한국음향학회지 제36권 제6호 pp. 407~412 (2017)
The Journal of the Acoustical Society of Korea Vol.36, No.6 (2017)
https://doi.org/10.7776/ASK.2017.36.6.407

pISSN : 1225-4428
eISSN : 2287-3775

주목 메커니즘 기반의 심층신경망을 이용한 음성 감정인식

Speech emotion recognition using attention mechanism-based deep neural networks

고상선,[1] 조혜승,[1] 김형국[1†]

(Sang-Sun Ko,[1] Hye-Seung Cho,[1] and Hyoung-Gook Kim[1†])

[1]광운대학교 전파공학과

(Received June 28, 2017; revised July 21, 2017; accepted November 29, 2017)

초 록: 본 논문에서는 주목 메커니즘 기반의 심층 신경망을 사용한 음성 감정인식 방법을 제안한다. 제안하는 방식은 CNN(Convolution Neural Networks), GRU(Gated Recurrent Unit), DNN(Deep Neural Networks)의 결합으로 이루어진 심층 신경망 구조와 주목 메커니즘으로 구성된다. 음성의 스펙트로그램에는 감정에 따른 특징적인 패턴이 포함되어 있으므로 제안하는 방식에서는 일반적인 CNN에서 컨벌루션 필터를 tuned Gabor 필터로 사용하는 GCNN(Gabor CNN)을 사용하여 패턴을 효과적으로 모델링한다. 또한 CNN과 FC(Fully-Connected)레이어 기반의 주목 메커니즘을 적용하여 추출된 특징의 맥락 정보를 고려한 주목 가중치를 구해 감정인식에 사용한다. 본 논문에서 제안하는 방식의 검증을 위해 6가지 감정에 대해 인식 실험을 진행하였다. 실험 결과, 제안한 방식이 음성 감정인식에서 기존의 방식보다 더 높은 성능을 보였다.
핵심용어: 음성 감정인식, 주목 메커니즘, 심층 신경망, 가버 필터

ABSTRACT: In this paper, we propose a speech emotion recognition method using a deep neural network based on the attention mechanism. The proposed method consists of a combination of CNN (Convolution Neural Networks), GRU (Gated Recurrent Unit), DNN (Deep Neural Networks) and attention mechanism. The spectrogram of the speech signal contains characteristic patterns according to the emotion. Therefore, we modeled characteristic patterns according to the emotion by applying the tuned Gabor filters as convolutional filter of typical CNN. In addition, we applied the attention mechanism with CNN and FC (Fully-Connected) layer to obtain the attention weight by considering context information of extracted features and used it for emotion recognition. To verify the proposed method, we conducted emotion recognition experiments on six emotions. The experimental results show that the proposed method achieves higher performance in speech emotion recognition than the conventional methods.
Keywords: Speech emotion recognition, Attention mechanism, Deep neural networks, Gabor filters

PACS numbers: 43.60.Bf, 43.72.Bs

I. 서 론

최근 마이크로 폰 센서가 탑재된 스마트 폰의 보급으로 인해 사용자의 음성 데이터의 수집이 용이해짐에 따라 음성 감정 인식에 대한 연구가 활발해지고 있다. 이와 함께 신경망 구조를 이용한 멀티미디어 인식에 대한 높은 성능이 보고되고 있으며 음성 감정인식에 대한 많은 연구들이 신경망 구조를 활용하여 진행되고 있다.

Mao *et al.*[1]은 음성 신호에 CNN(Convolution Neural Networks)을 적용하여 감정의 정보를 효과적으로 학습함을 보였다. 하지만 CNN만으로는 음성과 같은 시계열 데이터가 포함하는 시간적 흐름에 따른 정보를 고려하지 못한다는 한계가 있었다. 이에 신호의 시간적 속성을 고려해 학습하는 RNN(Recurrent Neural Netowkrs)

†**Corresponding author:** Hyoung-Gook Kim (hkim@kw.ac.kr)
Department of Radio Sciences and Engineering, Kwangwoon University, 20 Gwangun-ro, Nowon-gu, Seoul 01897, Republic of Korea
(Tel: 82-2-940-5574, Fax: 82-2-913-5006)

