한국음향학회지 제39권 제5호 pp. 468~475 (2020)
The Journal of the Acoustical Society of Korea Vol.39, No.5 (2020)
https://doi.org/10.7776/ASK.2020.39.5.468

pISSN : 1225-4428
eISSN : 2287-3775

잡음 환경에 강인한 기동어 검출을 위한 삼중항 손실 기반 도메인 적대적 훈련

Triplet loss based domain adversarial training for robust wake-up word detection in noisy environments

임형준,[1] 정명훈,[1] 김회린[1†]

(Hyungjun Lim,[1] Myunghun Jung,[1] and Hoirin Kim[1†])

[1]한국과학기술원 전기및전자공학부

(Received July 17, 2020; revised August 28, 2020; accepted August 29, 2020)

초 록: 단어의 특성을 잘 표현하는 음성 단어 임베딩은 기동어 인식에서 중요한 역할을 한다. 하지만 기동어 인식이 수행되는 환경에서 필연적으로 발생하는 다양한 종류의 잡음으로 인해 음성 단어 임베딩의 표현 능력이 손상될 수 있으며, 인식 성능의 저하를 초래할 수 있다. 본 논문에서는 음성 단어 임베딩에 영향을 줄 수 있는 환경적인 요인을 완화시키는 삼중항 손실 기반의 도메인 적대적 훈련 방식을 제안한다. 잡음 환경에서의 기동어 검출 실험을 통해 제안하는 방식이 기존의 도메인 적대적 훈련 방식을 효과적으로 개선하는 모습을 확인할 수 있었고, 잡음 환경에서의 기동어 검출을 위해 기존에 제안된 다른 방법과의 결합을 통해 제안하는 방식의 확장성을 확인할 수 있었다.
핵심용어: 기동어 검출, 음향 단어 임베딩, 도메인 적대적 훈련, 삼중항 손실

ABSTRACT: A good acoustic word embedding that can well express the characteristics of word plays an important role in wake-up word detection (WWD). However, the representation ability of acoustic word embedding may be weakened due to various types of environmental noise occurred in the place where WWD works, causing performance degradation. In this paper, we proposed triplet loss based Domain Adversarial Training (tDAT) mitigating environmental factors that can affect acoustic word embedding. Through experiments in noisy environments, we verified that the proposed method effectively improves the conventional DAT approach, and checked its scalability by combining with other method proposed for robust WWD.
Keywords: Wake-up word detection, Acoustic word embedding, Domain adversarial training, Triplet loss

PACS numbers: 43.72.Bs, 43.72.Ne

I. 서 론

기동어 검출(Wake-up Word Detection, WWD)은 장치를 필요할 때 활성화시켜 자원을 효과적으로 관리하는 기술로써, 최근 음성 관련 분야에서 많은 각광을 받고 있다. 애플의 "시리"와 아마존의 "에코"는 기동어 검출의 대표적인 예로, 입력되는 음향 신호로부터 "시리야", "알렉사"와 같은 특정 기동어 발화를 감지하여 장치를 활성화시킨다. 따라서 이와 같은 기동어 검출에서는 기동어 발화와 기동어가 포함되지 않은 일반적인 음성 발화를 잘 구별할 수 있는 능력이 필수적이다. 이와 관련하여 단어의 특성을 잘 표현해주는 방법에 대한 많은 연구들이 있었다. 고전적인 방식에서는 음성 단어 발성을 행렬 형태로 표현하는 방법이 많이 활용되었다.[1-3] 음성의 길이에 따라 행렬의 크기가 가변적이기 때문에 이와

†**Corresponding author:** Hoirin Kim (hoirkim@kaist.ac.kr)
School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-Gu, Daejeon 34141, Republic of Korea
(Tel: 82-42-350-7617, Fax: 82-42-350-7619)

같은 방법에서는 기동어 및 입력 음향 신호에 대한 행렬 사이의 거리를 동적 시간 워핑(Dynamic Time Warping, DTW)을 통해 계산하며, 이를 미리 정해둔 threshold와 비교하여 기동어 여부를 판단하게 된다. 하지만 DTW는 일반적으로 많은 계산량이 요구되기 때문에, 휴대 장치와 같은 제한된 자원에서 동작해야하는 기동어 검출에 적용되는 데에는 어려움이 존재한다. 최근에는 이와 같은 문제를 위해 음성 단어 발성을 벡터 형태로 나타내는 음성 단어 임베딩 방식들이 제안되었다.[4-7] 이 방법에서는 기동어 및 입력 음향 신호에 대한 거리를 벡터 사이의 코사인 거리로 계산하기 때문에 계산량 측면에서 장점을 보인다. 또한, 간단한 표현 방식에도 불구하고 최근의 연구들[4,5]에서는 기존의 행렬 형태의 표현 방식에 비해 좋은 성능이 보고되고 있다.

