 마지막 선형 계층과 활성화 함수를 변경하였고, p개의 인코더 블록 다음으로 잔차 연결 블록을 추가하였다.

단일 레이블 분류 문제로 변경하여 화자 분할을 진행하기 때문에 테스트 단계에서는 임계값 없이 예측된 결과의 최대값을 이용해 각 프레임의 레이블을 결정할 수 있다.

3.4 손실함수 조합

본 논문에서는 단일 레이블 분류 기반의 시스템 SL-SA-EEND 모델과 SL-Res-SA-EEND 모델의 학습을 위해 PIT 손실함수와 교차 엔트로피 손실함수를 조합하여 사용하였다. 예측된 단일 레이블에 대한

손실을 계산하기 위해 교차 엔트로피 손실함수를 사용하고 동시에, 정답 레이블의 화자 순열을 고려하기 위해 PIT 손실함수를 이용하였다.

Fig. 2는 전체 손실함수 계산과정을 나타낸다. 그림에서 'Multi to single'은 기존 정답 레이블을 단일 레이블로 변환하는 과정으로 Eq. (13)을 이용해 수행되며, 'Single to multi'는 예측된 단일 레이블을 다중 레이블 형태로 변환하는 과정으로 Eq. (13)의 과정을 반대로 수행한다.

먼저, 정답 레이블의 순열을 결정하기 위해 모델을 통해 예측된 단일 레이블에 대한 확률을 다중 레이블 형태로 변환한 후 PIT 손실함수를 계산한다. PIT 손실함수를 통해 결정된 정답 레이블 순열은 단일 레이블 형태로 변환하여 교차 엔트로피 손실함수 계산에 사용한다. 그 후 PIT 손실함수와 교차 엔트로피 손실함수 값을 더해서 최종 모델의 손실함수로 사용하였다. 교차 엔트로피 손실함수는 다음과 같이 계산하였다.

$$L_{CE} = \frac{1}{\mathrm{T} \cdot 2^{\mathrm{C}}} \sum_{t=1}^{\mathrm{T}} \sum_{c=1}^{2^{\mathrm{C}}} -(g_{t,c}^{\phi})' \log(z_{t,c}'). \tag{16}$$

$(g_t^{\phi})'$은 t번째 프레임에서 PIT 손실함수를 통해 결정된 정답 레이블 순열을 Eq. (13)을 이용해 단일 레이블로 변환한 값이며 z_t'는 SL-SA-EEND 모델 또는 SL-Res-SA-EEND 모델에서 추정된 t번째 프레임의 레이블을 나타낸다.

제안하는 화자 분할 모델의 전체 손실함수는 다음과 같이 합하여 계산하였다.

$$L = L_{PIT} + L_{CE}. \tag{17}$$

IV. 실험 및 결과

4.1 데이터베이스

2명의 화자가 존재하는 잡음 음성 데이터를 생성하기 위해 깨끗한 음성 샘플로 100 h 길이의 Librispeech 데이터베이스를 이용하였다.[10] 배경 잡음으로는 A Music, Speech and Noise(MUSAN) 데이터베이스[11] 중 배경 잡음을 사용하였고, Room Impulse Response(RIR) 필터는 Simulated Room Impulse Responses 데이터베이스의 10,000개 필터를 이용하였다.[12] 잡음 데이터 생성을 위해 Signal-to- Noise Ratio(SNR) 조건으로 (5 dB, 10 dB, 15 dB, 20 dB) 중에 무작위로 선택하여 설정하였으며, RIR 필터는 생성한 잡음 데이터 중에서 50 %의 확률로 무작위 선택하여 선택된 음성에만 적용하였다.

2명의 화자 음성을 혼합하기 위해서 기존 SA-EEND에서 제시한 과정을 그대로 이용하였으며, 화자 음성의 중첩 구간 비율이 약 34 % 정도가 되도록 설정하여 생성하였다. 또한 각 화자의 발화 수는 최소 5개에서 최대 10개까지 선택될 수 있도록 하였다. SA-EEND 모델의 실험과 동일하게 진행하기 위해 생성된 훈련 데이터는 100,000개, 검증 및 테스트 데이터 개수는 각각 500개로 설정하였다. 테스트 데이터 길이는 약 24 h 정도로 화자 분할을 위한 실제 음성 데이터베이스들과 유사하거나 더 긴 길이이다.[13,14] 또한 테스트 데이터의 화자들은 훈련, 검증 데이터와 겹치지 않게 구성하였다.

