일반논문 (Regular Paper)
방송공학회논문지 제25권 제5호, 2020년 9월 (JBE Vol. 25, No. 5, September 2020)
https://doi.org/10.5909/JBE.2020.25.5.742
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

공용 신경망의 다중 학습을 통한 음소와 감정 인식의 성능 향상

김 재 원[a)], 박 호 종[a)‡]

Performance Enhancement of Phoneme and Emotion Recognition by Multi-task Training of Common Neural Network

Jaewon Kim[a)] and Hochong Park[a)‡]

요 약

본 논문에서는 하나의 공용 신경망을 사용하여 음소와 감정을 모두 인식하는 방법과 공용 신경망 학습을 위한 다중 학습 방법을 제안한다. 공용 신경망은 동일한 동작을 수행하여 두 정보를 모두 인식하며, 이는 인간이 하나의 청각기관으로 여러 정보를 동시에 인식하는 구조에 해당한다. 다중 학습은 여러 정보를 위한 공통 모델링을 진행하므로 여러 정보에 대한 일반화된 학습을 진행시켜 기존의 정보별 개별 학습에서 나타나는 과적합을 감소시키고 인식 성능을 향상시킨다. 또한, 다중 학습에서 음소 인식에 가중치를 부여하여 음소 인식 성능을 추가 향상시키는 방법을 제안한다. 동일한 특성벡터와 신경망을 사용할 때, 제안한 다중 학습이 적용된 공용 신경망의 성능이 각 정보별로 학습시킨 개별 신경망에 비하여 우수한 것을 확인하였다.

Abstract

This paper proposes a method for recognizing both phoneme and emotion using a common neural network and a multi-task training method for the common neural network. The common neural network performs the same function for both recognition tasks, which corresponds to the structure of multi-information recognition of human using a single auditory system. The multi-task training conducts a feature modeling that is commonly applicable to multiple information and provides generalized training, which enables to improve the performance by reducing an overfitting occurred in the conventional individual training for each information. A method for increasing phoneme recognition performance is also proposed that applies weight to the phoneme in the multi-task training. When using the same feature vector and neural network, it is confirmed that the proposed common neural network with multi-task training provides higher performance than the individual one trained for each task.

Keyword : deep neural network, common recognition, multi-task training, emotion recognition, phoneme recognition

a) 광운대학교 전자공학과(Dept. of Electronics Engineering, Kwangwoon University)
‡ Corresponding Author : 박호종(Hochong Park)
E-mail: hcpark@kw.ac.kr
Tel: +82-2-940-5104
ORCID: https://orcid.org/0000-0003-1600-6610

※ 본 논문은 2019년도 정부(교육부)의 재원으로 한국연구재단의 지원(NRF-2016R1D1A1B03930923)을 받아 수행된 연구임. (The present Research has been conducted by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2016R1D1A1B03930923)).
· Manuscript received May 7, 2020; Revised July 1, 2020; Accepted September 9, 2020.

Ⅰ. 서 론

최근 기계학습 기반으로 음소(phoneme), 감정(emotion), 화자(speaker) 등의 음성 정보를 인식하는 기술이 널리 연구되고 있다[1-6]. 기존 연구는 각 정보별로 최적의 특성벡터(feature vector)와 인식기 구조를 설계하며, 그에 따라 각 정보 인식기는 서로 다른 특성벡터를 추출하여 개별적으로 학습되고 서로 다른 동작을 수행한다. 즉, 하나의 음성 입력에 대하여 각 정보별로 특화된 개별 인식기가 병렬로 동작하여 각 정보를 인식하며, 이 구조는 마치 인간이 여러 개의 청각기관을 가지며 각 청각기관이 서로 다른 정보를 각각 인식하는 것과 같다.

인간은 하나의 청각기관을 가지며, 하나의 신호를 입력하여 모든 정보에 대한 공통 분석을 진행하고 뇌의 마지막 단계에서 정보별 독립 동작을 수행하여 모든 정보를 동시에 인식한다[5]. 예로, 인간은 음성 신호를 청취하여 음소와 감정을 동시에 인식한다. 이 동작은 그림 1과 같이 하나의 특성 추출기(feature extractor)와 분류기(classifier)를 공용으로 사용하여 하나의 음성 입력으로부터 여러 정보를 동시에 인식하는 것에 해당하며, 이와 같은 공용 인식기는 여러 정보를 모두 인식하도록 학습되어야 한다.