기반의 방식들이 적용되기 시작하였고 그중에서도 LSTM(Long-Short Term Memory) 기반의 음성 감정인식 방식은 현재 관련 분야에서 기존의 방식들보다 높은 성능을 보이고 있다. 최근에는 CNN과 LSTM이 가지는 각각의 장점을 결합한 방식이 제안되고 있으며, 결합된 CNN과 LSTM으로부터 추출된 특징 벡터를 DNN(Deep Neural Networks)에 적용한 CLDNN방식으로 인식 성능을 향상시키는 시도가 이루어지고 있다.[2] 이와 함께 주목 메커니즘을 결합한 방식이 제안되고 있다.

주목 메커니즘이란 사람이 물체를 인식할 때 배경을 포함한 모든 정보를 사용하는 것이 아니라 특징적인 부분에 집중하는 점에 착안한 방식이다. 초기 주목 메커니즘은 신경망 기반 이미지 처리 분야에서 비교적 중요한 정보를 담고 있는 특정 부분에 가중치를 부여하여 이미지를 효과적으로 분석하기 위해 사용되었다. 최근에는 점차 그 적용 분야가 확장되어 인공 신경망 번역이나 음성 감정 인식에 적용하려는 연구가 시도되고 있다.[3]

이에 본 논문에서는 CNN, GRU(Gated Recurrent Unit), DNN을 결합한 심층 신경망 방식에 주목 메커니즘을 적용한 음성 감정인식 방식을 제안한다. 제안하는 방식에서는 일반적인 CNN 대신 Gabor 필터를 적용하는 GCNN[4]을 사용한다. 기존의 GCNN과는 다르게 본 논문에서는 음성 신호의 감정에 따른 스펙트로그램에서 나타나는 특징적인 패턴에 따라 설정된 tuned Gabor 필터를 적용하여 감정 인식에 효과적인 특징을 추출한다. 또한 제안하는 방식에서는 LSTM을 사용하는 기존의 방식과 달리 GRU를 사용한다. GRU는 LSTM과 비교해 상대적으로 내부 구조가 단순하여 연산량이 적고 과적합이 덜 일어나는 장점이 있다. 마지막으로, 주목 메커니즘을 통해 음성 신호에서 감정 정보가 많이 포함된 부분에 대한 가중치를 계산하여 더욱 효과적인 감정 인식이 가능하도록 한다. 이때, 제안하는 방식에서는 일반적인 순환 신경망 기반의 주목 메커니즘이 아닌 컨벌루션 레이어와 FC(Fully-Connected) 레이어로 구성된 주목 메커니즘을 적용한다.

II. 음성 감정인식 방식

Fig. 1은 본 논문에서 제안하는 음성 감정인식 방식의 전체 구조를 나타낸다. 본 방식은 Log-Mel 에너지 스펙트로그램 추출, GC(Gabor-Convolutional) 레이어, GRU 레이어로 구성되는 특징값 추출 모듈과 컨벌루션 레이어, FC 레이어로 구성되는 주목 메커니즘 모듈, 그리고 FC 레이어로 구성되는 분류 모듈로 이루어져 있다.

먼저 발화단위의 음성 신호가 입력되면 이에 대해 Log-Mel 에너지 스펙트로그램이 추출되고, 이는 GC 레이어로 입력된다. 이때 현재 시간 t에 대해 왼쪽으로 l개 프레임과 오른쪽으로 r개 프레임을 연결한 특징 시퀀스 $\mathbf{x} = [\mathbf{x}_{t-r}, ..., \mathbf{x}_t, ..., \mathbf{x}_{t+r}]$, $\mathbf{x}_t \in \mathbb{R}^F$가 입력된다. 여기서 F는 Log-Mel 에너지 스펙트로그램의 차원을 나타낸다.

GC 레이어는 일반적인 컨벌루션 레이어의 변형된 형태로, 컨벌루션 필터가 tuned Gabor 필터로 초기화되어 입력 특징 맵에 적용된다. tuned Gabor 필터는 감정에 따라 각도가 설정된 Gabor 필터를 의미하며 이를 통해 기존의 컨벌루션 레이어 보다 효과적으로 감정 인식에 특화된 정보를 추출할 수 있다.

