ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 47, No. 5, pp. 479-487, 2020. 5
https://doi.org/10.5626/JOK.2020.47.5.479

음성감정데이터베이스의 분석과 프레임 단위 특징과 발음단위 특징을 통합하는 Attention Mechanism을 이용한 음성 감정 인식 시스템의 개발

(Analysis of Speech Emotion Database and Development of Speech Emotion Recognition System using Attention Mechanism Integrating Frame- and Utterance-level Features)

김 도 경 † 김 윤 중 ††
(Dokyung Kim) (Yoonjoong Kim)

요 약 본 연구에서 음성신호로부터 프레임 단위의 특징과 발음 단위의 특징을 통합하고 감정의 정보를 분석하는 BLSTM(Bidirectional Long-Sort Term Memory) 레이어, Attention mechanism 레이어 및 심층 신경회로망으로 구성되는 모델을 제안하고, 음성 감정 데이터베이스 IEMOCAP(Interactive Emotional Dyadic Motion Capture) 레이블의 신뢰성 분석에 기초하여 모델의 성능을 분석하였다. IEMOCAP 데이터베이스에서 제공되는 레이블의 평가 자료에 기초하여 기본 데이터 셋, 감정 클래스의 분포를 균형화 시킨 데이터 셋, 3명 이상의 판정에 기초하여 신뢰성이 개선된 데이터 셋을 구성하고, 각각의 데이터 셋에 대하여 화자독립 교차검증실험을 수행하였다. 개선되고 균형화된 데이터 셋에 대한 실험에서 최대 67.23% (WA, Weighted Accuracy), 56.70% (UA, Unweighted Accuracy)의 스코어를 성취하였고 기본 데이터 셋의 실험에 비하여 6.47% (WA, 4.41% (UA) 개선됨을 확인하였다.

키워드: BLSTM(Bidirectional Long-Sort Term Memory) RNN(Recurrent Neural Network), 감정 인식, Attention Mechanism, 교차검증평가, 데이터 균형화

Abstract In this study, we propose a model consist of BLSTM (Bidirectional Long-Sort Term Memory) layer, Attention mechanism layer, and Deep neural network to integrate frame- and utterance-level features from speech signals model reliability analysis the labels in the speech emotional database IEMOCAP (Interactive Emotional Dyadic Motion Capture). Based on the evaluation script of the labels provided in the IEMOCAP database, a default data set, a data set with a balanced distribution of emotion classes, and a data set with improved reliability based on three or more judgments were constructed and used for performance of the proposed model using speaker independent

† 비 회 원 : 한밭대학교 컴퓨터공학과 학생
dgkim1007@naver.com
†† 종신회원 : 한밭대학교 컴퓨터공학과 교수
(Hanbat Nat'l Univ.)
yjkim@hanbat.ac.kr
(Coressponding Author)

논문접수 : 2019년 10월 25일
(Received 25 October 2019)
논문수정 : 2020년 2월 25일
(Revised 25 February 2020)
심사완료 : 2020년 3월 11일
(Accepted 11 March 2020)

정보과학회논문지 제47권 제5호(2020. 5)

cross validation approach. Experiment on the improved and balanced dataset achieve a maximum score of 67.23% (WA, Weighted Accuracy) and 56.70% (UA, Unweighted Accuracy) that represents an improvement of 6.47% (WA), 4.41% (UA) over the baseline dataset.

Keywords: BLSTM(Bidirectional Long-Sort Term Memory) RNN(Recurrent Neural Network), emotion recognition, attention mechanism, cross-validation test, dataset balancing

1. 서 론

감정 인식 테크놀로지는 음성이나 몸짓에 관련된 정보를 분석하여 감정 상태를 확인하는 기술이다. 문화에 따라 감정을 표현하는 제스처가 다를 수 있고 성인이 되면 감정을 제어하는 경향을 보이므로 제스처와 같은 표정보다 음성에 표현되는 감정의 정보가 더 일관성이 있다. 음성 감정 인식 시스템(SER, Speech Emotion Recognition System)은 주어지는 음성 신호로부터 특징을 추출하고 감정 모델을 정의하여 학습하고 분류한다. 특징추출 과정은 20~50msec의 짧은 프레임에서 LLD(Low Level Descriptor)를 추출하고 추출된 LLD에 통계 함수를 적용하여 발음 단위 특징인 HSF(High-level Statistical Functions)를 계산한다[1].