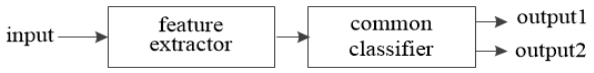

그림 1. 공용 인식기를 사용하는 다중 정보 인식 구조
Fig. 1. Structure of multiple information recognition using common recognizer

본 논문에서는 인간의 청각 동작을 모방하는 새로운 인식 구조를 개발하는 목표에 따라 하나의 공용 인식기를 사용하여 음소 인식과 감정 인식을 동시에 수행하는 기술을 개발한다. 이를 위해 공용 인식기가 두 정보를 모두 인식하도록 학습시키는 다중 학습(multi-task training) 방법과, 학습 과정에서 각 정보 사이에 성능을 조정하는 방법을 개발한다. 또한, 음소와 감정 인식에 필요한 정보 분석의 시간 해상도가 서로 다르므로, 공용 인식기가 두 가지 시간 해상도로 입력되는 특성벡터를 동시에 모델링 하도록 학습시킨다. 이에 따라 특성벡터에 상관없이 공용 인식기는 동일한 동작을 수행하여 두 정보를 모두 인식한다.

기계학습의 성능을 결정하는 중요한 요인 중에 하나가 학습의 일반화(generalization)이고, 일반화를 증대시키는 여러 방법이 개발되었다[7,8]. 기존의 일반화는 학습 결과가 학습 데이터에 의존하는 것을 방지하는 것이고, 이는 데이터에 대한 일반화이다. 일반화 개념을 확장하면 인식할 정보에 대한 일반화가 정의되고, 하나의 인식기가 서로 다른 정보를 동시에 인식하도록 학습시키면 여러 정보에 대하여 일반화된 학습이 진행되므로 학습 결과가 하나의 정보 인식에 존속되는 문제가 완화되고 학습의 일반화가 향상 될 수 있다. 따라서 제안하는 공용 인식기의 다중 학습을 통하여 기존의 개별 인식기에 비하여 성능을 향상시키는 것을 목표로 한다.

완전히 동일한 특성벡터를 사용하는 유사한 특성의 정보 인식에서, 인식기가 여러 정보 인식을 수행하도록 학습시키면 성능이 향상되는 것이 보고되었다[9]. 그러나 본 논문에서는 서로 다른 시간 해상도의 특성벡터를 사용해야 하고 음소와 감정은 유사한 음향 및 음성 정보가 아니므로, [9]에서 가정하였던 조건이 성립하지 않고 더 일반화된 환경에서 다중 학습에 의한 성능 향상을 검증한다. 또한, 본 논문의 다중 정보 인식은 하나의 음원에 해당하는 음성 신호로부터 서로 다른 기준에 따라 음향 특성을 분류하여 여러 종류의 음성 정보를 인식하는 것으로서, 여러 음원이 혼합된 신호에서 각 음원 정보를 인식하는 기존의 multi-label 분류와는 차이를 가진다[10]. 기계학습에서 다른 정보를 위해 이미 학습한 모델을 활용하는 전이학습(transfer learning) 기법이 있다[11]. 이 방법은 서로 다른 정보 인식을 위한 학습 결과를 공유하는 면에서 본 논문의 목표와 유사하지만, 기존 학습 결과를 새로운 학습을 위한 초기 데이터로 활용하는 것에 불과하고 본 논문과는 다르게 여러 정보 인식을 위한 다중 학습을 진행하지 않는다.

본 논문의 목표는 인식 성능 향상을 위한 새로운 특성벡터와 인식기 동작을 개발하는 것이 아니라, 기존 기술을 바탕으로 공용 인식기를 구현하고 다중 학습을 통하여 성능 향상을 얻는 것이다. 따라서 기존의 음성 정보 인식에 널리 사용되는 특성벡터와 심층 신경망(deep neural network, DNN)[7]을 그대로 사용하여 공용 인식기를 개발하고, 공용 인식기 학습을 위한 새로운 다중 학습 방법을 개발한다. 즉, 제안 기술은 DNN의 성능을 향상시키는 새로운 학습 방법

에 해당한다. 성능 평가를 통하여 하나의 공용 인식기가 음소와 감정을 모두 인식할 수 있는 것을 검증하고, 동일한 특성벡터와 DNN을 사용할 때 제안하는 공용 DNN의 성능이 기존 기술에 따라 각 정보별로 학습된 개별 DNN보다 향상되는 것을 확인한다.