GC 레이어로부터 출력된 특징 벡터는 GRU 레이어로 입력된다. GRU는 중요한 정보가 들어올 때마다 업데이트 게이트에서 현재 상태를 갱신하고 리셋 게이트에서 현재 상태를 삭제하면서 과거의 정보를 선택적으로 반영하므로 스펙트로그램의 시간적인 변화를 효과적으로 모델링 할 수 있다.

GRU 레이어를 통해 특징 벡터 $\mathbf{h} = [\mathbf{h}_1, ..., \mathbf{h}_T]$, $\mathbf{h}_t \in \mathbb{R}^D$가 출력된다. 여기서 T와 D는 각각 출력 특징 벡터의 개수와 차원을 나타낸다. 출력 특징 벡터 $\mathbf{h}$는 주목 메커니즘 모듈로 입력된다.

음성 신호에는 감정이 드러나는 강도가 강한 부분

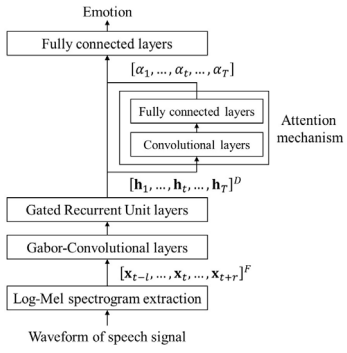

Fig. 1. Scheme of the proposed speech emotion recognition method.

Table 1. Primary patterns of the slopes of the early and late segment of six vocal emotions.

Emotion	Early segment	Late segment
Anger	Fast downward	Slow downward
Happy	Very fast upward	Fast downward
Fear	Fast upward	Downward
Sadness	Slow downward	Slow downward
Surprise	Very fast upward	Slow downward
Disgust	Horizontal	Upward

Table 2. Secondary patterns of the slopes of the early and late segment of six vocal emotions.

Emotion	Early segment	Late segment
Anger	Slow downward	Horizontal
Happy	Upward	Downward
Fear	Horizontal	Slow downward
Sadness	Upward	Slow downward
Surprise	Horizontal	Downward
Disgust	Slow downward	Slow downward

과 약한 부분이 혼재한다. 주목 메커니즘은 앞서 추출된 특징벡터에 대해 감정적으로 중요한 부분에 가중치를 적용하여 집중적으로 보기 위해 사용된다. 이때, 본 논문에서는 일반적인 순환 신경망 기반의 주목메커니즘이 아닌 컨벌루션 레이어와 FC레이어로 구성되는 주목 메커니즘을 사용한다. 먼저 GRU 레이어를 통과한 특징 벡터 $\mathbf{h}$ 가 컨벌루션 레이어로 입력된다. 이를 통해 특징 벡터가 가진 세부적인 정보는 지우고 전체의 큰 윤곽을 관찰함으로써 주목할 만한 정보를 얻을 수 있다. 컨벌루션 레이어를 통과한 특징벡터는 FC레이어로 입력되어 주목 가중치 파라미터 α_t가 계산된다.

다음으로, 주목 가중치 파라미터 α_t가 앞서 출력한 특징벡터 $\mathbf{h}$ 에 적용되어 가중치가 부여된 특징벡터가 계산된다. 최종적으로, 가중치가 부여된 특징벡터가 분류기 역할을 하는 다층의 FC 레이어로 입력된다. FC 레이어는 각 특징벡터를 감정 클래스에 효과적으로 매핑되도록 하며 마지막 FC 레이어에서는 softmax 함수를 통해 입력 특징에 대한 각 클래스들의 예측 확률이 출력되고 가장 높은 확률을 가진 감정 클래스로 인식된다.

2.1 Gabor-Convolutional 레이어

본 논문에서는 추출한 Log-Mel 에너지 스펙트로그램을 GCNN에 입력한다. GCNN은 기존의 방식과는 다르게 컨벌루션 레이어에서 랜덤 초기화된 사각형의 필터 대신 Gabor 필터를 사용한다.