4.2 실험 설정 및 성능 평가 지표

음성의 특징 추출을 위해 창 크기 및 홉 크기는 각각 25 ms, 10 ms로 설정하였고, 80차원의 로그 멜 필터뱅크 에너지를 사용하였다. 생성된 입력 특징은 프레임마다 앞뒤 7프레임씩 값을 연결하였고, 10개 프레임마다 1개 프레임을 선택하여 사용하도록 10씩 서브-샘플링을 수행하였다.

SA-EEND 기반의 모델은 모두 4개의 인코더 블록을 사용하였고, 인코더 블록의 어텐션은 256차원을 가지며 헤드의 개수는 4개로 설정하였다. 또한 피드포워드 네트워크의 유닛 개수는 1,024개로 설정하여 실험을 수행하였다. 최적화 알고리즘은 25,000 warm-up steps에 의해 학습률이 조정되는 'Adam' 알고리즘을 사용하였다.[15] 최종 모델은 마지막 10개 모델의 파라미터를 평균 내어 생성하였다.

화자 분할의 성능은 Diarization Error Rate(DER)로 평가하였다[16]. DER은 음성이 있는 구간이지만 음성이 없다고 판단한 구간에 대한 Miss error rate(MISS)와 음성이 없는 구간을 음성이 있다고 판단한 구간에 대한 False alarm error rate(FA), 화자를 혼동한 구간에 대한 Speaker confusion error rate(SPK)의 합으로

계산된다. 분할된 결과에 중간값 필터를 적용하고 칼라 허용 오차는 0.25 s로 설정한 후 DER 평가를 진행하였다.

성능 비교를 위해 기존의 대표적인 클러스터링 기반 화자 분할 기법 중 하나인 pyannote 2.0에 대한 성능 평가를 실시하였다.[17,18] 또한 제안한 모델의 성능 평가를 위해 기존 SA-EEND와 임계값 의존 문제에 대해 해결 방법을 제안했었던 SEND[6]와의 비교를 진행했다. SEND 모델의 단일 레이블 분류 기반 방법만 비교하기 위해 제안되었던 화자 특징 학습 네트워크 없이 다중 레이블을 예측한 후 단일 레이블로 변경하는 방법만 적용하여 화자 분할을 진행하였다.

4.3 실험 결과

기존 SA-EEND 모델과 같이 다중 레이블 분류 문제로 화자 분할을 수행했을 때 임계값에 따른 성능 변화를 관찰하기 위해 Table 1과 같이 실험을 수행하였다. 모델의 성능이 임계값에 따라 최대 1.7 %p만큼 차이가 나는 것을 확인할 수 있으며, 데이터 종류에 따라 최적의 결과를 보여줄 수 있는 임계값이 다를 것으로 예상된다.

Table 2에서는 베이스라인 시스템들과 본 논문에서 제안하는 모델의 성능 평가 결과를 볼 수 있다. pyannote는 클러스터링 기반의 화자 분할 시스템으로서 본 논문에서 사용한 잡음 음성 데이터에 대한 화자 분할 성능이 많이 떨어지는 것으로 보인다. SEND 모델은 추정된 다중 레이블을 단일 레이블로 변환하여 임계값 문제를 해결했지만, 화자 특징 학습 네트워크 없이 훈련을 진행했을 때는 기존 SA-EEND 모델보다도 DER의 성능이 떨어지는 것을 확인할 수 있었다.

Table 1. The DER performances of SA-EEND model according to threshold (%).

Threshold	MISS	FA	SPK	DER
0.3	0.2	3.2	1.8	5.2
0.4	0.3	2.1	1.9	4.3
0.5	0.4	1.4	1.9	3.8
0.6	**0.6**	**0.9**	**2.0**	**3.5**
0.7	1.0	0.6	1.9	3.6

Table 2. The DER performances of baseline system and proposed model (%).

Model	MISS	FA	SPK	DER
Pyannote	18.7	0.1	0.5	19.3
SA-EEND	0.6	0.9	2.0	3.5
SEND	0.5	1.1	2.5	4.1
SL-SA-EEND	**0.5**	**0.8**	**1.8**	**3.1**
SL-Res-SA-EEND	**0.4**	**0.7**	**1.8**	**2.9**

본 논문에서 제안하는 SL-SA-EEND 모델의 결과는 기존 시스템인 SA-EEND 모델과 비교했을 때, 분할 성능이 향상된 결과를 보여준다. 제안하는 단일 레이블 분류 기반의 화자 분할 방법이 임계값 없이 더 나은 화자 분할을 수행할 수 있음을 보여준다. 특히 화자 혼동 에러 비율인 SPK에서 2.0에서 1.8로 감소한 것으로 보아 제안하는 방법이 화자 혼동에도 효과적임을 보여준다. 잔차 구조를 적용한 SL-Res-SA-EEND 모델이 가장 높은 성능을 보이며 기존 SA-EEND 모델보다 DER이 약 20.7 %만큼 향상되었다. 이를 통해 SL-Res-SA-EEND 모델이 낮은 층의 인코더 학습에 효과적인 영향을 끼친 것을 확인할 수 있었다.