Ⅱ. 제안하는 공용 신경망의 다중 학습 방법

다중 학습의 목표는 모든 정보 인식에 대한 공통 모델을 얻는 것이고, 이를 위하여 모든 정보 인식에 공통적으로 적용할 수 있는 일반적인 범용 특성벡터를 사용해야 한다. 예로, 기존의 다중 학습에서는 동일한 특성벡터를 모든 인식기에 공통으로 적용하여 성능의 향상을 얻는다[9]. 따라서 본 논문에서는 기존의 각 정보 인식기가 사용하는 특화된 특성벡터가 아니라, 음성 신호의 분석에 널리 사용되는 범용 특성벡터를 사용한다. 그에 따라 Mel-spectrogram(M-spec)과 Mel-frequency cepstral coefficient(MFCC)를 기반으로 두 종류의 특성벡터를 정의하고 각각에 대한 동작을 검증한다[12,13]. 이를 통해 특성벡터의 종류에 관계없이 공용 DNN의 다중 학습에 의하여 성능 향상이 가능한 것을 검증한다.

프레임 단위로 입력 신호의 스펙트럼을 구하고, 주파수를 mel scale로 변환한다[12]. 동일한 mel 간격으로 42개 밴드를 정의하고 밴드의 로그 에너지를 구하여 42개의 M-spec을 구하고, 여기에 42-point discrete cosine transform을 적용하여 42개 MFCC를 얻는다[13]. 다음, M-spec과 MFCC의 프레임 변화 특성을 추출하기 위해 각각의 1차 프레임 델타(delta)와 2차 프레임 델타를 구하고[14], 최종적으로 각 프레임별로 M-spec과 MFCC 기반의 $42 \times 3 = 126$차원 특성벡터를 각각 정의한다. 아래에 설명하는 모든 동작은 두 종류의 특성벡터에 대하여 각각 독립적으로 수행하고 성능을 측정한다.

음소 인식은 높은 시간 해상도의 특성 모델링을 요구하고, 감정 인식은 상대적으로 낮은 시간 해상도의 특성 모델링을 요구한다. 따라서 제안하는 공용 인식기가 감정과 음소 인식을 동시에 수행할 때 하나의 시간 해상도로 설정된 동일한 특성벡터를 사용할 수는 없다. 이 상황에서 학습의 일반화 효과를 높이기 위하여 동일한 물리적 의미를 가지면서 시간 해상도만 달리하는 특성벡터를 설계한다. 즉, 그림 1의 특성벡터 추출단에서 동일한 특성벡터를 높은 시간 해상도와 낮은 시간 해상도에서 각각 구하고, 이를 하나의 공용 DNN에 입력한다. 공용 DNN은 높은 해상도의 특성벡터로부터 음소를 인식하고 낮은 해상도의 특성벡터로부터 감정을 인식하도록 학습하며, 그에 따라 입력 특성벡터의 해상도에 관계없이 공용 DNN은 동일한 동작을 수행하여 두 정보를 모두 인식한다.

표 1은 본 논문에서 사용하는 두 가지 프레임 구조를 보여주고, 입력 신호의 샘플링 주파수는 16 kHz이다. 음소 인식은 F1 프레임 단위로 특성벡터를 구하여 F1 프레임마다 한 번씩 인식 결과를 출력하고, 감정 인식은 F2 프레임 단위로 특성벡터를 구하여 F2 프레임마다 한 번씩 인식 결과를 출력한다. 동일한 특성의 특성벡터이지만 프레임 구조가 다르므로 실제 특성벡터 세부 내용은 다르고, 그럼에도 불구하고 두 프레임 구조의 특성벡터를 동시에 사용하여 하나의 공용 DNN을 학습시키고, 동일한 DNN 동작을 통하여 두 정보를 모두 인식한다.

표 1. 음소와 감정 인식을 위한 두 가지 프레임 구조
Table 1. Two frame structures for phoneme and emotion recognition

frame	frame length	hop length
F1	25 ms	10 ms
F2	100 ms	40 ms

그림 2는 본 논문에서 사용하는 4가지 DNN 구조를 보여준다. 각 층(layer)의 숫자는 층별 뉴런 수이고, 출력층의 뉴런 수는 각 정보 인식의 클래스 수와 같다. $S_{ref}(3)$과 $S_{ref}(4)$는 각 정보 인식별로 주어지는 개별 DNN이고 각각 3개와 4개의 은닉층을 가지며, 기존 기술에 따라 각 정보 인식기별로 독립적으로 학습하여 얻고, 이것의 성능이 제안하는 공용 DNN 성능의 평가 기준이 된다. $S_{prop}(3)$과 $S_{prop}(4)$는 음소와 감정 인식을 동시에 수행하는 공용 DNN이고, 출력층은 각 정보별로 분리되어 인식 결과를 출력한다. $S_{prop}(3)$에서 두 정보 인식기가 3개 은닉층을 모두 공유하고 마지막 은닉층과 출력층 사이의 연결 파라미터만 달리하며, 이 때의 성능을 $S_{ref}(3)$과 비교한다. $S_{prop}(4)$는 4개의 은닉층을 가지며 마지막 은닉층이 정보 인식기별로 분리된 구조이고,

