ISSN 2005-8063
2017. 6. 30.
Vol.9 No.2
pp. 77-83

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

http://dx.doi.org/10.13064/KSSS.2017.9.2.077

히스토그램 등화와 데이터 증강 기법을 이용한 개선된 음성 감정 인식*

Improved speech emotion recognition using histogram equalization and data augmentation techniques

허 운 행 · 권 오 욱**
Heo, Woon-Haeng · Kwon, Oh-Wook

Abstract

We propose a new method to reduce emotion recognition errors caused by variation in speaker characteristics and speech rate. Firstly, for reducing variation in speaker characteristics, we adjust features from a test speaker to fit the distribution of all training data by using the histogram equalization (HE) algorithm. Secondly, for dealing with variation in speech rate, we augment the training data with speech generated in various speech rates. In computer experiments using EMO-DB, KRN-DB and eNTERFACE-DB, the proposed method is shown to improve weighted accuracy relatively by 34.7%, 23.7% and 28.1%, respectively.

Keywords: emotion recognition, histogram equalization, data augmentation

1. 서론

음성은 말의 의미뿐만 아니라 사람의 감정도 전달할 수 있다. 보통 감정 인식을 할 때 음성신호에서 감정 인식에 영향을 주는 특징들을 입력 신호로부터 추출하여 이것을 파라미터로 설정해 모델을 도출해낸다. 이때, 화자간의 특성을 고려하지 않기 때문에 특징 값들이 화자마다 조금씩 차이가 생기게 된다. 이러한 화자간의 오차들에 의해 감정 분류 기준이 되는 특징 값들 또한 영향을 받게 된다. 동일 화자 내에서는 특징 값들의 범위가 일정하지만, 화자가 달라지면 특징 값의 분포도 다르게 나타난다[1]. 대표적인 예로, 화자의 성별에 따라 피치 분포 범위가 크게 달라진다. 이러한 화자별 특성 차이를 줄인다면, 각 화자에 따른 감정 모델의 차이로부터 야기되는 인식오류도 줄일 수 있을 것이다.

본 논문에서는 화자간의 특성 차이에서 발생하는 오차를 줄여 감정 인식 성능을 향상하고자 한다. 화자간의 특성에는 발화 속도, 음의 높이, 발화 크기 등이 있다. 먼저, 화자간의 발화 속도 차이에서 기인하는 학습 모델의 오차를 줄이기 위해서 다양한 발화 속도를 갖는 데이터들로 증강하여 학습하는 데이터 증강(data augmentation; DA) 기법[2]을 적용함으로써 학습 모델의 발화 속도에 대한 강인성을 높여준다. 음의 높이, 발화 크기 등의 화자 간 특징 분포 차이를 줄이기 위하여, 히스토그램 등화(histogram equalization; HE) 기법[3]을 이용하여 데이터베이스 각 화자의 분포를 학습 데이터 전체 화자들의 분포에 맞춰준다. 히스토그램 등화를 통해 데이터베이스의 모든 화자는 각 특징마다 동일한 특징 분포를 가진다.

*이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2015R1D1A3A01020817)
** 충북대학교, owkwon@cbnu.ac.kr, 교신저자
Received 24 January 2017; Revised 13 June 2017; Accepted 21 June 2017

2. 기존 방법

음성 신호에서 감정을 찾기 위해서는 우선 감정에 영향을 미치는 특징들을 찾아야 한다. 예를 들어, 감정이 격해지면 목소리가 커지고 음의 높이 변화가 생기는 것을 직관적으로 알 수 있다. 감정 인식에서 흔히 쓰는 특징들은 피치(pitch), 에너지(energy), mel-frequency cepstral coefficient (MFCC), 지터(jitter), 쉬머(shimmer), 영 교차율(zero crossing rate; ZCR) 등이 있다[4].

감정인식 특징들을 이용하여 감정을 분류하기 위해 패턴 분류기를 사용한다. 패턴 분류기로는 Gaussian mixture model (GMM), support vector machine (SVM), deep neural network (DNN)이 사용된다[5]. 본 논문에서는 패턴분류기로 SVM을 사용하였다. 이는 SVM은 GMM보다 패턴 분류 정확도가 높고, 본 연구에서와 같이 데이터베이스의 크기가 작은 경우에는 DNN보다 더 나은 성능을 보이기 때문이다.