이처럼 단어 표현 방식에는 많은 발전이 있었지만, 그와 같은 표현들이 실제 기동어 검출이 동작하는 환경에서 유효한지에 대한 연구들은 상대적으로 부족하며, 대부분의 연구들에서는 훈련 과정에서 단순히 잡음이나 잔향이 섞인 다중 조건(multi-condition) 데이터를 활용하는데 그치고 있다.[3-4] 또한 실제 발생되는 잡음이나 잔향과 같은 요인들을 훈련 과정에 모두 포함시키는 것은 불가능하며, 훈련 및 평가 환경 사이의 불일치를 피할 수 없다. 이와 같은 어려움을 해결하기 위해 본 논문에서는 최근 도메인 적응 분야에서 뛰어난 성능을 보이는 도메인 적대적 훈련(Domain Adversarial Training, DAT)[8-14]을 통해 환경적인 요인에 대한 영향을 최소화하는 음성 단어 임베딩을 생성하는 방법을 제안한다. 제안하는 방법에서는 도메인 사이의 상대적인 관계를 학습하도록 도메인 삼중항 손실을 기존 DAT의 도메인 네트워크에 도입함으로써 훈련-평가 사이의 환경 불일치 문제를 완화시켰다.

제안하는 방법에 대한 효과성을 입증하기 위해 Aurora4 코퍼스를 이용한 잡음 환경에서의 기동어 검출 실험을 수행하였다. 더불어 제안하는 방법의 확장성을 확인하기 위해 훈련 과정에서 환경적인 요인에 대한 고려 없이 오직 단어 표현만을 강화시켜 환경적인 요인에 대한 문제를 해결하는 Lim *et al.*[6]의 방법과의 통합 실험을 수행하였다.

II. 삼중항 손실 기반 도메인 적대적 훈련

Fig. 1에 나타난 바와 같이, 제안하는 방법의 구조는 세 개의 인코더로 구성되어있다. 먼저, 입력 특징 벡터 $\mathbf{x}=(\mathbf{x}_1, \cdots, \mathbf{x}_T)$는 Long Short-Term Memory(LSTM) 기반의 공유 인코더 G_s를 통해 인코딩 행렬 $\mathbf{H}^s=[\mathbf{h}_1^s, \cdots, \mathbf{h}_T^s]$로 변환된다.

$$\mathbf{h}_t^s = G_s(\mathbf{x}_t;\theta_s)\in\mathbb{R}^d, \tag{1}$$

여기서 T는 프레임 개수, d는 인코딩 벡터의 차수를 나타낸다. 공유 인코더를 통해 변환된 인코딩 행렬은 단어 인코더 G_w와 도메인 인코더 G_d에 각각 입력되어 다음과 같이 음성 단어 임베딩 $\mathbf{h}^w$와 도메인 임베딩 $\mathbf{h}^d$를 각각 생성한다.

$$\mathbf{h}^w = G_w(\mathbf{h}_T^s;\theta_w)\in\mathbb{R}^d, \tag{2}$$

$$\mathbf{h}^d = \frac{1}{T}\sum_t G_d(\mathbf{h}_t^s;\theta_d)\in\mathbb{R}^d. \tag{3}$$

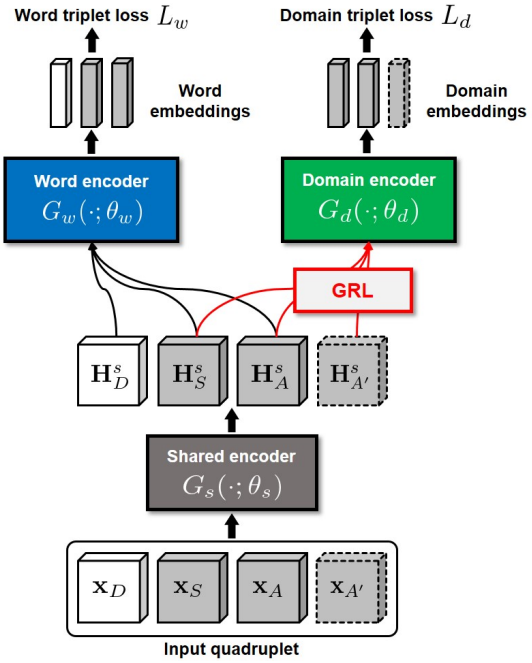

Fig. 1. (Color available online) Overall architecture of triplet loss based domain adversarial training (tDAT).

