발화 길이 기반 가중 평균을 이용한 화자 임베딩 등록 방식

김수경, 이건희
HDC Labs

kimsk5818@hdc-labs.com, Gunhee_Lee@hdc-labs.com

Speaker Enrollment Embeddings Using Duration-Based Weighted Averaging

SuKyung Kim, Gunhee Lee
HDC Labs

요 약

짧은 발화로부터 추출된 화자 임베딩은 제한된 음소 정보로 인해 불확실성이 증가하며, 이를 기존의 단순 평균 방식으로 결합할 경우, 등록 임베딩의 신뢰도가 저하될 수 있다. 이러한 문제를 해결하기 위해, 본 논문에서는 발화 길이 기반 가중 평균 등록 방식을 제안한다. 실험 결과, 제안한 방법은 다양한 등록 발화 길이 조건에서 기존 단순 평균 방식 대비 일관된 성능 향상을 보였으며, 특히 짧은 발화가 다수 포함된 조건에서 가장 큰 개선 효과를 확인했다.

Ⅰ. 서 론

화자 인식 시스템은 음성 기반 보안 인증, 개인화 음성 서비스 등 다양한 상용 서비스에 활용되고 있다. 이러한 시스템은 사용자의 음성을 사전에 등록하는 등록 과정을 통해 각 화자의 기준 음성 표현을 확보한 이후, 입력 음성과 등록 정보를 비교하여 화자를 식별한다. 이때 등록 단계에서 생성되는 화자 임베딩 벡터는 이후 모든 인식 과정의 기준이 되므로, 등록 임베딩의 품질은 전체 시스템 성능에 직접적인 영향을 미친다.

화자 임베딩 벡터 추출에 있어서 등록 발화의 길이는 화자 정보를 얼마나 충분히 관측할 수 있는지를 결정하는 정보의 양으로 해석할 수 있으며, 이는 추출된 임베딩 벡터의 품질과 밀접한 관련이 있다. 일반적으로 등록 발화의 길이가 길수록 음색, 조음 특성, 발화 습관 등 화자 고유의 특성이 더 충분히 반영될 수 있어, 보다 정확하고 안정적인 임베딩 벡터를 추출할 수 있다. 반면, 발화 길이가 짧은 경우에는 관측 가능한 음소 정보가 제한되며, 이로 인해 임베딩 추정의 신뢰도가 저하될 수 있다.

기존 연구들에서도 발화 길이와 화자 인식 성능 간의 관계가 지속적으로 분석되어 왔다. [1]에서는 사전에 학습된 화자 임베딩이 화자 정보뿐만 아니라 발화 길이와 같은 발화 속성 정보를 일부 포함하고 있으며, 발화 길이에 따라 임베딩 공간상의 분포 특성이 달라질 수 있음을 실험적으로 분석하였다. 또한 짧은 발화 조건에서 화자 인식 성능이 현저히 저하된다는 점은 여러 연구에서 공통적으로 보고되었다. 특히, [2]는 짧은 발화로부터 추출된 화자 임베딩은 제한된 음소 정보로 인해 불확실성이 증가하며, 이러한 불확실성이 화자 검증 시스템의 성능 저하를 유발하는 주요 요인임을 지적하였다. 이에 따라, [3]과 같이 짧은 발화 조건을 명시적으로 고려한 임베딩 학습 기법들이 제안되었다.

이처럼 발화 길이는 화자 임베딩의 안정성과 신뢰도에 직접적인 영향을 미치며, 이는 최종 화자 인식 성능으로 이어진다. 그럼에도 불구하고, 기존의 다중 발화 기반 화자 등록 방식에서는 일반적으로 모든 등록 발화를 동일한 가중치로 평균하여 하나의 등록 임베딩을 생성한다. 이러한 방식은 길이가 짧고 정보량이 상대적으로 적은 발화가 포함되더라도, 길이가 긴 발화와 동일한 중요도로 취급한다는 한계를 가진다. 그 결과, 불확실성이 큰 짧은 발화에서 추출된 임베딩이 등록 임베딩에 그대로 반영되어 전체 임베딩의 분산을 증가시키고, 화자 식별 성능을 저하시킬 가능성이 있다.

본 논문에서는 이러한 문제를 해결하기 위해, 다중 발화 등록 환경에서 발화 길이에 따라 차등적인 가중치를 부여하는 발화 길이 기반 가중 평균 등록 방식을 제안한다. 제안하는 방법은 길이가 긴 등록 발화에 더 높은 비중을 부여함으로써, 불안정한 짧은 발화의 영향을 효과적으로 줄이고, 보다 안정적이고 신뢰도 높은 화자 등록 임베딩을 생성하는 것을 목표로 한다.