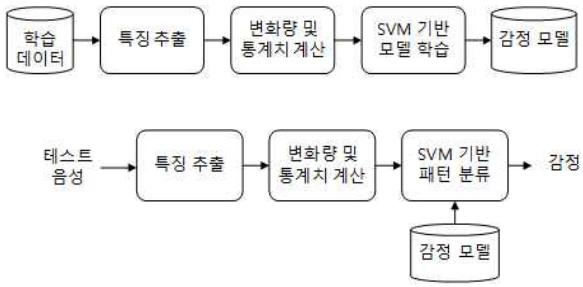

그림 1. 일반적인 감정 인식의 구조도
(a) 모델 학습 (b) 감정 인식
Figure 1. Block diagram of general emotion recognition
(a) Model learning (b) Emotion recognition

<그림 1>은 SVM 패턴 분류기를 사용한 감정 인식기의 구조도이다. SVM을 사용하려면 발화 당 하나의 특징벡터가 필요하다. 음성 신호에서 프레임 별 특징을 추출하여 통계치(최댓값, 최솟값, 평균, 표준편차, 첨도, 회귀계수 등)를 계산한다. 이러한 통계치를 연결함으로써 한 발화를 하나의 특징벡터로 표현한다. 학습 모델을 만들 때는 학습 데이터의 특징 벡터들을 SVM 패턴 분류기의 입력으로 넣어 각 감정 모델을 만들고, 테스트할 때는 입력 음성의 특징 벡터를 SVM 패턴 분류기의 감정 모델과 비교하여 감정들을 분류한다.

특징 추출 부분에서는 openSMILE 프로그램[6]을 이용하여 발화에서 MFCC 1~12차 계수, 피치, 에너지, ZCR, 하모닉대잡음비(harmonic-to-noise ratio; HNR)로 구성되는 총 16개의 특징을 추출한다. 특징을 추출할 때 사용한 환경 설정 파일(config file)은 INTERSPEECH 2009 (IS09) emotion challenge[7]에서 사용했던 IS09 384차 특징 벡터의 환경 설정 파일의 통계치 추출 부분인 low-level descriptors (LLD)를 삭제하여 재설정하였다. 윈도우 크기는 25 ms이고 10 ms씩 프레임을 이동하였다. <표 1>은 IS09 384차 특징 벡터에 대한 표이다. <표 1>의 특징 벡터는 16개의 특징과 각 특징의 미분 값을 통해 32차로 구성되고, 32차에 12개의 통계치를 적용하여 전체 384차를 얻을 수 있다.

표 1. IS09 384차 특징
Table 1. IS09 384-dimensional features

특징(32차)	통계치(12차)
(△)ZCR	평균, 표준편차, 첨도, 왜도, 최댓값, 최솟값, 최댓값 위치, 최솟값 위치, 최댓값과 최솟값 사이의 범위, 회귀 분석 파라미터(기울기, y 절편, 평균 제곱 에러)
(△)HNR	
(△)Energy	
(△)Pitch	
(△)MFCC 1-12	

SVM 패턴 분류기는 커널 함수를 이용해 비선형 분류를 한다. 이때 커널 함수는 radial basis 함수[8]를 사용하였다. 추출된 특징들을 SVM 패턴 분류기의 입력으로 사용하기 위해서 변화량 및 통계치를 계산한다. 변화량과 <표 1>의 통계치를 계산해 384차 특징 벡터를 얻어서 SVM의 입력으로 사용한다.

3. 제안 방법

3.1. 전체 구조

그림 2. 제안 방법의 구조도
(a) 모델 학습 (b) 감정 인식
Figure 2. Block diagram of the proposed method
(a) Model learning (b) Emotion recognition

제안 방법의 구조도는 <그림 2>와 같다. 모델 학습 시, <그림 1>의 구조에 음영 블록인 발화 기반 데이터 증강과 히스토그램 등화가 추가되었다. 발화 속도 기반 데이터 증강을 통해 학습 데이터를 증강시키고 특징을 추출한다. 히스토그램 등화 모델은 각 화자 데이터에 대한 화자별 히스토그램의 누적분포함수(cumulative distribution function; CDF)와 테스트 화자를 제외한 모든 화자의 학습 데이터들에 대한 누적분포함수로 구성된다. 히스토그램 등화 모델을 이용한 히스토그램 등화 과정을 통해 새로운 추정 값으로 특징 값을 대체한다. 추정된 특징 값은 변화량 및 통계치를 계산하여 SVM의 입력인 특징 벡터가 된다. 증강된 학습 데이터의 특징 벡터를 이용하여 각 감정 모델을 만든다.

감정 인식 시, 학습된 히스토그램 등화 모델과 감정 모델을

이용한다. 먼저 테스트 음성의 특징을 추출한다. 추출된 특징 값은 히스토그램 등화 모델과 테스트 데이터의 화자에 대한 히스토그램 누적분포함수를 이용한 히스토그램 등화 과정을 통해 새로운 추정 값으로 대체된다. 대체된 테스트 음성들의 특징 값을 변화량과 통계치를 계산해 특징 벡터로 만들어준다. SVM에서 테스트 음성의 특징 벡터로 감정 모델과 패턴을 비교해 최종 결과인 감정을 얻을 수 있다.