공유 인코더의 경우와 같이 단어 및 도메인 인코더는 LSTM으로 이루어진다. 단어를 고정된 차원의 벡터로 표현하는 Eq. (2)의 단어 임베딩의 경우 Reference [5]에서와 같이 단어 인코더의 마지막 은닉 상태를 사용하였다. 특히 Eq. (3)의 도메인 임베딩의 경우 도메인 인코더의 모든 은닉 상태에 대한 평균값을 고려함으로써 도메인 정보를 보다 효과적으로 나타낼 수 있도록 하였다.

입력 사중항 X에 대해, 전체 네트워크를 훈련하기 위한 목적 함수 E는 아래의 식과 같이 단어 및 도메인에 대한 삼중항 손실 함수 L_w와 L_d로 정의된다.

$$E(X;\Theta) = L_w(X_w;\Theta) - \lambda L_d(X_d;\Theta), \tag{4}$$

$$L_w(X_w;\Theta) = \max\{0, m + d(\mathbf{h}_A^w, \mathbf{h}_S^w) - d(\mathbf{h}_A^w, \mathbf{h}_D^w)\}, \tag{5}$$

$$L_d(X_d;\Theta) = \max\{0, m + d(\mathbf{h}_A^d, \mathbf{h}_S^d) - d(\mathbf{h}_A^d, \mathbf{h}_{A'}^d)\}, \tag{6}$$

여기서 $\Theta = \{\theta_s, \theta_w, \theta_d\}$는 전체 모수 집합, $\lambda \in [0,1]$는 도메인 삼중항 손실 함수에 대한 가중치, m은 삼중항 손실에서의 margin, $d(\cdot,\cdot)$은 코사인 거리를 각각 나타낸다. $X_w = \{\mathbf{x}_A, \mathbf{x}_S, \mathbf{x}_D\}$와 $X_d = \{\mathbf{x}_A, \mathbf{x}_S, \mathbf{x}_{A'}\}$는 Table 1과 같이 X의 부분 집합으로, 각각 단어 및 도메인 삼중항을 나타낸다. 단어 삼중항의 경우 기준이 되는 단어($\mathbf{x}_A$)에 대해 같은 단어($\mathbf{x}_S$, same word label)와 다른 단어($\mathbf{x}_D$, different word label)로 구성하였으며, 단어의 상대적인 관계만을 나타내기 위해 모두 같은 도메인을 사용하였다. 마찬가지로 도메인 삼중항의 경우 기준이 되는 도메인($\mathbf{x}_A$)에 대해 같은 도메인($\mathbf{x}_S$, same domain label)과 다른 도메인($\mathbf{x}_{A'}$, different domain label)로 구성하였으며, 도메인의 상대적인 관계만을 나타내기 위해 모두 같은 단어를 사용하였다. 기존 DAT[8]의 경우 도메인을 분류하는 목적으로 교차 엔트로피 함수를 최소화하도록 도메인 네트워크가 훈련되는 것이 일반적이지만, 제안하는 방법에서는 도메인 임베딩 사이의 상대적인 관계를 나타내는 삼중항 손실 함수를 최소화하도록 도메인 네트워크를 훈련하여 도메인 임베딩이 훈련 데이터에 나타나지 않은 미지의 도메인에 대해서도 일반화될 수 있도록 하여 훈련-평가 불일치 문제를 해결하고자 했다.

Table 1. Configuration of input quadruplet X.

Input	Word label	Domain label	X_w	X_d
$\mathbf{x}_A$	Same	Same	✓	✓
$\mathbf{x}_S$	Same	Same	✓	✓
$\mathbf{x}_D$	Different	Same	✓	
$\mathbf{x}_{A'}$	Same	Different		✓

목적 함수 E를 최적화하는 것은 다음과 같이 두 개의 최적화 문제로 나타낼 수 있다.

$$(\hat{\theta}_s, \hat{\theta}_w) = \arg\min_{\theta_s,\theta_w} E(X;\theta_s,\theta_w,\hat{\theta}_d), \tag{7}$$

$$\hat{\theta}_d = \arg\max_{\theta_d} E(X;\hat{\theta}_s,\hat{\theta}_w,\theta_d). \tag{8}$$

각각의 모수들은 다음과 같은 규칙으로 갱신된다.

$$\theta_s \leftarrow \theta_s - \mu\left(\frac{\partial L_w}{\partial \theta_s} - \lambda\frac{\partial L_d}{\partial \theta_s}\right), \tag{9}$$

$$\theta_w \leftarrow \theta_w - \mu\frac{\partial L_w}{\partial \theta_w}, \tag{10}$$

$$\theta_d \leftarrow \theta_d - \mu\lambda\frac{\partial L_d}{\partial \theta_d}, \tag{11}$$