Ⅱ. 본론

본 논문에서는 다중 발화 기반 화자 등록 과정에서 등록 발화 간 길이 차이로 인해 발생하는 임베딩 품질 저하 문제를 완화하기 위해, 발화 길이에 기반한

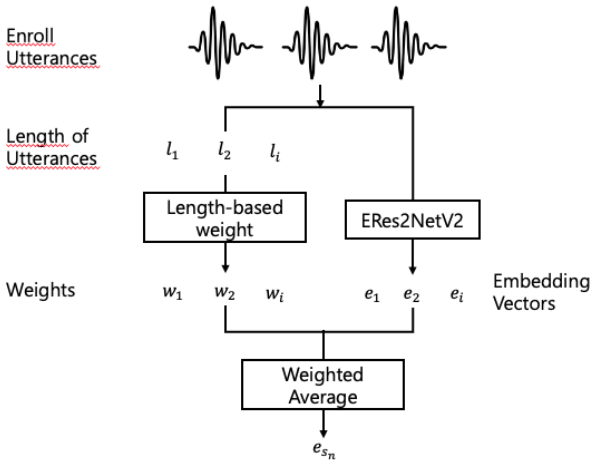

Figure1. 길이 기반 가중 평균 등록 프로세스

가중 평균 화자 등록 임베딩 생성 방법을 제안한다.

Figure 1 은 제안하는 다중 등록 발화로부터 최종 화자 등록 임베딩을 생성하는 과정을 단계별로 나타낸다. 주어진 등록 발화들을 입력으로, 각 발화의 길이 l_i 를 계산한다. 그리고 길이 기반 가중치를 식(1)과 같이, 계산하고 ERes2NetV2 모델을 사용해 각 등록 발화에서 추출한 임베딩 벡터들과 함께 식(2)와 같이 가중 평균하여 화자 별로 하나의 등록 임베딩 벡터 e_{s_i} 를 도출한다.

$$w_i = \frac{l_i}{L}, \left(L = \sum_{j=1}^{N} l_j\right) \cdots(1)$$

$$e_{s_i} = \sum_{j=1}^{N} w_j e_j \cdots(2)$$

Ⅲ. 실험

다중 발화 기반 화자 등록 환경에서 길이 기반 가중 평균이 성능에 미치는 영향을 확인하기 위해 서로 다른 길이 조합의 4 가지 조건에서 단순 평균 등록 방식과 길이 기반 가중 평균 했을 때의, 화자 인식 성능을 비교했다. 데이터는 [4]voxceleb 테스트셋을 사용했으며 인식에 사용된 화자는 총 40 명이다. 4 가지 등록 발화 길이 조건은 다음과 같다.: 조건 A [0.3, 0.4, 0.5, 0.6, 0.7]/조건 B [0.3,0.3,0.3,0.3,0.7] /조건 C[0.8, 0.9, 1.0, 1.1, 1.2] /조건 D[0.8, 0.8, 0.8, 0.8, 1.5]

[표 1]을 보면, 모든 조건에서 제안한 길이 기반 가중 평균 방식이 단순 평균 방식보다 더 높은 정확도를 보였으며, 성능 향상의 폭은 등록 발화의 길이 분포에 따라 다르게 나타났다. 등록 발화가 매우 짧고 하나의 긴 발화가 포함된 조건 B 의 경우, 가중 평균 방식이 긴 발화에 더 큰 비중을 부여함으로써 약 5.7%의 성능 향상을 보였다. 조건 C, D 의 경우 전체 발화 길이가 충분히 길어 단순 평균 방식도 비교적 기본적인 화자 정보가 확보되어 가중 평균 방식의 추가적인 성능 향상 폭이 작게 나타났다. 결과를 전반적으로 분석하자면 제안한 길이 기반 가중 평균 방식이 등록 발화 간 길이 편차가 클수록, 짧은 발화의 비중이 높을 수록 더 큰 성능 향상이 나타난다는 것을 알 수 있다.

	단순 평균	길이 기반 가중 평균
조건 A	79.38%	79.93%
조건 B	59.01%	64.70%
조건 C	85.73%	85.92%
조건 D	87.74%	87.80%

표 1. 등록 발화 길이에 따른 화자 분류 정확도

Ⅳ. 결론

본 논문에서는 짧은 발화로부터 추출된 화자 임베딩 벡터의 정확도 보완을 위한 길이 기반 가중 평균 등록 방식을 제안했다. 실험 결과, 제안한 방법은 다양한 등록 발화 길이 분포 조건에서 기존 단순 평균 방식 대비, 일관된 성능 향상을 보였으며, 특히 짧은 발화가 다수 포함된 조건에서 가장 큰 성능 개선을 확인했다. 이를 통해 등록 발화가 짧고, 길이 편차가 큰 실제 서비스 환경에서 제안한 방법이 효과적일 수 있음을 입증하였다.

참 고 문 헌

[1] WANG, Shuai; QIAN, Yanmin; YU, Kai. “What does the speaker embedding encode?”. arXiv preprint arXiv:2512.18286, 2025.

[2] Poddar, Arnab; SAHIDULLAH, Md; SAHA, Goutam. “Speaker verification with short utterances: a review of challenges, trends and opportunities”. IET Biometrics, 2018, 7.2: 91-101.

[3] Jee-won Jung, Hee-Soo Heo, Hye-jin Shim, and Ha-Jin Yu,

“Short utterance compensation in speaker verification via cosine-based teacher-student learning of speaker embeddings,”

Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019.

[4] Nagrani, Arsha, et al. "Voxceleb: Large-scale speaker verification in the wild." Computer Speech & Language 60, 2020.