Gabor 필터는 파라미터 조정을 통해 여러 가지 변형과 회전이 가능하기 때문에 이를 기반으로 특징의 다양한 방향을 찾는데 유용하다. Chang과 Morgan[4]은 이러한 Gabor 필터의 특성을 기반으로 음성 스펙트로그램의 특징적 패턴을 더 효과적으로 추출하기 위해 Gabor 필터를 컨벌루션 레이어에 적용한 GCNN 구조를 제안하였다. 해당 구조는 모든 방향에 대한 필터링을 수행하기 위해 360°를 일정 간격으로 분할하여 설정한 59개의 Gabor 필터뱅크를 사용하였으며 이는 음성 인식에 대해 일반적인 CNN보다 향상된 성능을 보였다.

하지만 음성 신호를 자세히 관찰한 결과, 같은 문장이라도 스펙트로그램이 감정에 따라 다른 패턴을 보이는 것을 확인하였다. 이에 본 논문에서는 감정별 스펙트로그램의 특징적 패턴을 찾기 위해 감정에 따라 미리 설정된 tuned Gabor 필터를 사용한다.

감정에 따른 음성 스펙트로그램의 특징적 패턴은 문장 전체가 아닌 일부분에서 나타나므로 감정이 가장 뚜렷하게 나타나는 패턴을 포함하는 구간을 주요 패턴, 부차적으로 나타나는 패턴을 포함하는 구간을 부가 패턴이라고 정의한다. Fig. 2는 영어 문장 “Will you tell me why?”에 대한 fear, anger의 2가지 감정의 발화 스펙트로그램과 제안한 방식을 통해 출력되는 주목 가중치 파라미터를 나타낸다.

Fig. 2(a)의 fear 감정의 스펙트로그램에서 주요 패턴의 초반부는 매우 빠르게 상승하다가 후반부에는 천천히 하강하는 형태를 보이며, 부가 패턴의 초반

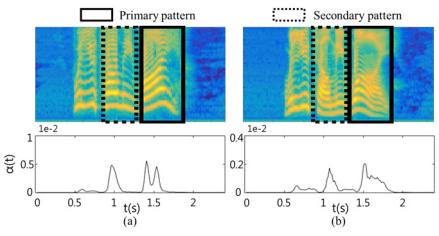

Fig. 2. (a) Fear, (b) Anger. Top: spectrogram of sentence, bottom: attention weight $\alpha(t)$ over time.

부에는 수평적으로 진행되다가 후반부에는 보통의 속도로 하강한다. 또한 특징적 패턴의 부근에서 주목 가중치 파라미터가 상대적으로 높게 나타남을 확인할 수 있다. 이러한 감정 별 스펙트로그램의 특징 패턴을 정의하여 Tables 1, 2에 나타내었다.

또한 Fig. 2의 아래 그림에서 각 신호에 따라 출력된 주목 가중치 파라미터를 확인할 수 있다. 출력된 가중치는 신호의 묵음 구간에 대해 매우 작은 값을 보이는 한편 본 논문에서 지정한 특징적 패턴의 부근에서는 다른 부분에 비해 높은 값을 보인다. 이를 통해 감정 별 특징적 패턴이 나타나는 특징 벡터에 가중치를 둔 효과적인 분석이 가능하다.

본 논문에서는 총 6개의 감정을 분류하기 위해 12개의 tuned Gabor 필터를 설정하였다. Tuned Gabor 필터는 스펙트로그램 방향에 따라 상승과 하강을 각각 양수와 음수로, 각도에 따라 매우 빠른, 빠른, 보통, 느림을 각각 60°, 45°, 30°, 15°의 크기로 정의하였으며 수평은 0°로 정의하였다. GC 레이어에서, 입력된 Log-Mel 에너지 스펙트로그램에 12개의 tuned Gabor 필터가 적용되며 이를 통해 특징 맵이 생성된다. 생성된 특징 맵은 pooling 레이어를 통과하여 시간-주파수축에 대해 각각 차원이 축소된다. 본 과정을 통해 음성 스펙트로그램의 대표적인 특징 패턴을 더욱 간결하고 명확하게 추출할 수 있다.