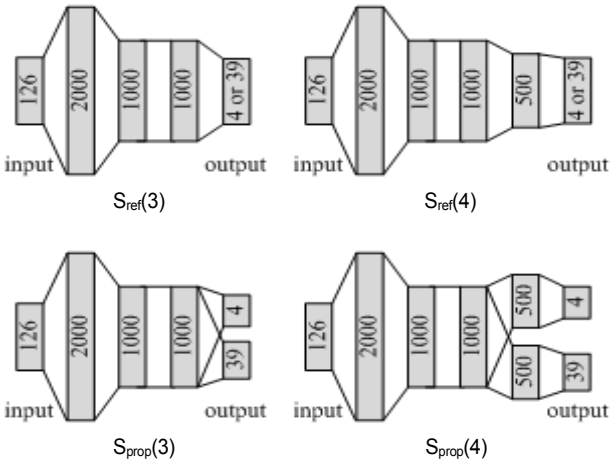

그림 2. 음소와 감정 인식을 위한 DNN 구조
Fig. 2. DNN structures for phoneme and emotion recognition

$S_{prop}(3)$에 비하여 정보 인식기별 독립 동작을 더 포함하고 있으며 이 때의 성능을 $S_{ref}(4)$와 비교한다. 모든 DNN에서 $S_{prop}(4)$의 출력층을 제외한 모든 층은 완전 접속(fully connected) 구조를 가지고, $S_{prop}(4)$의 출력층은 마지막 은닉층의 각 분할층과 완전 접속된다.

기존의 개별 DNN 학습은 각 정보별로 레이블링 되어 있는 학습 데이터 세트를 사용하여 수행한다. 그러나 본 논문에서 요구하는 음소와 감정이 동시에 레이블링 되어 있는 개방형 음성 데이터 세트는 없다. 따라서 본 논문에서는 기존에 사용하는 각 정보 인식별 학습 데이터 세트를 그대로 사용하고, 공용 DNN이 두 정보를 모두 인식하도록 학습시키는 방법을 개발한다.

제안하는 공용 DNN을 위한 다중 학습 방법의 핵심은 배치(batch) 단위로 진행하는 각 정보 인식별 교차 학습이고, 그림 3이 $S_{prop}(3)$에 대한 다중 학습 구조를 보여준다. 첫 단계로, 음소 인식을 위한 학습 데이터 세트에서 F1 프레임 단위로 특성벡터 $\boldsymbol{X}_P$를 구하고, 감정 인식용 학습 데이터 세트에서 F2 프레임 단위로 특성벡터 $\boldsymbol{X}_E$를 구한다. 각 특성벡터를 중복되지 않으며 랜덤하게 선택하여 배치를 구성하고, 한 번의 배치 단위 업데이트에서 활용하는 신호 길이를 동일하게 하기 위하여 음소 인식기 학습의 배치 크기는 4096으로 하고 감정 인식기 학습의 배치 크기는 1024로 한다.

$S_{prop}(3)$의 네트워크 파라미터를 음소와 감정 인식기에 대하여 번갈아 업데이트한다. 이 과정에서 각 인식기는 $S_{ref}(3)$ 구조의 인식기 학습에서와 동일한 비용 함수를 사용한다. 한 배치에 해당하는 4096개의 $\boldsymbol{X}_P$를 $S_{prop}(3)$에 입력하고, 음소 인식에 대한 비용 함수만을 바탕으로 파라미터를 업데이트한다. 이 때, 감정 인식의 출력은 구할 필요 없고, 마지막 은닉층과 감정 인식 출력층 사이의 파라미터는 업데이트 하지 않는다. 다음, 업데이트된 파라미터가 적용된 $S_{prop}(3)$에 한 배치에 해당하는 1024개의 $\boldsymbol{X}_E$를 입력하고, 감정 인식에 대한 비용 함수만을 사용하여 파라미터를 업데이트하고, 마지막 은닉층과 음소 인식 출력층 사이의 파라미터는 업데이트 하지 않는다. 다시, 업데이트된 파라미터