추가된 블록인 발화 속도 기반 데이터 증강은 발화 속도를 변환하여 여러 발화 속도를 고려한 데이터를 만들기 위한 것이다. 증강된 데이터는 여러 발화 속도를 가지므로 이 데이터를 이용한 학습 모델은 발화 속도에 강인한 모델이 된다. 히스토그램 등화 과정은 각 특징의 화자별 히스토그램의 누적분포함수를 통해 각 특징들의 분포와 분포 범위를 같게 만들어 주는 새로운 특징 값을 추정하여 대체한다. 이를 통해 화자별 특성에 의한 특징 분포에 의한 오차를 줄일 수 있다.

테스트는 화자 교차 검증법을 통해 진행된다. 그러므로 학습 시, 각 테스트 화자에 따른 히스토그램 등화 모델과 감정 모델을 각 테스트 화자마다 별도로 생성해주어야 한다.

3.2. 발화 속도 변환 알고리듬

발화의 속도 변환의 핵심은 1배속의 데이터를 다른 배속의 데이터로 바꿨을 때 데이터 손실이나 왜곡이 없어야 한다는 것이다. 발화 속도를 변환할 때, 단순히 시간 축을 늘이거나 줄이면 주파수 영역에서 왜곡이 생겨 피치가 바뀌게 되어 음성이 변조된 것처럼 들린다. 이러한 현상을 막기 위해서 파형 유사도 중첩 가산(waveform similarity overlap add; WSOLA) 알고리듬[9]을 이용한다. 본 논문에서는 WSOLA 알고리듬을 이용하는 SoX 프로그램[10]으로 발화 속도 기반 데이터 증강을 하였다.

WSOLA를 이용하여 발화 속도가 변환된 신호는 아래의 식과 같이 표현할 수 있다[9].

$$y(n)=\sum_{k} v(n-kS)x(n+kS(\alpha-1)+\delta_k) \quad (1)$$

$x(n)$은 원래 신호, $y(n)$은 발화 속도가 변환된 출력 신호, k는 인덱스, α는 시간 비율 인자(time scale factor), S는 윈도우의 절반 길이를 나타낸다. 출력 신호 $y(n)$에서 $v(n)$은 해닝 윈도우이다. δ_k는 유사도를 높이기 위한 변수이고, 중첩 가산 과정에서 중첩되는 부분의 유사도가 가장 높은 값을 가지도록 결정된다. δ_k를 결정하기 위해서 상호 상관계수를 이용한다.

<그림 3>은 WSOLA 적용 과정이다. L_k는 kS를 말하고, R_k는 $x(n)$에서 $k-1$ 번째 신호에서 윈도우 절반 길이인 S만큼 이동한 신호를 말한다. C_k는 $x(n)$에서 시간 비율 인자 α에 의해 결정되는 중첩 위치에서의 신호를 말한다. $y(n)$의 $k-1$ 번째 중첩 위치 L_{k-1}의 A 신호는 $x(n)$의 A' 신호이다. A' 신호는 유사도를 고려해 δ_{k-1} 만큼 이동한 $k-1$ 번째 신호이다. $y(n)$에서 B 신호는 A 신호와 윈도우의 절반 길이 S 만큼 겹쳐 있다. 그래서 A'신호에서 윈도우의 절반 길이 S 만큼 이동한 R_k 신호와 $x(n)$의 k번째 중첩 위치 신호인 C_k의 유사도가 가장 높은 δ_k를 결정한다. 중첩 위치에서 δ_k 만큼 이동한 B' 신호는 B 신호가 된다. B 신호가 결정되면 다음 $k+1$ 번째 신호를 결정하기 위해서 B'에서 S 만큼 이동한 R_{k+1} 신호를 이용하여 위와 같은 과정으로 WSOLA를 한다.

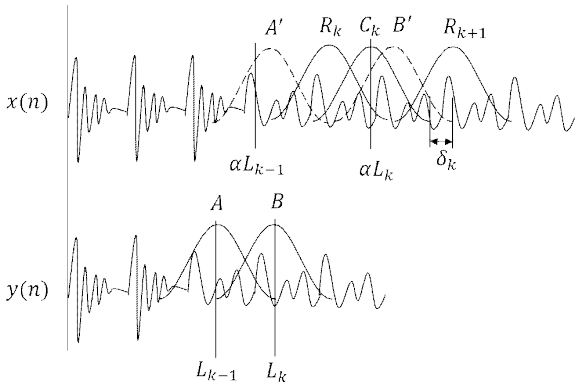

그림 3. 파형 유사도 중첩 가산
Figure 3. WSOLA

δ_k 값을 탐색할 때 임계치 $\delta_{\max}$를 설정해 k 번째 중첩 위치 αL_k 근방에서 R_k와 C_k의 유사도 값을 가장 높게 갖게 하는 δ_k를 결정하게 한다. SOX 프로그램에서는 14.68 ms 로 설정되었다.

