문장 종속 화자 검증 시스템을 위한 음성인식기 기반 Pooling 기법

문성환, 강우현, 한민현, 김남수

서울대학교 전기정보공학부 뉴미디어통신공동연구소

{shmun, whkang, mhhan}@hi.snu.ac.kr, nkim@snu.ac.kr

ASR-based pooling method for text-dependent speaker verification

Sung Hwan Mun, Woo Hyun Kang, Min Hyun Han and Nam Soo Kim

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

요 약

시스템에서 특정 문장만을 고려하는 문장 종속 화자 검증 시스템은 화자의 신원 정보 뿐 만 아니라 문장 정보도 함께 식별해야 하는 테스크이다. 본 논문에서는 문장 종속 화자 검증 시스템에서 적합한 딥러닝 기반 정보 처리를 위하여 음성인식기를 활용한 pooling 기법을 제안한다. 제안하는 기법은 음성인식기를 사용하여 문자 단위의 확률 분포를 추정하고 이를 pooling 단계에서 활용하여 화자 검증 시 화자 및 문장 정보를 함께 고려할 수 있는 알고리즘이다. 실험을 통해 문장 종속 화자 검증 시스템에서 제안하는 기법의 타당성과 유효함을 검증하였다.

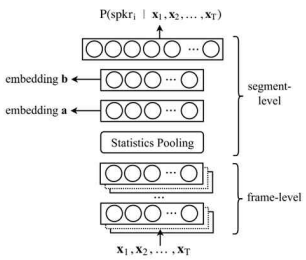
Ⅰ. 서 론

화자인식은 입력된 음성이 등록된 화자 중 어떤 사람의 목소리인지를 식별하는 분야이다. 화자인식은 목적에 따라 화자 식별과 화자 검증으로 나눌 수 있으며, 그 중 화자 검증은 특정 화자의 목소리가 맞는지 여부를 판별하는 테스크이다. 또한 시스템을 “Hi Bixby”와 같이 고정된 문장만을 고려하는 문장 종속 시스템과 임의의 문장에 대해 화자를 검증하는 문장 독립 시스템으로 세분화 할 수 있다. 문장 종속 시스템의 경우 화자와 문장이 모두 식별 되어야하기 때문에 순차적(sequential) 정보 및 컨텍스트 정보가 고려되어야 하며, 문장 독립 시스템의 경우 본 정보를 억제하고 화자 정보만을 추출해야 효과적인 화자 검증을 수행할 수 있다.

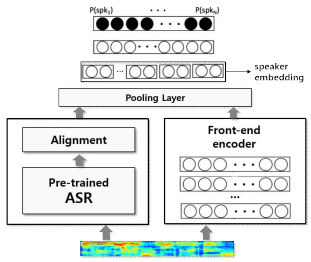
최근 다양한 딥러닝 기술들이 발전되고 대용량 공공 데이터 셋이 제공됨에 따라 컴퓨터비전 및 자연어처리 분야의 시스템 성능이 크게 향상되었다. 화자인식 분야에서도 딥러닝 기법 기반의 화자 임베딩(speaker embedding)을 적용하여 눈부신 성능 향상을 이루었으며 보다 효과적인 speaker embedding 추출을 위한 다양한 연구들(e.g., 심층심경망 구조, 손실 함수 모델링, pooling 기법 등)이 진행되고 있다. 그 중 speaker embedding 추출을 위한 pooling 기법은 프레임 단위의 피처를 고정된 차원의 벡터로 요약하는 단계이며, 이 과정에서 화자인식에 유효한 정보를 얼마나 잘 집계하는지가 성능에 직접적으로 연결된다. 본 논문에서는 컨텍스트 정보를 고려함으로써 문장 종속 화자 검증 시스템에 적합한 pooling 알고리즘을 제안 하고자 한다.

Ⅱ. Statistics pooling: x-vector 기법

최근 화자인식 분야에서 딥러닝 기반의 기법들을 적용하면서 신경망 구조, 대용량 데이터 셋, 목적 함수, pooling 기법 등의 다양한 연구가 활발히 이루어지고 있다. 그 중 pooling 기법은 프레임 단위의 입력 음성에 대한 프레임 단위의 출력 음성을 고정된 차원의 단일 벡터로 요약 하는 단계로, 전통적으로는 프레임 단위의 평균을 사용하여 고정된 차원의 벡터, 즉 speaker embedding을 추출하였다. 이 과정에서 단순히 동일한 가중치(평균)로 단일 벡터를 추출할 경우 정보 손실이 발생할 수 있기 때문에 이를 보완하기 위한 다양한 기법들이 발표되었다. 그 중 x-vector라 알려진 statistics pooling 기법 [1]은 평균 뿐 만 아니라, 프레임 단위의 출력에 대한 통계량(표준편차)를 함께 계산하고, 결합하여 affine layer를 통과하는 방식으로 학습 시 pooling 단계에서 프레임 간의 variability를 고려한



〔그림 1〕 x-vector



〔그림 2〕 제안하는 시스템 구조

기법이다. 학습 시에는 두 affine layer를 통과하며, 추론 시에는 마지막 layer를 추출하여 speaker embedding(x-vector)으로 사용 한다 [그림 1].