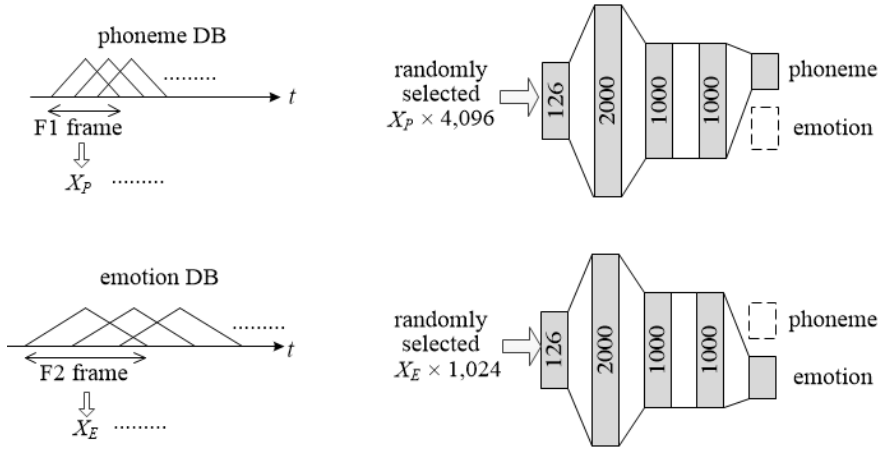

그림 3. 제안하는 공용 DNN의 다중 학습 방법
Fig. 3. Method of proposed multi-task training for common DNN

가 적용된 $S_{prop}(3)$에 4096개의 새로운 $\boldsymbol{X}_P$를 입력하고 음소 인식 성능을 기준으로 파라미트를 업데이트 한다. 위의 과정을 반복하면 마지막 은닉층과 출력층 사이의 파라미터를 제외한 모든 파라미터는 두 정보 인식에 공통적으로 학습되고, 궁극적으로 음소와 감정 인식 사이에 일반화된 학습 결과를 가지는 공용 $S_{prop}(3)$을 얻는다. 음소와 감정 인식기를 위한 두 학습 데이터 세트의 시간 길이를 동일하게 맞추었고, 각 학습 데이터 세트는 동일한 수의 배치를 가진다. 따라서 위 과정을 각 학습 데이터 세트의 모든 배치에 교차 적용하여 한 epoch에 대한 학습을 진행하고, 이 때 각 정보 인식기는 동일한 수의 파라미터 업데이트를 수행한다.

$S_{prop}(4)$에 대한 교차 학습도 동일한 방식으로 진행된다. 4096개의 $\boldsymbol{X}_P$를 $S_{prop}(4)$에 입력하고 음소 인식에 대한 비용 함수에 따라 파라미터를 업데이트 한다. 이 때, 정보별로 분리된 마지막 은닉층과 출력층에서 감정 인식과 관련된 모든 뉴런을 제거한 DNN을 학습시킨다. 다음, 1024개의 $\boldsymbol{X}_E$를 $S_{prop}(4)$에 입력하고, 마지막 은닉층과 출력층에서 음소 인식과 관련된 모든 뉴런을 제거한 DNN을 감정 인식에 대한 비용 함수에 따라 학습시킨다. 이 과정을 반복하여 두 정보 인식 사이에 일반화된 학습 결과를 가지는 $S_{prop}(4)$를 구한다.

이와 같이 각 정보 인식기별로 번갈아 학습을 할 때, 각 정보 인식이 요구하는 네트워크 파라미터의 목표점이 다르므로 각 목표점 사이에서 무의미한 파라미터 업데이트가 반복될 수 있다. 예로, 두 정보 인식이 요구하는 파라미터 업데이트 방향이 서로 반대이면, 교차 파라미터 업데이트를 진행할 때 실제로는 제자리에 머무르게 될 수 있다. 그림 4는 $S_{ref}(3)$을 각 정보 인식기별로 개별 학습할 때와 $S_{prop}(3)$을 두 정보 인식에 대하여 교차 학습 할 때의 각 정보 인식기별 학습 곡선이다. 과적합을 방지하기 위한 학습 종료 조건을 적용한 결과이며, 개별 DNN과 공용 DNN 사이에는 학습 특성의 차이가 없으며, 공용 DNN을 교차 학습을 하여도 지속적으로 의미 있는 학습이 진행된다. 즉, 네트워크 파라미터 공간의 각 위치에서 두 정보의 인식 성능 사이에 극단적 차이는 없고, 교차 학습을 수행하면 한 인식기만 학습시키는 결과는 발생하지 않고, 각 인식기에 대한 최적은 아니지만 두 인식기 사이에 균형을 가지는 일반화된 학습이 가능하여 성능 향상을 얻을 수 있다. 만일 학습 종료 조건을 지나 추가 학습을 진행하여도 $S_{ref}(3)$와 $S_{prop}(3)$의 학습 사이에는 의미 있는 차이가 나타나지 않는 것을 확인하였다.