감정을 모델링하기 위한 방법으로 과거에는 Hidden Markov Model을 이용한 방법[2]들이 주로 사용되어 왔으나 최근 심층 신경회로망(DNN, Deep Neural Network)과 RNN(Recurrent Neural Network)의 등장으로 감정 인식의 음성신호와 같은 시계열 데이터 처리 시스템의 연구에 괄목할 만한 진전이 이루어지고 있다.

Lee[3]의 연구에는 BLSTM(Bidirectional Short-Long Memory)을 이용하는 아키텍처를 제안하고 39차 프레임 단위 특징으로부터 고 수준의 감정 정보를 학습하고 ELM(Extreme Learning Machine)으로 문장 수준의 표현에 매핑시키는 방법으로 63.89%의 성능을 얻었다.

또 다른 연구[4]에서는 CNN(Convolutional Neural Network)과 RNN(Recurrent Neural Network)으로 시간 영역의 음성신호를 감정의 연속 순환 모델에 매핑시키는 end-to-end 음성 감정 인식 방법에 Attention Mechanism을 추가하여 인식률을 향상하는 모델을 제안하였다.

Chernykh[5]는 감정이 발음의 일부 프레임에 포함되어 있다고 가정하고 BLSTM과 CTC(Connectionist Temporal Classification) 구조로 감정을 심층 학습하는 모델을 제안하고 IEMOCAP[6] 감정 데이터베이스로 평가하여 정확도 54%의 성능을 보여주었다.

Mirsamadi[7]는 LSTM RNN, Pooling, 가중치 Pooling, 단순 집중(attention)등 다양한 구조를 비교 분석하고 IEMOCAP에서 실험을 수행하여 SVM 방법의 결과와 성능을 비교하였다. Zhou의 연구[8]에서는 시간 영역 프레임으로부터 SAE(stacked Auto Encoder) 또는 DBN (Deep Belief Network) 모델을 학습하고 sigmoid로 분류하였다. 베를린 감정 음성 데이터베이스(EMO-DB)[9]의 536발음 파일을 사용하여 최대 65%의 성능을 보였다. 이와 같은 연구에서 프레임 단위 특징, 발음 단위 특징을 사용하고 LSTM RNN, DNN, 단순 집중구조의 모델의 조합을 이용하고 있으며 EMO-DB나 IEMOCAP의 데이터 셋에 대하여 60% 수준의 인식률을 보이고 있다.

본 연구에서는 통합 감정 시스템을 제안하고 실험을 통하여 개선된 성능을 보이고자 한다. 음성 감정 신호로부터 프레임 단위 특징(FLF, Frame Level Feature)을 추출하고 BLSTM과 Attention Machinism[11,12]으로 감정 정보를 계산한다. 이 감정 정보는 발음단위 특징 HFS(High level Statistics Functions)과 FCN(Fully Connected Neural Network)으로 결합되어 신뢰성이 추가되는 감정으로 계산된다. 본 연구에서 사용되는 프레임 단위 특징은 프레임 수준의 MFCC(Mel-Frequency Cepstral Coefficient) 및 미분 값이고, 발음단위 특징은 MFCC, filter bank 에너지로 구성되는 LLD(low-level descriptors)로부터 계산되는 발음전체 구간의 HSF이다. 즉 LLD의 평균, 분산, 최대, 최솟값, mode의 통합체이다.

IEMOCAP 음성 감정 데이터베이스에 11종의 감정들로 구성되어 있으며 감정의 분포는 균일하지 못하다. 감정의 종류는 기존 연구들[7,10]과의 연구에서와 같은 방법으로 일관성을 유지하기 위하여 화남, 행복, 중립, 슬픔의 4종을 대상으로 하였고, 감정의 개체 수가 모두 최대 감정 수와 같아지도록 임의로 추출하여 보충하는 감정 분포를 균일화한 데이터 셋, 3명 이상의 판정에 기초하여 신뢰성이 개선된 데이터 셋을 구성하여 실험하였다. 화자독립, 교차 검증실험 방법으로 실험을 하였다.

