Transformer 기반 화자 분할을 위한 도착 순서 기반 Sort Loss의 특성 분석

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

Analysis of Arrival-Order-Based Sort Loss in Transformer-Based Speaker Diarization

Chanyeong Moon, Beom Jun Woo and Nam Soo Kim
Department of Electrical and Computer Engineering and INMC, Seoul National Univ

요 약

여러 화자가 동시에 발화하는 환경에서 각 화자가 언제 발화하고 있는지를 추정하는 화자 분할(speaker diarization) 분야에서는, 최근 여러 분야를 이어붙인 파이프라인 처리 기반의 다단계 방식에서 벗어나 하나의 신경망을 통해 화자 분할을 직접 학습하는 End-to-End Neural Diarization(EEND) 모델에 대한 연구가 활발히 이루어지고 있다. 기존 EEND 계열의 연구들은 화자 레이블 간 순열 문제를 해결하기 위해, 신경망이 화자 간 대응 관계를 자동으로 결정하도록 하는 permutation-invariant 학습 방식을 주로 사용해왔다. 한편, 최근에는 화자의 도착 순서(arrival order)를 기준으로 화자 레이블의 순서를 고정적으로 정렬하여 학습하는 Sortformer가 제안되었으며, 기존 EEND 방법들에 비해 우수한 성능을 보이며 주목받고 있다. 이러한 접근은 permutation을 고려하지 않는 대신, 정적인 화자 순서를 도입한다는 점에서 기존 방식과 근본적으로 다른 가정을 내포하고 있다. 이에 따라, arrival order 기반의 정렬 방식이 다양한 모델 구조에서도 일반적으로 유효한지에 대한 체계적인 고찰이 필요하다. 본 논문에서는 Transformer와 Conformer 구조를 각각 화자 분할 모델로 사용하여, arrival order 기반 sort loss가 모델 구조에 따라 어떠한 영향을 미치는지를 비교·분석하였다.

Ⅰ. 서 론

화자 분할은 다수의 화자가 동시에 발화하는 음성 신호로부터 각 화자의 발화 구간을 식별하는 문제로, 회의 기록 자동화, 방송 음성 분석, 음성 인식 전처리 등 다양한 응용에서 핵심적인 역할을 수행한다. 전통적인 화자 분할 시스템은 음성 활동 검출 (Voice Activity Detection), 화자 인식 (Speaker Recognition), 군집화 (Clustering) 와 같은 개별 구성 요소를 순차적으로 결합하는 방식으로 설계되어 왔다. 이러한 구조는 단계별 최적화가 가능하다는 장점이 있으나, 시스템 복잡도가 높고 각 단계의 오류가 누적될 수 있다는 한계를 가진다. 이러한 한계를 보완하기 위해, 최근에는 단일 신경망을 통해 화자 분할을 직접 학습하는 End-to-End Neural Diarization(EEND) 접근이 주목받고 있다. EEND 모델은 프레임 단위의 화자마다 발화 구간을 직접 예측함으로써 후처리 의존도를 낮추지만, 출력 화자 레이블의 순서가 임의적으로 바뀔 수 있는 순열 문제를 본질적으로 포함한다. 기존 연구들은 이를 해결하기 위해 주로 permutation-invariant 학습 방식을 채택해왔으며, 이는 화자 순서에 대한 사전 가정이 필요 없다는 장점을 제공한다. 최근 제안된 Sortformer는 화자의 위치 정보를 나타내는 positional embedding을 추가적으로 입력으로 사용하여, 화자 도착 순서를 기준으로 화자 레이블의 순서를 고정한 학습 방식을 도입하였다. 이를 통해 화자 레이블 간 순열 문제를 명시적으로 고려하지 않는 대안을 제시하였으며, 비교적 단순한 학습 구조에도 불구하고 경쟁력 있는 성능을 보여 주목받고 있다. 그러나 이러한 접근은 정적인 화자 순서를 가정한다는 점에서, 모델 구조에 따라 서로 다른 영향을 미칠 가능성을 내포한다. 특히 arrival order 기반 정렬 방식이 다양한 신경망 구조 전반에 걸쳐 안정적인 성능을 보장하는지에 대해서는 아직 충분한 분석이 이루어지지 않았다. 본 논문에서는 EEND-EDA 모델의 인코더로 사용된 Transformer 구조와 함께 Conformer 구조를 적용하여, arrival order 기반 sort loss가 모델 구조에 따라 화자 분할 성능과 학습 특성에 미치는 영향을 비교·분석한다. 이를 통해 정적 화자 순서를 가정한 loss 설계가 서로 다른 모델 구조에서도 보편적으로 적용될 수 있는지에 대한 실험적 근거를 제시하고자 한다.