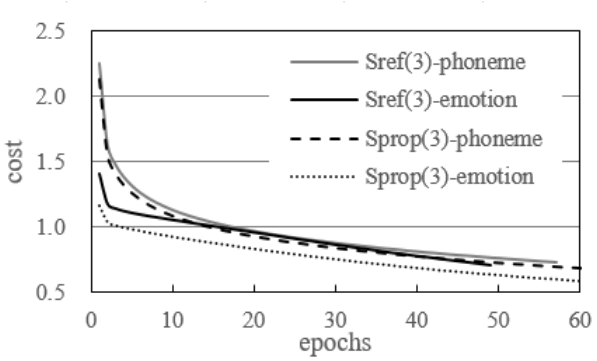

그림 4. MFCC 기반의 특성벡터를 사용할 때 $S_{ref}(3)$과 $S_{prop}(3)$의 학습 곡선
Fig. 4. Learning curve of $S_{ref}(3)$ and $S_{prop}(3)$ when using MFCC-based feature vector

Ⅲ. 성능 평가

M-spec과 MFCC 기반의 두 가지 특성벡터에 대하여 각 DNN의 인식 성능을 측정하고, 공용 DNN의 다중 학습에 의한 성능 향상과 교차 학습 과정에서 가중치 적용을 통한 정보 인식기 사이의 성능 조정을 검증하였다. 감정 인식은 IEMOCAP 데이터 세트를 사용한다[15]. 본 논문에서는 happy, sad, angry, neutral 등 총 4가지 감정으로 분류하고, 데이터 세트 분할은 학습, 검증, 평가를 6 : 2 : 2 비율로 한다. 음소 인식은 TIMIT 데이터 세트를 사용하고[16], 총 39가지 음소로 분류한다. TIMIT에서 주어지는 train set, development set, core test set을 각각 학습, 검증, 평가 데이터로 사용한다.

모든 DNN에서 은닉층은 ReLU, 출력층은 softmax 활성화 함수를 사용하고, 모든 인식기의 학습 방법을 통일시켜 동일한 조건에서 각 DNN의 성능을 비교하였다. He 초기화를 사용하고[17], 일반화 향상을 위해 확률 0.5의 drop-out을 모든 은닉층에 적용하였다[8]. 모든 인식기는 교차 엔트로피(cross entropy) 비용 함수를 사용하고, Adam 최적화를 사용하여 학습하고[18], 각 정보별로 학습률과 학습 종료 등의 모든 조건을 동일하게 하였다. 동일한 검증 데이터에 대하

여 조기 종료 (early stopping)의 patience를 10으로 하였고, 그 결과 학습의 epoch 수는 $S_{ref}(3)$의 감정 인식과 음소 인식에서 각각 49와 57이고, 두 인식을 동시에 수행하는 $S_{prop}(3)$에서 63이다. 두 가지 시간 해상도의 특성벡터를 사용하므로 교차 학습에서 각 정보별 배치 크기를 조정하여 각 정보에 대한 업데이트 시간 빈도를 통일시켰다. 또한, IEMOCAP과 TIMIT 데이터 세트의 길이가 다르므로, 동일한 조건의 학습을 위해 두 학습 데이터 세트의 길이를 통일시켰다. 따라서 각 학습 데이터 세트가 포함하는 배치의 수는 동일하고, 한 epoch에 대한 교차 학습을 진행하면 각 정보 인식기에 대하여 동일한 수의 파라미터 업데이트를 실시하고, 최종적으로 동일한 epoch 수의 학습을 실시하면 동일한 횟수의 업데이트를 실시하고 두 인식기는 동일한 조건의 학습을 하게 된다.

학습된 DNN을 사용하여 각 정보 인식 성능을 측정하는 과정은 다음과 같다. 음소 인식용 평가 데이터에서 F1 프레임 단위로 특성벡터를 구하여 각 DNN에 입력하고 F1 프레임마다 음소를 인식하여 성능을 구한다. 감정 인식용 평가 데이터에서 F2 단위로 특성벡터를 구하여 각 DNN에 입력하여 F2 프레임마다 출력층 softmax 값을 구하고, 각 파일 단위로 프레임 softmax 합을 기준으로 정보를 인식하여 성능을 구한다. $S_{prop}(3)$과 $S_{prop}(4)$는 각각 하나의 고정된 동작으로 두 정보 인식을 모두 수행할 때의 성능에 해당하며, 인간의 청각기관이 짧은 구간 단위로 신호를 분석하고 이와 동시에 긴 구간 단위로 신호를 분석하여 모두 뇌에 전달하고 각 정보를 인식하는 것과 동일한 동작이다.