논문의 구성은 다음과 같다. 2장에서 논문에서 채택된 LSTM과 집중 메커니즘에 대하여 소개하고, 3장에서 제안 모델의 구성에 관하여 기술한다. 4장에서 음성 감정 데이터베이스, 실험용 데이터 셋, 프레임 단위 특징 및 음성 단위 특징을 기술하고 5장에서 제안 모델의 비교 실험 및 성능 평가에 대해 논하고 6장에서는 결론에 대해 기술한다.

2. 관련연구

음성신호에서는 프레임 특징과 주변 프레임 특징들의 패턴이 감정의 특성을 표현하게 되므로 이러한 시계열

데이터의 의존성을 잘 다룰 수 있는 LSTM 모델을 본 연구에 채택하였다. 단순 RNN 모델에서 시계열 데이터의 사건들 간 지연이 존재하고 지연이 커지게 되면 이들을 학습할 때 발생될 수 있는 가중치 변수의 폭발 또는 소멸(exploding and vanishing gradient) 문제가 발생한다. 이 문제를 해결하기 위한 목적으로 개발된 결과물이 LSTM이다[11]. Felix Gers[12]는 LSTM의 아키텍처에 망각 게이트(forget gate)를 추가하여 LSTM 아키텍처가 음성 및 필기체 인식과 같은 task에서 결정적임을 보여주었다. 다음은 LSTM 아키텍처의 다이어그램이다. 백색으로 표시된 도형은 FCN 유닛이고, 황색으로 표시된 도형은 요소 단위 연산(element wise operation)을 의미한다. LSTM의 계산 과정은 식 (1)부터 식 (6)에 기술하였고 개념을 그림 1에 도식화하였다.

Fig. 1 LSTM은 입력 시퀀스에 대하여 상태 값은 동적으로 계산한다. T개의 특징으로 구성되는 입력 시퀀스를 $\{x_t | t=[1,T]\}$이라고 하자. 시간 t에서 LSTM의 동작은 다음과 같이 망각 게이트 σ_f, 입력 게이트 σ_i 및 출력 게이트 σ_o에 의하여 상태가 제어된다. 그림 1과 식 (1)에서와 같이 망각 게이트 σ_f에서 계산되는 망각 비율 f_t로 전 상태 C_{t-1}의 크기가 제어된다. 즉, 전 단계의 출력 h_{t-1}과 입력 x_t을 결합(concatenate)하고 히든 파라메터 W, b로 구성되는 FCN으로 변환하고 시그모이드 활성화함수 σ_f으로 망각 비율을 계산한다. 이와 같이 상태의 량은 망각게이트 f_t로 입력 량은 입력게이트 i_t로 출력량은 출력 게이트 o_t로 제어한다. 현 상태(기억) C_t의 크기는 전 상태 C_{t-1}와 잠정상태 $\tilde{C}_t$ 상태 크기를 f_t와 i_t으로 일종의 인터폴레션(interpolation)하는 방식으로 계산하여 필요한 내용만 기억할 수 있게 해준다.

$$f_t = \sigma_f(W_f[h_{t-1};x_t]+b_f) \quad (1)$$

$$i_t = \sigma_i([W_i[h_{t-1};x_t]+bi) \quad (2)$$

$$\tilde{C}_t = \tanh(W_c[h_{t-1};x_t]+b_c) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (4)$$

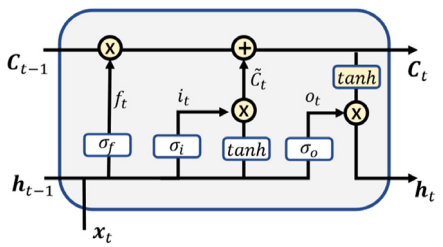

그림 1 LSTM 장치의 개념도
Fig. 1 Conceptual diagram of the LSTM unit

$$o_t = \sigma_0(W_0[h_{t-1};x_t]+b_0) \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

따라서 LSTM은 필요한 내용만 기억하는 방식으로 음성 특징 열의 의존성을 학습할 수 있고 의존성의 패턴은 감정에 따라 고유하게 대응될 수 있으므로, 시계열 특징패턴으로부터 감정을 학습하고 인식하는 감정 인식 모델에서 중요한 역할을 수행한다.