2.2 주목 메커니즘

주목 메커니즘은 사람이 신호를 인식하는데 사용하는 주의 집중 효과를 반영한 방식이다. 음성 신호에 주목 메커니즘을 적용한 대표적인 분야로는 RNN 기반의 인공 신경망 번역이 있다.

기존의 인코더 디코더 구조의 인공 신경망 번역은 encoding된 입력 신호의 정보를 decoder로 전달할 때 신호의 시간 순서에 따른 중요도를 고려하지 않고 전달하므로, 이러한 비효율성을 개선하기 위해 주목 메커니즘을 적용한 방식이 제안되었다.[5] 해당 방식은 인코딩에서 출력된 시퀀스 벡터와 디코더의 시퀀스 벡터를 기반으로 각 시간 단위에서 주목 가중치를 계산하고, 이를 인코더의 출력 시퀀스 벡터에 적용한 가중 합 벡터인 컨텍스트 벡터를 구해 디코더에 반영하는 방식이다. 이때 주목 가중치는 인코더와 디코더의 시간 순서를 모두 고려하여 계산된다. 이는 주목 메커니즘을 이용해 음성 신호의 맥락에 따른 중요도를 계산하고 이를 반영함으로써 더욱 효율적인 번역을 수행하였다.

위 방식을 변형하여 Mirsamadi *et al*.[3]은 RNN과 주목 가중치를 이용한 음성 감정인식을 제안하였다. 해당 방식은 RNN 기반의 many-to-one 모델링을 이용한 방식으로 프레임 별 모델링 후 출력되는 시퀀스 $\boldsymbol{s}_t$에 대해 주목 가중치 α_t를 적용한 가중 합을 계산하여 발화 단위의 감정 분류를 수행한다. 발화 단위의 분류를 수행하므로 인공 신경망 번역 방식과는 다르게 출력 시간 순서를 고려하지 않으며 오로지 입력된 시퀀스 벡터의 중요도만 고려한다. 따라서 컨텍스트 벡터를 생성하지 않고 주목 가중치만을 계산하여 이를 RNN의 출력 시퀀스에 적용해 가중치가 부여된 벡터 $\boldsymbol{z}$를 출력한다. 이는 다음과 같이 나타낼 수 있다.

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k}^{T}\exp(e_k)}, \tag{1}$$

$$e_t = score(\boldsymbol{u}^T \boldsymbol{s}_t), \tag{2}$$

$$\boldsymbol{z} = \sum_{t=1}^{T} \alpha_t \mathbf{h}_t, \tag{3}$$

여기서 t는 입력 시퀀스 벡터의 시간 순서, $score(\bullet)$는 $\boldsymbol{s}_t$의 중요도를 계산하는 함수를 나타낸다. 위 방식에서는 주목 가중치 계산 시 주목 파라미터 벡터인 $\boldsymbol{u}$를 사용하였으며 음성 감정 인식에 있어 기존의 RNN 기반의 방식보다 향상된 성능을 보였다.

본 논문의 주목 메커니즘 방식은 앞서 설명한 방식과 비교해 두 가지 중요한 차이를 보인다. 본 방식에서는 RNN에서 출력되는 시퀀스 벡터가 아닌 CNN 기반의 출력 특징 벡터를 이용하며, 주목 가중치 파라미터를 FC 레이어를 이용하여 모델링한다.

특징 추출 모듈을 통해 출력된 특징 벡터 $\mathbf{h} \in \mathbb{R}^{T \times D}$는 주목 메커니즘 모듈로 입력된다. 이때 $\mathbf{h}$의 현재시간 t에 대해 τ개의 특징벡터를 연결한 시퀀스가 컨벌루션 레이어로 입력된다. 컨벌루션 레이어를 거치며 입

력된 벡터에 대해 중요한 정보가 집중적으로 분포한 위치를 더욱 뚜렷하게 나타내는 특징맵이 출력된다. 출력된 특징 맵은 평탄화되어 특징벡터 $\boldsymbol{s}$ 가 생성된다. $\boldsymbol{s}$ 에 대해 제안하는 방식에서는 Eq. (2)의 $score(\bullet)$로서 크기 (T, T)의 2단 FC 레이어를 사용한다.