표 2는 각 특성벡터와 DNN별로 감정과 음소 인식의 평균 정확도(mean accuracy)를 보여준다. S_{ref} 성능은 기존 기술에 따라 각 정보 인식기별로 DNN을 개별 학습할 때의 성능으로서 제안한 S_{prop} 성능 비교의 기준이고, S_{prop} 성능에서 괄호의 수는 기준 성능 대비 성능 향상을 %p로 표기한 것이다. S_{prop}을 사용하면 감정 인식 성능은 향상되는 경향을 가지고 음소 인식 성능은 하락하며, M-spec과 MFCC 사이에 뚜렷한 차별성은 없다. S_{prop}에서 감정 인식 성능이 향상된 이유는 교차 구조의 다중 학습에서 감정과 음소 인식을 동시에 고려하여 일반화된 학습을 하였기 때문에 감정 인식만으로 학습한 S_{ref}에서 발생하는 과적합이 감소하였기 때문으로 판단된다. 반면, 음소 인식은 감정 인식에 비하여 분류 클래스가 많아 상대적으로 많은 학습이 필요하며, 감정 인식과 동일한 조건의 학습으로 인하여 감정 인식에 비해 충분한 학습이 안 되어 성능이 저하된 것으로 추정된다. 즉, S_{prop}에서 두 정보 인식 사이에 균형 있는 학습이 안 되고, 감정 인식 방향으로 치우친 학습 결과를 얻는다.

표 2. 각 특성벡터와 DNN에서의 감정과 음소 인식의 평균 정확도(%)
Table 2. Mean accuracy(%) of emotion and phoneme recognition for each feature vector and network

feature / network	Mel-spectrogram		MFCC	
	emotion	phoneme	emotion	phoneme
$S_{ref}(3)$	66.30	70.97	68.30	69.73
$S_{prop}(3)$	67.48 (+1.18)	70.12 (-0.85)	68.03 (-0.27)	68.62 (-1.11)
$wS_{prop}(3)$	69.29 (+2.99)	70.41 (-0.56)	69.11 (+0.81)	69.56 (-0.17)
$S_{ref}(4)$	66.58	71.28	66.39	70.36
$S_{prop}(4)$	67.84 (+1.26)	70.28 (-1.00)	68.48 (+2.09)	69.15 (-1.21)
$wS_{prop}(4)$	69.20 (+2.62)	71.28 (0.00)	68.57 (+2.18)	70.42 (+0.06)

교차 구조의 다중 학습에서 정보 사이의 균형 있는 학습을 위해 학습 과정에서 가중치를 적용하는 방법을 시도하였다. S_{prop}의 성능 분석에 의하면 다중 학습 과정에서 감정 인식 학습에는 이미 충분한 일반화가 진행되어 성능이 향상되었으므로, 음소 인식 학습에서 감정 인식 학습에 비하여 더 강한 일반화가 발생하도록 차별적 학습을 하면 음소 인식 성능이 향상될 것이다. 이를 위하여 교차 학습 과정에서 음소 인식을 위한 학습 단계에서는 확률 0.5의 drop-out을 적용하고 감정 인식을 위한 학습 단계에서는 drop-out을 적용하지 않는다. 이렇게 학습된 DNN을 wS_{prop}이라 하며, 표 2에서 보듯이 모든 경우에서 wS_{prop}의 음소 인식 성능이 S_{prop}보다 향상되며, 감정 인식 성능도 S_{prop}에 비하여 더 향상되었다. 따라서 차별적 drop-out 적용을 통하여 두 정보 사이의 균형 있는 학습이 진행되어 일반화 성질이 더욱 향상된 학습 결과를 얻었으며, 두 정보 인식의 성능이 S_{prop} 대비 모두 증가하는 것을 확인하였다. $S_{prop}(4)$와 $wS_{prop}(4)$의 성능은 각각 $S_{prop}(3)$과 $wS_{prop}(3)$보다 우수한 경향을 보이며, 은닉층이 추가되어 모델링 능력이 향상되고 4번째 은닉층이 정보별로 분리되어 정보별 독립 동작이 포

함되었기 때문이다.

결론적으로, 각 특성벡터에 대하여 $wS_{prop}(4)$는 각 정보별로 $S_{ref}(4)$에 비하여 동등 이상의 인식 성능을 가진다. 만일 음소와 감정 인식의 평균 성능을 비교하면, $S_{ref}(3)$-MFCC 조합을 제외한 모든 경우에서 공용 DNN이 개별 DNN 보다 우수한 평균 성능을 가진다. 동일한 수의 은닉층을 가지는 S_{ref}, S_{prop}, wS_{prop}는 모두 동일한 수의 뉴런을 사용하여 각 정보를 인식한다. 따라서 공용 DNN은 신경망 자원의 증가 없이 두 가지 정보를 모두 인식하는 기능을 가지며, 이와 같은 기능 향상뿐만 아니라 제안하는 다중 학습을 통하여 기존의 개별 DNN보다 향상된 평균 성능을 제공할 수 있다. 따라서 본 논문의 기술을 적용하면 기존 기술에 따른 특성벡터와 DNN을 그대로 사용하면서 두 정보 인식을 모두 수행하는 기능을 제공하고 두 정보의 평균 인식 성능을 향상시킬 수 있는 것을 확인하였다.