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

최근 화자 분할 분야에서는 여러 처리 단계를 순차적으로 결합하는 기존 방식보다, 하나의 신경망을 통해 화자 발화 구간을 직접 추정하는 종단형 학습 접근이 주된 연구 흐름으로 자리 잡고 있다. 이 가운데 EEND(End-to-End Neural Diarization)는 입력 음성으로부터 화자별 활동 정보를 직접 예측하는 구조로, 다양한 화자 분할 연구의 기반 모델로 활용되고 있다 [1]. 다만, EEND는 출력 구조가 고정되어 있어 화자 수가 변하는 실제 대화 환경에서는 적용에 제약이 따른다. 이를 보완하기 위해 제안된 EEND-EDA(End-to-End Neural Diarization with Encoder-Decoder Attractor)는 화자 수를 사전에 가정하지 않고, 입력 음성으로부터 화자 표현을 요약하는 attractor를 생성하는 방식으로 모델을 확장하였다 [2]. 이러한 attractor 기반 접근은 가변 화자 수 환경에서도 유연하게 동작할 수 있도록 하여, 종단형 화자 분할 모델의 활용 범위를 넓혔다.

Ⅲ. 화자 도착 시간 순서를 통한 화자 순서 고정 (Arrival-time-order-based speaker ordering)

종단형 화자 분할 모델은 입력 음성으로부터 각 화자의 발화 여부를 프레임 단위로 동시에 추정한다. 이때 출력 화자 레이블의 순서는 임의적이며, 동일한 화자 분할 결과라도 화자 인덱스의 순열에 따라 서로 다른 출력 표현이 존재할 수 있다. 이러한 순열 문제(permutation problem)를 해결하기 위해 기존 EEND 계열 모델들은 permutation-invariant training (PIT)을 사용해왔다.

프레임 수를 T, 화자 수를 K라 하고, 정답 화자 분할 행렬을

$Y \in \{0,1\}^{K\times T}$, 모델 예측 결과를 $P \in [0,1]^{K\times T}$라 하면, PIT 기반 손실은 다음과 같이 정의된다.

$$L_{PIT}(Y,P) = \min_{\pi\in\Pi}\frac{1}{KT}\sum_{k=1}^{K}\sum_{t=1}^{T} BCE(y_{\pi(k),t}, p_{k,t}).$$

위 수식의 Π는 화자 인덱스 {1, …, K}에 대한 모든 순열의 집합이며, BCE(·)는 binary cross-entropy 손실 함수이다. PIT는 화자 순서에 대한 사전 가정을 필요로 하지 않는다는 장점이 있으나, 모든 순열을 고려해야 하므로 계산 복잡도가 증가하고, 다른 태스크와의 결합이 어렵다는 한계를 가진다.

EEND-EDA는 기존 EEND 구조를 확장하여, 화자 수가 사전에 주어지지 않은 환경에서도 화자 분할을 수행할 수 있도록 attractor 기반 구조를 도입하였. 프레임별 임베딩 $\{e_t\}_{t=1}^{T}$ EDA 모듈은 sequence-to-sequence 방식으로 화자별 attractor $\{a_k\}_{k=1}^{K}$를 생성하고, 각 프레임에서의 화자 발화 확률은 다음과 같이 계산된다.

$$p_{k,t} = \sigma(a_k^T e_t).$$

여기서 $\sigma(\,)$는 sigmoid 함수이다. EEND-EDA 역시 화자 attractor의 순서가 사전에 정의되지 않기 때문에, 학습 시 PIT 기반 손실을 사용하여 화자 순열 문제를 해결한다. 즉, EEND-EDA는 구조적으로 화자 순서를 고정하지 않으며, 순열 불변성을 손실 함수 수준에서 처리한다는 특징을 가진다.