V. 결 론

본 논문에서는 다중 레이블 분류 문제로 화자 분할을 수행하면서 발생하는 임계값 의존 문제를 해결하기 위해 단일 레이블 분류 기반의 화자 분할 시스템에 대해 연구를 진행했다. 또한 낮은 층의 인코더 블록의 효과적인 학습을 위해 추가적인 잔차 연결 구조를 구성하여 전이 학습을 수행하였다. 그 결과, 2명의 화자가 존재하는 잡음 음성 데이터에 대해 기존 베이스라인 모델과 비교하여 향상된 분할 성능을 보여주었으며, 잔차 연결 구조를 추가하였을 때 가장 높은 분할 성능을 보였다. 향후에는 2명 이상의 화자가 존재하는 음성에 대해서도 단일 레이블 분류 기법을 적용하여 최적의 임계값을 찾는 과정 없이 화자 분할을 수행하는 방안에 관해서 연구하고자 한다.

감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2021R1F1A1063347).

References

1. D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, “Speaker diarization using deep neural network embeddings,” Proc. ICASSP, 4930-4934 (2017).
2. Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, “Speaker diarization with LSTM,” Proc. ICASSP, 5239-5243 (2018).
3. M. Diez, L. Burget, S. Wang, J. Rohdin, and H. Cernocký, “Bayesian HMM based x-Vector clustering for speaker diarization,” Proc. Interspeech, 346-350 (2019).
4. I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, “Target-speaker voice activity detection: a novel approach for multispeaker diarization in a dinner party scenario,” Proc. Interspeech, 274-278 (2020).
5. Y. C. Liu, E. Han, C. Lee, and A. Stolcke, “End-to-end neural diarization: From transformer to conformer,” Proc. Interspeech, 3081-3085 (2021).
6. Z. Du, S. Zhang, S. Zheng, and Z. Yan, “Speaker embedding-aware neural diarization: A novel framework for overlapping speech diarization in the meeting scenario,” arXiv preprint arXiv:2203.09767 (2022).
7. Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, “End-to-end neural speaker diarization with permutation-free objectives,” Proc. Interspeech, 4300-4304 (2019).
8. Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, “End-to-end neural speaker diarization with self-attention,” Proc. ASRU, 296-303 (2019).
9. Y. Yu, D. Park, and H. K. Kim, “Auxiliary loss of transformer with residual connection for end-to-end speaker diarization,” Proc. ICASSP, 8377-8381 (2022).
10. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” Proc. ICASSP, 5206-5210 (2015).
11. D. Snyder, G. Chen, and D. Povey, “Musan: A music, speech, and noise corpus,” arXiv preprint arXiv:1510.08484 (2015).
12. T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” Proc. ICASSP, 5220-5224 (2017).
13. J. Carletta, “Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus,” Lang. Resour. Eval. **41**, 181-190 (2007).
14. A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, “The ICSI meeting corpus,” Proc. ICASSP, 364-367 (2003).
15. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Proc. NIPS, 5998-6008 (2017).
16. J. G. Fiscus, J. Ajot, and J. S. Garofolo, *The Rich Transcription 2007 Meeting Recognition Evaluation* (Springer, Maryland, 2007), pp. 373-389.
17. H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W, Bouaziz, and M. P. Gill, “Pyannote. audio: Neural building blocks for speaker diarization,” Proc. ICASSP, 7124-7128 (2020).
18. H. Bredin and A. Laurent, “End-to-end speaker segmentation for overlap-aware resegmentation,” Proc. Interspeech, 3111-3115 (2021).

저자 약력

▸ 정 재 희 (Jaehee Jung)

2021년 2월 : 인천대학교 컴퓨터공학부 공학사
2021년 3월 ~ 현재 : 인천대학교 컴퓨터 공학과 석사과정

▸ 김 우 일 (Wooil Kim)

1996년 2월, 1998년 8월, 2003년 8월 : 고려대학교 전자공학과 학/석/박사
2004년 8월 ~ 2005년 8월 : Carnegie Mellon University 박사후연구원
2005년 8월 ~ 2012년 8월 : University of Texas at Dallas 연구원, 연구교수
2012년 8월 ~ 현재 : 인천대학교 컴퓨터공학부 조교수, 부교수, 교수