여기서 μ는 학습률을 나타낸다. 단어 및 도메인 인코더의 경우 단순히 Eqs. (4)와 (5)의 단어 및 도메인 삼중항 손실을 최소화 하는 방향으로 훈련되지만, 공유 인코더의 경우 단어 삼중항 손실을 최소화하는 동시에 도메인 삼중항 손실을 최대화하는 방향으로 훈련된다. 즉, 공유 인코더는 단어는 잘 구분하는 동시에 도메인은 잘 구분하지 못하도록 적대적으로 훈련된다. 기존 DAT[8]에서와 마찬가지로 공유 인코더의 모수 θ_s를 확률적 경사하강법으로 최적화하기 위해 역전파시에만 gradient의 부호를 바꿔주는 GRL(Gradient Reversal Layer)[15]을 활용하였다(Fig. 1).

III. 성능 평가

본 논문에서 제안하는 방법의 성능 평가를 위해 등록 및 검증 단계로 구성된 기동어 검출 실험을 수행하였다. 등록 단계에서 사용자는 자신이 원하는 기동어를 선정 및 발성하게 되고, 발성된 발화에 대해 Eq. (2)로 주어지는 음성 단어 임베딩 $\mathbf{h}^w$를 생성한다. 등록에는 총 3회의 기동어 발성을 사용하였다.

검증 단계에서는 연속적으로 입력되는 음향 신호 내에 사용자가 등록했던 기동어 발성이 존재하는지 여부를 판단한다. 일반적으로 기동어 인식은 사용자가 임의의 시간에 발성하는 기동어를 검출해내야 하므로 매 순간 기동어의 존재 여부를 판단해야 한다. 이를 위해 특정 길이(≈1 s)의 윈도우를 이동시키면서(≈0.1 s) 연속적으로 검증을 수행했으며, 등록된 기동어에 대한 음성 단어 임베딩들과 윈도우에 대한 음성 단어 임베딩 사이의 코사인 거리의 평균값을 정해진 threshold와 비교하여 기동어 여부를 판단하였다. 일단 기동어가 감지되어 장치가 활성화되면 기동어 검출이 종료되는 실제 상황을 고려하여 기동어를 포함하는 연속적인 윈도우들 중에서 한 번의 검증만을 수행하도록 했다. 즉, Fig. 2에서와 같이 기동어를 일정부분 포함하는 세그먼트들에 대해 각각 코사인 거리를 계산한 뒤 최솟값만을 활용하여 검증을 수행한다. 여기서 기동어를 포함하는 윈도우는 아래의 식에 의해 정해진다.

$$l_{overlap} > k_{tol} \times l_{wakeup}, \tag{12}$$

여기서 $l_{overlap}$은 윈도우내에 기동어가 포함된 부분에 대한 길이를, l_{wakeup}은 기동어 구간에 대한 길이를 나타내며 아래와 같이 계산된다.

$$l_{overlap} = \min\{t_{window,end}, t_{wakeup,end}\} - \max\{t_{window,start}, t_{wakeup,start}\}. \tag{13}$$

$$l_{wakeup} = t_{wakeup,end} - t_{wakeup,start}, \tag{14}$$

여기서 $t_{window,start}$, $t_{window,end}$ 는 윈도우의 시작점과 끝점을 $t_{wakeup,start}$, $t_{wakeup,end}$는 기동어의 시작점과 끝점을 각각 나타낸다. Eq. (12)의 k_{tol}은 기동어 구간이 잘리는 것을 허용하는 정도를 나타낸 상수값으로, $k_{tol} = 1.0$일 경우에는 기동어 구간이 온전히 포함된 윈도우만을 사용하는 것을 의미한다. 본 논문에서는 $k_{tol} = 0.8$을 사용하였다.