이와 같은 이유로 감정 인식과 같은 시계열 데이터를 처리하는 LSTM RNN은 심층 신경망에서 보다 우수한 성능을 보인다. 그러나 프레임 간의 거리가 멀어지면 의존성의 관계(기억능력)가 약화되는 장기 의존성 문제가 발생한다. 이를 해결하기 위하여 신경망을 이용한 번역 과정에서 전 구간의 의존성을 균일하게 계산하는 집중 메커니즘(Attention Mechanism)이 제안되었다[13]. 이 방법은 다양한 분야에 응용이 확산되었고 감정 인식 분야에서도 시도되고 있다.

$$s(h_t) = v_a^{\top}\tanh(W_a h_t) \quad (7)$$

$s(h_t)$는 softmax로 정규화되어 어텐션 벡터 α_t로 계산되고 다시 h_t와 연산되어 문맥벡터 c가 구해진다.

$$c = \sum_{t=1}^{T}\alpha_t h_t \quad (8)$$

$$\alpha_t = \frac{e^{s(h_t)}}{\sum_{i}^{T} e^{s(h_i)}} \quad (9)$$

학습과정에서 입력 시퀀스 h_t는 LSTM출력이고 문맥벡터 c는 목적 감정에 가장 상관관계가 있는 일부의 h_t 값들이 가중되어 계산되도록 학습된다. 즉 시퀀스 h_t중에서 중요한 요소들의 가중치 (어텐션벡터) α_t가 크게 되어 문맥정보 c의 값에 반영된다.

3. 제안된 통합 음성 감정 인식 시스템

본 연구에서 제안한 통합 음성 감정 인식기(MESR, Merged Sound Emotion Recognizer)의 모델은 2장에서 소개한 양방향 LSTM과 집중 메커니즘으로 구성되며, 전체 구성도는 그림 2와 같다.

하나의 음성 감정 신호(wave file)은 특징추출 과정(Feature Extraction)에서 프레임 단위 특징(FLF)과 발음 단위 특징(HSF)으로 변환되고, FLF는 BLSTM 레이어와 어텐션 레이어를 통과하여 음성데이터의 감정을 의미하는 감정의 문맥 벡터가 된다. 이 벡터는 HSF와 FCN으로 결합되고 4차원 벡터의 로짓(logit)을 계산한다. 로짓은 softmax로 정규화되고 라벨(label)과 cross entropy로 연산되어 로스(loss)로 계산된다. 학습과정에서 이 로스를 최소화하기 위하여 모델을 구성하는 모든

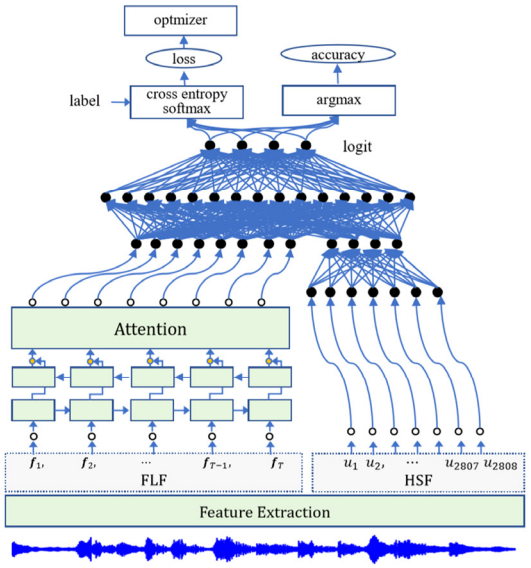

그림 2 제안된 MSER시스템 개념도
Fig. 2 Diagram of the proposed MSER

표 1 제안된 시스템 사양
Table 1 Specification of the proposed system

layer	shape	description
Frame-level features	(-1, -1, 39)	FLF, mfcc
BLSTM	(-1, -1, 39)	input shape
	32	hidden unit
	(-1,-1, 64)	output shape
Attention	32	attention size
	(-1,64)	output shape
Utterance-level features	(-1, 2808)	HSF
	(-1, 32)	FCN, relu
Merged DNN	(-1, 128)	FCN, relu
	(-1, 32)	FCN, relu
	(-1,4)	FCN, softmax