FC레이어를 통해 출력되는 e_t를 바탕으로 Eq. (1)를 통해 주목 가중치 파라미터 α_t를 계산할 수 있다. 계산된 α_t는 Eq. (3)과 같이 앞서 출력된 특징벡터 $\mathbf{h}$ 에 적용되며 이로부터 가중치가 부여된 특징벡터 $\boldsymbol{z}$ 가 출력된다. 출력된 $\boldsymbol{z}$ 는 다시 다층의 FC 레이어로 입력하여 분류를 수행하기 위한 학습을 추가적으로 진행한 후 마지막 레이어를 통해 최종 감정인식 결과가 출력된다.

III. 실험

3.1 실험환경 및 측정방식

실험에서는 SAVEE 데이터베이스를 사용하였다.[6] 해당 데이터는 남성 네 명의 목소리로 구성되며 감정 클래스는 분노, 혐오감, 공포, 행복, 슬픔, 놀라움의 6가지로 구성된다. 감정마다 15개의 문장, 화자마다 총 90개의 영어 문장으로 녹음된 데이터로 구성되며 각 데이터는 평균 4 s의 길이, 모노, 44.1 kHz의 샘플링 레이트, 16 비트의 깊이로 구성되어 있다.

본 논문에서는 객관적 평가를 위해 잭나이프 기법을 사용했다. 데이터를 화자 별로 4 set으로 나누어 3 set은 학습 데이터로 1 set은 테스트 데이터로 성능 측정 실험을 진행했으며, 4회의 실험 결과의 평균치로 최종 결과를 나타내었다.

실험에 사용된 분류기는 GCNN, LSTM, GRU, DNN의 조합들로 사용하였으며 특징값은 40개의 멜 밴드, 10 ms의 윈도우를 사용하여 추출한 Log-Mel 에너지를 사용하였다. 실험에서 GCNN의 입력 특징으로는 현재 프레임에서 왼쪽으로 14프레임, 오른쪽으로 5프레임을 포함한 20개 프레임으로 구성된 20×40 크기의 특징 맵을 사용하였다. GCNN에서, tuned Gabor 필터의 크기는 7×7로, Max pooling은 2의 크기를 적용한다. 주목 모듈의 컨벌루션 레이어의 필터는 5x5 크기로, Max pooling은 5의 크기로 적용하였다. 활성 함수로는 ReLU(Rectified Linear Unit) 함수를, 특징 맵은 1개씩 추출하였다.

LSTM은 2개의 히든 레이어를 사용했으며 각 레이어의 뉴런은 128개를 사용하였다. 그리고 GRU는 3개의 히든 레이어를 사용했으며 각 레이어의 뉴런 수는 200개를 사용하였다. 최종 FC 레이어는 4개의 층을 사용하였으며 각각의 뉴런 수는 128, 32, 32, 7개로 사용하였다. FC 레이어 중 앞의 세 개 레이어는 ReLU 함수로 활성화 시켰으며 마지막 레이어는 softmax 함수를 사용하였다. 모든 신경망은 ASGD(Asynchronous Stochastic Gradient Descent)의 최적화 전략을 사용하여 교차 엔트로피를 기준으로 학습하였다.

3.2 실험 결과

Table 3은 실험 결과를 나타내며, Fig. 3은 제안한 방식에 대한 실험결과의 혼동행렬을 나타낸다. 실험결과는 Chang과 Morgan[4]이 제안한 PNS 특징 기반의 GCNN을 이용한 방식을 baseline으로 사용하여 단계별로 비교하였다. 또한 Mirsamadi *et al.*이 제안한 방식과의 비교 실험을 진행하였다.

Log-Mel 에너지를 특징을 바탕으로 제안한 GCNN으로 인식한 결과는 75.53 %으로 baseline과 비교하여 약 2.3 %가 향상되었다. GCNN-LSTM 방식의 감정 인식률은 LSTM 대신 GRU를 적용하여 분류한

Table 3. Experimental Results.