3.1 실험 설정

음성 인식 분야에서 폭넓게 활용되고 있는 WSJ 코퍼스[16] SI-284셋으로부터 무작위로 선택한 50만 개의 삼중항(12538 개의 고유 단어)으로 깨끗한 환경의 훈련 데이터를 구성하였다. 본 논문에서는 도메인을 잡음 환경으로 한정했으며, 이를 위해 QUT 잡음 데이터베이스[17]를 5 dB ~ 15 dB의 Signal to Noise Ratio (SNR)로 섞어 50만 개의 삼중항을 갖는 다중 조건 훈련 데이터를 생성하였다. 먼저 clean 환경의 삼중항 $\{\mathbf{x}_A, \mathbf{x}_S, \mathbf{x}_D\}$에 대해 랜덤하게 선택된 잡음을 1종을 동일하게 섞은 다음, 다른 종류의 잡음 1종을 추가로

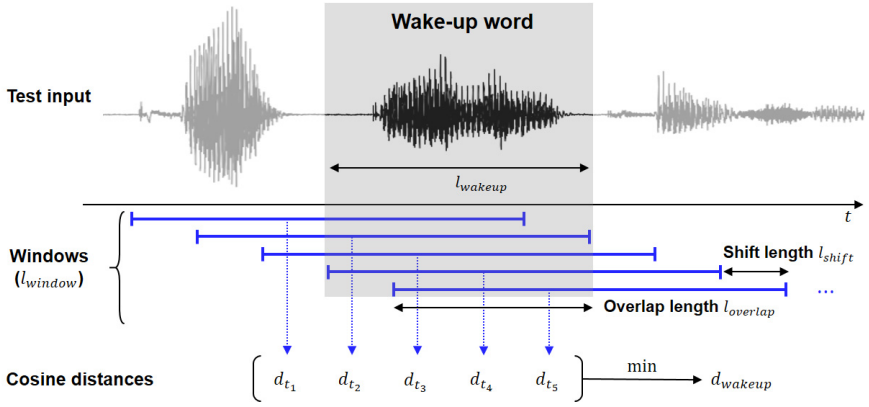

Fig. 2. (Color available online) An example of calculating the distance of wake-up word.

선택하여 $\mathbf{x}_A$에 섞어 $\mathbf{x}_{A'}$를 생성하였다. Eq. (4)에서의 가중치 λ를 결정하기 위한 개발 데이터로는 WSJ dev93셋에 마찬가지로 QUT 잡음을 섞은 데이터를 활용하였다. 여기서 multi-condition 훈련 데이터에는 'HOME' 잡음을, 개발 데이터에는 'CAFE', 'STREET', 'CAR', 'REVERB' 잡음을 각각 활용하여 훈련 데이터와 개발 및 시험 데이터 사이에 잡음의 종류가 겹치지 않도록 하였다. 특히 시험 데이터로는 WSJ에 여섯 가지 종류(car, babble, restaurant, street, airport, train)의 잡음 환경이 10 dB ~ 20 dB의 SNR로 섞인 Aurora4 코퍼스[18] dev1206 셋을 사용하였다. 기동어의 경우 Reference [6]에서와 같이 시험 데이터에서 최소 4회의 빈도수를 보이는 단어들 중에서 무작위로 선정하였으며, Table 2에 사용된 기동어를 나열하였다.

입력으로는 25 ms의 프레임 크기(10 ms overlap)에서 계산된 40차의 Mel-filterbank 로그 에너지를 사용하였다. 전체 네트워크는 128개의 셀로 구성된 단방향 LSTM 4개 층으로 구성하였으며, 공유 인코더는 2개 층, 단어 및 도메인 인코더는 각각 1개 층을 할당했다. 네트워크의 모수들은 random한 초기값을 갖도록 하였고, TensorFlow 툴킷[19]의 ADAM[20] 알고리즘으로 훈련하였다$\left(\beta_1 = 0.9,\ \beta_2 = 0.99,\ \epsilon = 1 \cdot 10^{-8}\right)$. 총 20번의 반복 학습을 진행하는 동안 학습률은 0.001로 고정하였고, 미니배치는 128개의 샘플을 사용하였다. 여기서 하나의 샘플에는 하나의 기동어를 포함시켰다. Eq. (4)의 가중치 λ는 개발 데이터의 실험 결과를 통해 0.01로 결정하였다.

제안된 방식의 효과를 비교하기 위해 다음과 같이 기준 방법들을 구성하였다.

- Multitask learning: Fig. 1에서 GRL이 제외된 구조로, $E(X;\Theta) = L_w(X_w;\Theta) + \lambda L_d(X_d;\Theta)$를 최적화하도록 학습된 모델을 의미한다. 사용된 도메인 손실 L_d의 종류에 따라 MT(교차 엔트로피 함수)와 tMT(삼중항 손실 함수)로 나타냈다.
- Domain adversarial training: 기존의 방법에서와 같이 도메인 손실 L_d에 교차 엔트로피 함수를 사용한 방식을 의미하며, DAT로 나타냈다.

Table 2. Selected wake-up words for Aurora4 corpus (in alphabetical order, #total = 24).