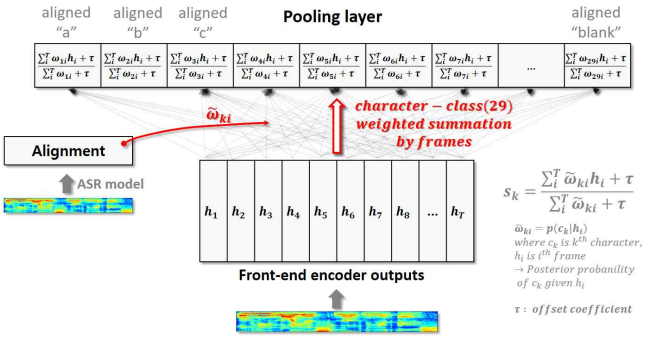
Ⅲ. ASR-based pooling 기법

음성 정보는 순차적 정보이며, 화자 정보 이외의 다양한 정보(e..g., 잡음, 공백, 환경, 녹음기기, 언어 등) 또한 포함하고 있다. 때문에, 보다 효과적인 화자 인식을 위해서는 중요한 정보(화자 정보, 문장 정보-문장 종속 경우)는 더 강조하여 처리하는 접근 방식이 필요하다. 최근 Short-duration Speaker Verification(SdSV) 챌린지에서 다 언어(영어, 페르시아어) 데이터 셋에서 효과적인 문장 종속 화자 검증을 위해 문자 단위의 pooling 기법을 활용한 기법이 제안되었다 [2]. 본 논문에서는 [2]에서 적용한 pooling 기법을 활용하여 단일 언어(영어) 문장 종속 화자 검증 시스템을 구축한다. [2]에서는 프레임 단위의 출력을 집계하는 과정에서 화자의 발화 정보를 고려하기 위해 CTC기반의 사전 학습된 ASR 모델을 활용 한다. ASR모델의 softmax 레이어 출력을 통해 추정한 posterior는 다음과 같이 정의 된다:

$$\pi_{k,i} = P(C = c_k | \boldsymbol{h_i}) \quad (1)$$

식 (1)에서, $C = \{c_k | c_k = k^{th} character, 1 \le k \le K\}$ 이며, K는 학습에 사용된 총 문장의 수, h_i는 i번째 프레임 레벨 출력, T는 총 프레임 수이다 ($1 \le i \le T$). ASR모델 학습을 위해 CTC loss를 사용했으므로 K는 26개의 알파벳(a-z)과 공백(space), 생략부호(apostrophe) 및 black를 포함한 29로 설정된다. 추정된 posterior를 사용한 pooling은 아래와 같다:

$$\boldsymbol{v_k} = \frac{\sum_{i=1}^{T} \pi_{k,i} \boldsymbol{h_i} + \tau}{\sum_{i=1}^{T} \pi_{k,i} + \tau} \quad (2)$$



〔그림 3〕 제안하는 기법

$$v = \left({v_1}^T \mid \ \ldots \ \mid {v_K}^T \right)^T \qquad (3)$$

여기서 τ는 발산을 막기 위한 상수 값이다. 추정된 posterior를 통해 문자별 계산을 개별적으로 처리한 후 v로 결합하는 과정을 갖는다. v는 affine layer 및 softmax를 통과하며, affine layer를 speaker embedding으로 사용한다 [그림 2, 3].

본 기법은 프레임 단위의 출력을 집계하는 과정에서 문자 단위의 개별적 처리 과정을 갖기 때문에 추론 단계에서 speaker embedding 간의 유사도 계산 시 특정 발음 간의 특징 비교를 가능하도록 하며, 이를 통해 문장 종속 화자 검증 시스템에서 화자 정보 및 문장 발화 정보를 동시에 고려하여 비교 분석 할 수 있다 [2] [그림 3].