Method	Accuracy
Baseline	73.23 %
Log-Mel Energy + GCNN	75.53 %
Log-Mel Energy + GCNN-LSTM	77.65 %
Log-Mel Energy + GCNN-GRU	79.02 %
Log-Mel Energy + GCNN-GRU-DNN	79.76 %
RNN-weighted pool with attention[3]	80.35 %
Log-Mel Energy + GCNN-GRU-DNN + Attention mechanism	82.15 %

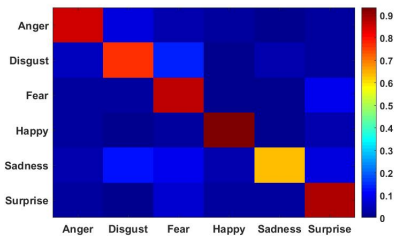

Fig. 3. Confusion matrix of the proposed method.

GCNN-GRU의 인식률은 약 1.37 % 정도 더 높게 나타났다. 또한 GCNN-GRU 구조에 DNN 구조를 추가한 결과는 DNN을 사용하기 전보다 0.74 % 정도 향상된 결과를 보였다. Mirsamadi *et al.*이 제안한 방식의 인식률은 80.35 %로 상대적으로 높게 나타났는데, 이는 해당 방식이 간단한 구조임에도 불구하고 다양한 저레벨 특징값을 사용했기 때문으로 생각된다. 최종적으로, 본 논문에서 제안한 방식인 GCNN-GRU-DNN에 주목 메커니즘을 적용한 구조의 인식률은 82.15 %로 비교한 방식들 중 가장 뛰어난 인식률을 나타내었다. 또한 Fig. 3을 보면, 행복, 놀라움의 감정의 인식률이 높은 반면 혐오감, 슬픔의 인식률은 상대적으로 낮게 나타난다. 이는 Tables 1, 2에서 보여지는 행복, 놀라움의 감정의 특징적 패턴이 다른 감정들보다 강하게 나타나는 반면, 혐오감, 슬픔의 특징적 패턴은 상대적으로 약하게 나타나는 동시에 패턴의 특성 자체가 일반적인 경우와 유사한 경우가 많기 때문에 나타나는 현상으로 보여진다.

IV. 결 론

본 논문에서는 주목 메커니즘 기반의 심층신경망을 이용한 음성 감정인식을 제안하였다. 제안한 방식에서는 tuned Gabor 필터를 사용하는 GCNN과 GRU, DNN을 결합하고 컨벌루션 레이어와 FC레이어로 구성된 주목 메커니즘을 적용하여 음성 신호에서 감정적으로 현저하게 변화하는 부분에 주목하여 학습을 진행할 수 있었다. 또한 실험에서는 6가지 감정에 대한 인식을 통해 제안한 방식이 기존 방식에 비해 성능 향상을 이루어냄을 확인하였다. 향후 연구에서는 본 논문에서 제안한 주목 메커니즘을 기반으로 신호의 특징적인 부분을 스스로 관찰하여 더욱 효과적으로 학습하는 방식에 대한 연구를 진행할 예정이다.

감사의 글

본 논문은 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2015R1D1A1A01059804).

References

1. Q. Mao, M. Dong, Z. Huang, and Y. Zhan, “Learning salient features for speech emotion recognition using convolutional neural networks,” IEEE Trans. Multimedia, **16**, 2203-2213 (2014).
2. T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, “Convolutional, long short-term memory, fully connected deep neural networks.” in IEEE ICASSP, 4580-4584 (2015).
3. S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” in IEEE ICASSP, 2227-2231 (2017).
4. S. Y. Chang and N. Morgan, “Robust CNN-based speech recognition with gabor filter kernels,” in Interspeech, 905-909 (2014).
5. D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv:1409.0473 (2014).
6. S. Haq and P. J. B. Jackson, “Speaker-dependent audio-visual emotion recognition,” in AVSP, 53-58 (2009).

저자 약력

▸ 고 상 선 (Sang-Sun Ko)

2015년: 광운대학교 전자융합공학과 학사
2015년 ~ 현재: 광운대학교 전파공학과 석사 과정

▸ 조 혜 승 (Hye-Seung Cho)

2015년: 광운대학교 전자융합공학과 학사
2015년 ~ 현재: 광운대학교 전파공학과 석박통합 과정

▸ 김 형 국 (Hyoung-Gook Kim)

2007년 ~ 현재: 광운대학교 전자융합공학과 교수