Sortformer[3]는 기존 EEND 계열과 달리, positional embedding을 추가적인 입력으로 활용하여, 화자 순서를 학습 과정에서 명시적으로 고정하는 접근을 취한다. 이를 위해 각 화자의 첫 발화 시점(arrival time)을 기준으로 화자 순서를 정의한다. 따라서 PIT loss와 같이 화자의 순서가 바뀌었을 때를 정의하지 않아도 되므로 고정된 순서의 라벨이 주어지고 이에 따라 loss를 흘려주면 된다. 이를 수식으로 표현하면 다음과 같다.

$$L_{sort}(Y,P) = \frac{1}{KT}\sum_{k=1}^{K}\sum_{t=1}^{T} BCE(y_{k,t}^{fixed}, p_{k,t}).$$

이 때, $y_{k,t}^{fixed}$의 경우 이미 순서가 배정된 k번째 화자에 대한 t번째 프레임에 대한 값을 의미한다. Sort Loss는 화자 순열에 대한 최소화 연산을 포함하지 않으며, 모델이 출력 단계에서 화자 채널을 도착 순서에 맞게 정렬하도록 유도한다.

이러한 차이는 모델 구조와의 상호작용 측면에서 중요한 영향을 미칠 수 있다. 특히 화자 순서를 암묵적으로 학습해야 하는 Transformer 기반 구조에서는, arrival-order 기반 정렬 가정이 학습 안정성 및 성능에 영향을 줄 가능성이 있다. 이에 본 논문에서는 EEND-EDA 구조를 기반으로 Transformer 및 Conformer 인코더를 각각 적용하여, arrival-order 기반 sort loss의 효과를 모델 구조별로 비교·분석한다.

IV. 실험

본 실험에서는 arrival-order 기반 sort loss가 모델 구조에 따라 화자 분할 성능에 미치는 영향을 분석하기 위해, EEND-EDA 모델을 기반으로 인코더 구조만을 변경하여 실험을 진행하였다. 인코더로는 기존 EEND-EDA에서 사용된 Transformer 구조와, 국소적 문맥 정보를 함께 활용하는 Conformer 구조를 각각 적용하였다. 각 모델은 화자 인식에서 주로 사용되는 simulation 데이터셋인 sim2spk를 사용하여 학습하였으며, 손실 함수로는 PIT loss와 sort loss를 각각 적용하여 비교하였다.

평가는 diarization error rate(DER)를 기준으로 수행하였으며, DER는 miss rate, false alarm, speaker confusion의 합으로 계산된다. 평가는 손실 함수와 모델 구조에 따른 차이만을 비교할 수 있도록 하였다.

Model	Loss	DER	MISS	False Alarm	Speaker Confusion
Trasnformer (Baseline)	PIT Loss (Baseline)	4.89	2.24	2.40	0.24
	Sort Loss	28.97	3.73	23.45	1.79
Conformer	**PIT Loss**	**3.14**	**1.21**	**1.71**	**0.22**
	Sort Loss	4.46	1.53	2.60	0.33

〔표 1〕 실험 결과 (DER〔%〕)

표 1의 실험 결과에서 전반적으로 PIT loss가 sort loss 대비 우수한 성능을 보임을 확인할 수 있다. 특히 Transformer 인코더를 사용하여 sort loss를 적용한 경우 DER가 급격히 증가하는 현상이 관찰되었으며, 이는 화자 도착 순서에 따라 화자 순서를 고정하는 방식이 Transformer 기반 모델의 성능 열화에 직접적인 영향을 미칠 수 있음을 시사한다. 이러한 현상은 Transformer가 positional encoding에 주로 의존하여 순서 정보를 표현하는 구조적 특성으로 인해, 국소적인 시간 패턴에 대한 inductive bias를 충분히 내재화하기 어렵기 때문에 발생한 것으로 해석할 수 있다. 반면, Sortformer는 Conformer 인코더를 사용함으로써 이러한 한계를 완화하였으며, 본 실험 결과와도 일관된 경향을 보인다. 다만, Sortformer가 EEND-EDA 기반 모델보다 우수한 성능을 보이는 것은 sort loss 자체의 효과라기보다는 사전 학습된 feature extractor와 추가적인 파라미터 사용에 따른 영향으로 해석하는 것이 타당하다.

V. 결론

본 논문에서는 arrival-order 기반 sort loss가 모델 구조에 따라 상이한 성능을 보임을 확인하였다. 실험 결과, Transformer 인코더에서는 성능 저하가 두드러진 반면, Conformer 기반 모델에서는 비교적 안정적인 결과를 보였다. 이는 정적 화자 순서를 가정한 loss 설계가 인코더의 구조적 특성에 크게 의존함을 시사한다.

ACKNOWLEDGMENT

이 논문은 2026년도 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] Park, T., Medennikov, I., Dhawan, K., Wang, W., Huang, H., Koluguri, N. R., ... & Ginsburg, B. Sortformer: A Novel Approach for Permutation-Resolved Speaker Supervision in Speech-to-Text Systems. In Forty-second International Conference on Machine Learning.