American, analysts, brokerage, company, corporation, department, double-quote, employment, February, hyphen, important, industry, investors, outstanding, percent, petroleum, production, question-mark, September, Shearson, spokesman, trading, western, Westinghouse

3.2 실험 결과

Fig. 3은 제안하는 방법 및 기준 방법들의 기동어 검출 성능을 Receiver Operating Characteristic(ROC) 곡선[21]으로 나타냈으며, 특히 기동어 검출에서 요구되는 낮은 오검출율(False Alarm Rate, FAR)에서의 재현율(recall)에 초점을 맞췄다. 'AUC'는 ROC 곡선 아래의 면적을 의미하며, 전체적인 성능을 나타낸다. 재현율과 오검출율은 다음과 같이 정의된다.

$$Recall = \frac{TP}{TP + FN}, \tag{15}$$

$$FAR = \frac{FP}{FP + TN}. \tag{16}$$

그림에서 보는 바와 같이, 제안하는 tDAT 방식은 전반적으로 다른 기준 방법들을 능가하는 모습을 확인할 수 있었다. MT에 비해 DAT의 성능이 우수한 것을 확인할 수 있었으며, 도메인에 대해 적대적인 학

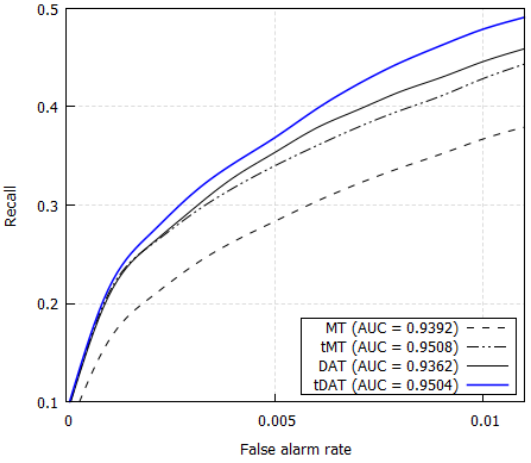

Fig. 3. (Color available online) ROC curve comparing the performance of the baslines (black lines) and the proposed tDAT (blue line) on the Aurora4 corpus.

Table 3. Performance comparison for each type of noise from the Aurora4 corpus. $R@[x]$ denotes recall at x false alarm rate.

Method	$R@[0.01, 0.005]$					
	Car	Babble	Restaurant	Street	Airport	Train
MT	0.457 / 0.300	0.428 / 0.273	0.356 / 0.252	0.430 / 0.315	0.411 / 0.273	0.428 / 0.314
tMT	0.538 / 0.377	0.475 / 0.329	0.458 / 0.316	0.504 / 0.370	0.469 / 0.335	0.498 / 0.346
DAT	0.542 / 0.387	0.505 / 0.358	**0.475 / 0.337**	0.500 / 0.358	0.511 / **0.380**	0.490 / 0.337
tDAT (Proposed)	**0.610 / 0.434**	**0.537 / 0.373**	0.469 / 0.322	**0.540 / 0.381**	**0.547** / 0.371	**0.530 / 0.365**

Table 4. Performance comparison of ISAN trained in various ways for each type of noise from the Aurora4 corpus. $R@[x]$ denotes recall at x false alarm rate.

Method	$R@[0.01, 0.005]$					
	Car	Babble	Restaurant	Street	Airport	Train
ISAN (clean)	0.719 / 0.529	0.594 / 0.493	0.549 / 0.424	0.670 / 0.570	0.694 / 0.584	0.683 / 0.569
+ MT	0.781 / 0.669	0.741 / 0.592	0.679 / 0.520	0.752 / 0.616	0.749 / 0.613	0.749 / 0.611
+ tMT	0.851 / 0.758	0.815 / 0.700	0.721 / 0.603	**0.822** / 0.720	0.800 / 0.687	0.806 / 0.698
+ DAT	0.822 / 0.724	0.774 / 0.670	0.698 / 0.571	0.777 / 0.669	0.787 / 0.673	0.787 / 0.683
+ tDAT (Proposed)	**0.851 / 0.781**	**0.834 / 0.745**	**0.758 / 0.632**	0.811 / **0.722**	**0.847 / 0.768**	**0.823 / 0.758**

습 방식이 환경적인 요인을 보다 효과적으로 대처할 수 있다는 것을 확인할 수 있었다. MT와 DAT 모두 도메인 삼중항 손실 함수를 사용했을 때 성능 향상을 관찰할 수 있었다. 이를 통해 학습-평가 사이의 환경 불일치가 있을 경우 도메인의 상대적인 관계를 통한 학습이 효과가 있음을 확인할 수 있었다. tMT의 경우 제안하는 tDAT에 비해 AUC 측면에서 좋은 성능을 보이기도 했지만, 낮은 FAR 영역에 초점을 맞추었을 때에는 그렇지 못한 모습을 보였으며, 이를 통해 제안하는 tDAT 방식이 기동어 검출에 보다 적합한 방식임을 확인할 수 있었다. Aurora4의 각 잡음에 대해 성능을 정리한 Table 3에서도 이와 같은 경향성을 확인할 수 있었다. 특정 잡음 환경에서 기준 방법들은 제안하는 방법에 비해 좋은 성능을 보이기도 했으나, 모든 잡음 환경에 대해서 일반화되지는 못했다.