Ⅳ. 결 론

본 논문에서는 기계학습 기반의 음소와 감정 인식의 성능 향상을 위하여 학습에서의 일반화를 향상시키는 새로운 방법을 제안하였다. 하나의 공용 인식기를 사용하여 두 정보를 모두 인식하고, 두 정보 인식에 대한 교차 구조의 다중 학습을 통하여 두 정보 사이의 균형 있는 학습을 진행시키고 일반화가 증가된 학습 결과를 얻는다. 이 때, 두 정보는 서로 다른 시간 해상도의 특성벡터를 사용하며, 공용 인식기는 특성벡터를 구분하지 않고 항상 동일한 동작을 수행하여 두 정보를 인식한다. 또한, 다중 학습 과정에서 음소 인식에 가중치를 부여하여 음소 인식에서의 일반화를 추가로 향상시켰다. 동일한 특성벡터와 동일한 DNN을 사용할 때, 제안한 공용 인식기의 평균 성능이 기존 기술에 따라 각 정보별로 학습한 개별 인식기에 비하여 우수한 것을 확인하였다.

참 고 문 헌 (References)

[1] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," *Proc. on IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 6645-6649, May 2013, doi:10.1109/ICASSP.2013.6638947.

[2] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech communication*, vol. 41, no. 4, pp. 603-623, Nov. 2003, doi:10.1016/S0167-6393(03)00099-2.

[3] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437-1462, Sept. 1997, doi:10.1109/5.628714.

[4] W. J. Jang, H. W. Yun, S. H. Shin, H. J. Cho, W. Jang, and H. Park, "Music genre classification using spikegram and deep neural network," *J. of Broadcast Engineering*, vol. 22, no. 6, pp. 693-701, Nov. 2017, doi:10.5909/JBE.2017.22.6.693.

[5] S. H. Shin, H. W. Yun, W. J. Jang, and H. Park, "Extraction of acoustic features based on auditory spike code and its application to music genre classification," *IET Signal Processing*, vol. 13, no. 2, pp. 230-234, Apr. 2019, doi:10.1049/iet-spr.2018.5158.

[6] S. Han, J. Kim, S. An, S. Shin, and H. Park, "Speech feature extraction based on spikegram for phoneme recognition," *J. of Broadcast Engineering*, vol. 24, no. 5, pp. 735-742, Sept. 2019, doi:10.5909/JBE.2019.24.5.735.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Cambridge and London, 2016.

[8] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929-1958, Jan 2014, doi:10.5555/2627435.2670313.

[9] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp.41-75, 1997, doi:10.1023/A:1007379606734.

[10] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," *Proc. on Int. Joint Conf. on Neural Networks*, pp. 1-7, July 2015, doi:10.1109/IJCNN.2015.7280624.

[11] S. J. Pan, and Q. Yang, "A survey on transfer learning," *IEEE Trans. on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, Oct. 2009, doi:10.1109/TKDE.2009.191.

[12] B. Logan, "Mel frequency cepstral coefficients for music modeling," *ISMIR*, vol. 270, pp. 1-11, Oct. 2000.

[13] ETSI, *Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Extended front-end feature extraction algorithm; Compression algorithm; Back-end speech reconstruction algorithm*, ETSI ES 202 211, v1.1.1, Nov. 2003.

[14] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall, pp. 423-424, 2001.

[15] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J, N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335, Dec. 2008, doi:10.1007/s10579-008-9076-6.

[16] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351-356, Aug. 1990, doi:10.1016/0167-6393(90)90010-7.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *Proc. on IEEE Int. Conf. on Computer Vision*, pp. 1026-1034, 2015, doi:10.1109/iccv.2015.123.
[18] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, Dec. 2014.

저 자 소 개

김 재 원

- 2019년 2월 : 광운대학교 전가공학과 학사
- 2019년 3월 ~ 현재 : 광운대학교 전자공학과 석박통합과정
- ORCID : https://orcid.org/0000-0002-6496-842X
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

박 호 종

- 1986년 2월 : 서울대학교 전자공학과 공학사
- 1987년 12월 : Univ. of Wisconsin-Madison 공학석사
- 1993년 5월 : Univ. of Wisconsin-Madison 공학박사
- 1993년 9월 ~ 1997년 8월 : 삼성전자 선임연구원
- 1997년 9월 ~ 현재 : 광운대학교 전자공학과 교수
- ORCID : https://orcid.org/0000-0003-1600-6610
- 주관심분야 : 오디오/음성 신호처리, 3D 오디오, 음악정보처리