파라메터들이 갱신된다. 인식 과정에서는 평가용 데이터로부터 로짓을 계산되고 4개의 값 중에서 최댓값에 해당하는 감정을 판정한다. 화살표는 구성 요소 간의 데이터 흐름이고, 모델의 전체 파라메타들은 표 1에 제시하였다.

4. 데이터 셋

본 연구에서 사용한 감정 데이터는 IEMOCAP 데이터베이스[6]이다. 5개의 세션으로 구성되어 있으며, 10명의 배우가 혼성 쌍으로 감정을 즉흥연기 또는 대본 연기한 내용을 녹화하고 평가자가 평가하여 제작된 데이터베이스이다. 감정 데이터는 화남, 행복, 슬픔, 평상, 흥분, 두려움, 놀라움, 역겨움, 기타, 기타의 11종으로 10,039개 음성 파일로 구성되어 있으며 5개 세션으로 분리되어 있다. 각 음성파일의 감정 판정은 5명의 평가자의 결과 중 다수의 감정을 그 음성파일의 레이블로 결정하였다. 데이터베이스는 기본정보로 각 파일과 판정 레이블의 쌍으로 제공하고, 각 평가자의 평가결과도 스크립트로 제공한다. 다른 연구의 결과와 일관성을 유지하기 위하여 4개의 감정 화남, 평상, 행복, 슬픔의 4490개의 파일을 사용하였다.

데이터베이스가 제공되는 감정 레이블의 신뢰도는 평가자의 주관적 판정이므로 절대적이라 할 수 없다. 따라서 다양한 분석을 위하여 그림 3과 같이 IEMOCAP 데이터베이스를 재구성한다. 4개 감정의 샘플로 이루어지는 데이터 셋 DB4를 생성하고, 학습 및 훈련용 데이터 셋 SDF(Speaker Dependent Full), 감정의 분포를 균형화한 SDB(Speaker Dependent Balance)를 구성한다. SDF, SDB는 화자종속실험에 사용된다. 화자독립 5회 교차실험(fold-cross validation) 용 데이터 셋 SICVF (Speaker Independent Cross Validation Full)을 생성한다.

데이터 셋의 감정 레이블 신뢰도를 제고하기 위하여 3인 이상의 평가자가 동일하게 판정한 감정의 샘플만으로 구성되는 개선된 데이터 셋 DB3를 생성한다. 화자독립 5회 교차검증실험용 데이터 셋 SICV3F(Speaker Independent Cross Validation 3 Full) 및 균형화된 데이터 셋 SICV3B(Speaker Independent Cross Validation 3 Balance)를 생성한다.

화자독립 데이터 셋은 K 교차 검증 기법으로 데이터 셋 DB4의 Session1~Session5에서 Session1이 테스트 셋이 되면 나머지 Session들은 학습 셋이 되어 SICVF1이 되고 나머지 SICVF2 - SICVF5도 같은 방법으로 해서 5개의 화자독립 교차검증용 데이터 셋이 만들어진다. 또한, 개선된 데이터 셋 DB3로부터 같은 방법으로 화자독립 교차검증용 데이터 셋 SICV3F를 생성한다.

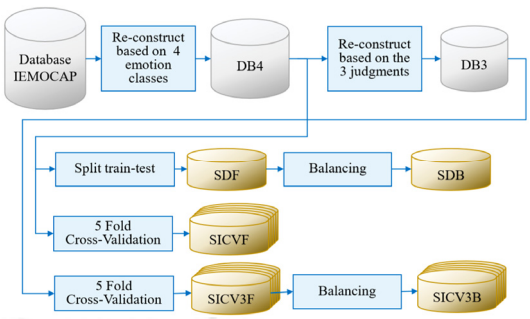

그림 3 데이터 셋의 생성절차
Fig. 3 Process dataset creation