IV. 실험

실험 및 검증을 위하여 구축한 모델 구조는 [그림 2]와 같다. 입력 음성을 위한 acoustic feature는 20ms 길이 및 10ms 간격의 hamming window를 통한 64차원의 Log Mel Filter Bank를 추출하여 사용하였으며, 300 프레임 단위로 랜덤 cropping 하였다. 프레임 단위 입력을 처리하는 네트워크로는 [1]에서 제안한 5층의 TDNN기반 구조를 채택하였으며 pooling layer 이후 세그먼트 단위 처리 네트워크는 2층의 LC/FC layer를 사용하였다. 마지막 layer를 제외한 각 층의 비선형 함수는 ReLU를 사용하였으며 최종 층은 화자 크기의 softmax layer를 통과한다. 단일 언어(영어) 문장 종속 화자 검증을 위한 화자 임베딩 네트워크 학습은 RSR2015를 사용하였으며 학습 및 테스트 프로토콜은 [표 1]과 같다. [2]에서 제안한 풀링 방식을 위하여 활용 될 음성인식기를 위한 학습은 [3]에서 제안한 CTC기반의 네트워크를 사용하며, Librispeech corpus를 학습하여 사용한다.

실험에서 사용한 성능 지표는 Equal Error Rate(EER)를 사용했으며, 문장 종속 화자 검증 시스템에서의 Target/Imposter 설정에 따라 4가지 type, Target-Correct(TC), Target-Wrong(TW), Imposter-Correct(IC), Imposter-Wrong(IW)의 성능을 검증하였다.

실험 결과는 총 3가지 모델에 대하여 비교, 분석 하였다. 첫 번째 모델은 화자인식에서 널리 사용되고 있는 딥러닝 기반의 x-vector 기법[1] 이며, 두 번째 모델은 x-vector 기법에 GMM을 활용하여 pooling 과정에서 mixture의 alignment 정보를 활용한 기법이다. 마지막 모델은 [2]에서 제안하는 pooling 기법을 활용한 ASR모델 기반 pooling 기법으로 특정 단의 발음 정보를 pooling과정에서 개별적으로 처리하고, 결합하는 방식이다. [표 2]의 결과에서 보듯, 제안하는 기법이 4가지 시나리오에서 모두 기존 기법들의 성능보다 향상된 결과를 보이는 것을 확인 할 수 있었다. 이는 제안하는 기법이 화자의 정보 뿐 만 아니라 발화 문장의 context 정보도 함께 고려하기 때문에 문장 종속 화자 검증 시스템에서 효과적으로 작용했다고 볼 수 있다. 또한 본 기법은 기존의 화자 검증 시스템에서 사용하는 프레임 워크에서 추정한 문자 단위 posterior만을 활용하는 구조 이

		# Total	# Male	# Female
Training (Part1~ Part3)		194 명 (9 세션) (127,241 발화)	100 명 (9 세션) (65,579 발화)	94 명 (9 세션) (61,662 발화)
Evaluation (Part1)	Enroll	106 명 (3 세션) (9,538 발화)	57 명 (3 세션) (5,128 발화)	49 명 (3 세션) (4,410 발화)
	Test	106 명 (6 세션) (19,066 발화)	57 명 (6 세션) (10,245 발화)	49 명 (6 세션) (8,810 발화)

〔표 1〕 학습 데이터 셋

Models	Target Trial Type	Non-Target Trial Type	EER[%]
x-vector	TC	TW	13.35
		IC	3.57
		IW	2.07
		TW, IC, IW	2.38
GMM-aligned x-vector	TC	TW	4.64
		IC	3.21
		IW	0.81
		TW, IC, IW	1.06
Proposed	TC	TW	**0.80**
		IC	**2.25**
		IW	**0.12**
		TW, IC, IW	**0.36**

〔표 2〕 실험 결과

므로 다양한 화자인식 기술들에 쉽게 적용할 수 있다는 장점이 있다.

V. 결론

본 논문에서는 문장 종속 화자 검증 시스템에서 활용 가능한 음성인식기 기반 pooling 기법을 활용한 프레임 워크를 제안한다. 본 기법은 CTC loss를 통해 사전 학습된 음성인식기를 활용하여 문자 단위의 확률 분포를 추정하며, 이를 사용함으로써 프레임 단위 출력 특징들의 pooling 단계에서 문장 정보를 고려 할 수 있도록 문자단위의 처리 과정을 갖는다. 실험을 통해 문장 종속 화자 검증 시스템에서 본 기법이 기존의 기법과 대비하여 효과적으로 동작하는 것을 검증하였다.

CKNOWLEDGMENT

이 논문은 2020년도 정부(경찰청)의 재원으로 지원받아 수행된 연구결과임 [과제명: 성문분석을 통한 실시간 화자검색 기술 개발 / 과제번호: PA-J00001-2017-101]

참 고 문 헌

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329 - 5333.

[3] S. Mun, W. Kang, M. Han and N. Kim, "Robust Text-Dependent Speaker Verification via Character-Level Information Preservation for the SdSV Challenge 2020," in *Proc INTERSPEECH*, 2020. (submitted)

[3] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, "Jasper: An end-to-end convolutional neural acoustic model," in *Proc. INTERSPEECH*, 2019, pp. 71 - 75.