제안하는 방법의 확장성을 확인하기 위해 이전에 제안된 Interlayer Selective Attention Network(ISAN)[6]을 제안하는 tDAT 및 앞에서와 같은 모든 기준 방법들을 통해 훈련시키는 추가 실험을 진행하였다. ISAN은 잡음 환경에서의 기동어 검출을 위해 제안된 방법으로, 인간이 자극에 반응할 때 자극의 중요한 특징에 초점을 맞추고 관련 없는 것을 억제한다는 내

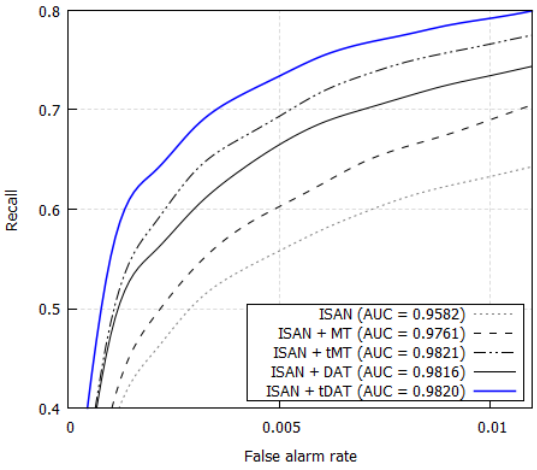

Fig. 4. (Color available online) ROC curve comparing the performance of ISAN trained in various ways on the Aurora4 corpus.

용의 선택적 주의 이론을 음향 단어 임베딩에 적용한 방식이다. 이 방법에서는 다른 단어들과 구별되는 단어 내의 중요한 부분들만을 강조하여 임베딩의 단어 구별 능력을 향상시킴과 동시에 환경의 영향을 최소화하도록 했으며, 특히 훈련 과정에서 환경적인 요인을 고려하지 않았지만 환경에 강인한 모습을 보였다는 점에서 의의가 있다. 여기서 단어 내의 중요도는 문자 및 단어 수준 인코딩 정보 사이의 유사도가 사용된다.

앞서 언급했던 기준 방법들 및 제안하는 tDAT를 이용하여 훈련한 ISAN에 대한 기동어 검출 실험 결과를 Fig. 4에 나타내었다. 그림에서 보는바와 같이 제안하는 tDAT 방식은 ISAN의 성능을 가장 효과적으로 향상시켰다. 앞의 결과에서와 마찬가지로 도메인 인코더를 삼중항 손실로 최적화했을 때 좋은 성능을 나타냈다. 흥미로운 점은 tMT가 DAT에 비해 성능 향상의 폭이 컸다는 것인데, 이를 통해 삼중항 손실이 학습-평가 환경 불일치 상황에서 기존의 교차 엔트로피 손실에 비해 효과적이라는 사실을 다시 한 번 확인할 수 있었다. Aurora4의 각 잡음에 대해 을 Table 4에 정리하였으며, 앞에서와 마찬가지로 특정 잡음 환경에서 제안하는 방법이 다소 낮은 성능을 보이기도 했지만, 전체적으로는 우수한 성능을 보이는 것을 확인할 수 있었다.

IV. 결 론

본 논문에서는 잡음 환경에서의 기동어 검출을 위한 방법으로 삼중항 손실 기반의 도메인 적대적 학습 방식을 제안하였다. 기존의 도메인 적대적 학습 방식에서 도메인 네트워크를 훈련하는 교차 엔트로피 손실을 도메인 사이의 상대적인 관계를 학습하기 위해 삼중항 손실로 대체시킴으로써 훈련-평가 사이의 환경적인 불일치 문제를 완화시켰다. 제안하는 방법에 대한 성능 평가를 위해 잡음 환경에서의 기동어 검출 실험을 수행하였으며, 잡음 데이터를 훈련에 활용하는 다른 기준 방법들에 비해 높은 성능을 확인하였다. 더불어 단어 표현 능력을 향상시키는 ISAN 방식과의 통합 실험을 수행하여 제안하는 방법의 확장성을 확인하였다.

감사의 글

본 연구는 산업통상자원부의 산업기술혁신사업으로부터 지원을 받아 수행된 연구임(No. 10063424, ‘실내용 음성대화 로봇을 위한 원거리 음성인식 기술 및 멀티 태스크 대화처리 기술 개발’).

References

1. Y. Zhang and J. R. Glass “Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams,” Proc. ASRU. 398-403 (2009).
2. G. Mantena and K. Prahallad, “Use of articulatory bottle-neck features for query-by-example spoken term detection in low resource scenarios,” Proc. ICASSP. 7128-7132 (2014).
3. H. Lim, Y. Kim, Y. Kim, and H. Kim, “CNN-based bottleneck feature for noise robust query-by-example spoken term detection,” Proc. APSIPA. 1278-1281 (2017).
4. G. Chen, C. Parada, and T. N. Sainath, “Query-by-example keyword spotting using long short-term memory networks,” Proc. ICASSP. 5236-5240 (2015).
5. S. Settle and K. Livescu, “Discriminative acoustic word embeddings: Recurrent neural network-based approaches,” Proc. SLT. 503-510 (2016).
6. M. Jung, H. Lim, J. Goo, Y. Jung, and H. Kim, “Additional shared decoder on Siamese multi-view encoders for learning acoustic word embeddings,” Proc. ASRU. 629-636 (2019).
7. H. Lim, Y. Kim, J. Goo, and H. Kim, “Interlayer selective attention network for robust personalized wake-up word detection,” IEEE Signal Process. Lett. **27**, 126-130 (2020).
8. Y. Ganin, H. Ajakan, H. Larochelle, F. Laviolette, and V. Lempitsky, “Domain-adversarial training of neural networks,” J. Mach. Learn. Res. **17**, 2096-2030 (2016).
9. E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial discriminative domain adaptation,” Proc. CVPR. 7167-7176 (2017).
10. Z. Pei, Z. Cao, M. Long, and J. Wang, “Multi-adversarial domain adaptation,” Proc. AAAI. 3934-3941 (2018).
11. R. Wang, M. Utiyama, A. Finch, L. Liu, K. Chen, and E. Sumita, “Sentence selection and weighting for neural machine translation domain adaptation,” IEEE/ACM Trans. Audio, Speech, Lang. Process. **26**, 1727-1741 (2018).
12. A. Tripathi, A. Mohan, S. Anand, and M. Singh, “Adversarial learning of raw speech features for domain invariant speech recognition,” Proc. ICASSP. 5959-5963 (2018).
13. S. Sun, C. F. Yeh, M. Y. Hwang, M. Ostendorf, and L. Xie, “Domain adversarial training for accented speech recognition,” Proc. ICASSP. 4854-4858 (2018).
14. S. Mirsamadi and J. H. Hansen, “Multi-domain adversarial training of neural network acoustic models for distant speech recognition,” Speech Commun. **106**, 21-30 (2019).
15. Y. Ganin and V. Lempitsky, “Unsupervised domain

adaptation by backpropagation," Proc. ICML. 1180-1189 (2015).
16. D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," Proc. Workshop Speech and Natural Lang. 357-362 (1992).
17. D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," Proc. Interspeech, 3110-3113 (2010).
18. H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," Proc. ISCA ITRW ASR. 181-188 (2000).
19. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Lenvnberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wiche, Y. Yu, and X. Zheng, "Tensor Flow: Large-scale machine learning on heterogeneous systems," Proc. USENIX OSDI. 265-283 (2016).
20. D. Kingma and J. Ba, "Adam: A method for stochastic optimization," Proc. ICLR. 1-15 (2015).
21. K. Hajian-Tilaki, "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation," Caspian J. Intern. Med. **4**, 627-635 (2013).

저자 약력

▸ 임 형 준 (Hyungjun Lim)

2013년 : 중앙대학교 전자전기공학부 학사
2013년 ~ 현재 : KAIST 전기및전자공학부 석박사 통합과정

▸ 정 명 훈 (Myunghun Jung)

2018년 : 한양대학교 융합전자공학부 학사
2018년 ~ 현재 : KAIST 전기및전자공학부 석박사 통합과정

▸ 김 회 린 (Hoirin Kim)

1984년 : 한양대학교 전자공학과 학사
1987년 : KAIST 전기및전자공학부 석사
1992년 : KAIST 전기및전자공학부 박사
1987년~1999년 : ETRI 선임연구원
2000년 ~ 현재 : KAIST 전기및전자공학부 교수