3.3. 히스토그램 등화

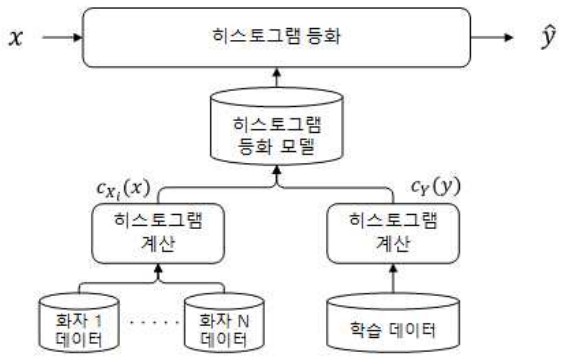

그림 4. 히스토그램 등화
Figure 4. Histogram equalization

히스토그램 등화는 <그림 4>와 같은 구조로 이뤄져있다[2]. N은 데이터베이스의 모든 화자 수를 말한다. 먼저, 각 화자에 대한 데이터를 이용한 화자별 히스토그램을 계산하여 누적분포함수 $c_{X_i}(x)$를 만든다. 10명의 화자라면 히스토그램의 누적분포함수를 c_{X_1}부터 $c_{X_{10}}$까지 만들 수 있다. 테스트 화자를 제외한 모든 화자의 학습데이터를 모아 히스토그램의 누적분포함수 $c_Y(x)$를 만든다. 히스토그램 등화 모델은 누적분포함수 $c_{X_i}(x)$와 $c_Y(x)$

로 구성된다. 화자 i에 대한 히스토그램 등화는 다음 식과 같이 표현할 수 있다.

$$\hat{y} = c_Y^{-1}(c_{X_i}(x)) \quad (2)$$

화자에 대한 누적분포함수와 전체 학습 데이터에 대한 누적분포함수의 출력 값을 비교하여 새로운 값을 추정하는 식이다. 이것을 쉽게 <그림 5>를 통해 이해할 수 있다. 각 화자의 데이터를 사상하기 위해 해당 데이터의 화자의 정보를 이용해 x 값의 누적 분포 함수의 출력 값 $c_{X_i}(x)$을 찾는다. 그 출력 값과 학습 데이터의 누적분포함수 출력 값이 같은($C_{X_i}(x) = C_Y(\hat{y})$) 출력을 갖는 $\hat{y}$를 구한다. 이러한 과정을 통해 각 특징 16개의 프레임 당 출력된 값들을 모두 새로운 추정 값 $\hat{y}$로 사상할 수 있다.

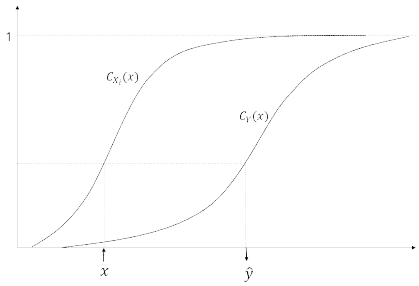

그림 5. 누적 분포 함수 사상 과정
Figure 5. Mapping process of CDF

추정된 값으로 대체된 특징 값은 학습 데이터에 대한 누적분포함수 $c_Y(x)$의 분포와 범위를 따른다. 테스트 시, 화자 교차 검증법을 이용하므로 테스트 화자가 바뀔 때 마다 새로운 학습 데이터에 대한 누적분포함수 $c_Y(x)$를 구한다. 그러므로 히스토그램 등화를 통해 같은 누적분포함수 $c_Y(x)$를 이용해 추정된 특징 값은 서로 다른 화자에 대해 같은 분포와 범위를 가지게 되지만, 테스트 화자가 달라진 새로운 누적분포함수 $c_Y(x)$에 대해 추정된 특징 값과의 분포와 범위는 다르다.

4. 실험 결과

4.1. 데이터베이스

음성 데이터베이스는 독일어 데이터베이스(EMO-DB)[11], 한국어 데이터베이스(KRN-DB)[12], eNTERFACE 데이터베이스[13]을 이용했다.

EMO-DB는 7개의 감정(화남, 중립, 두려움, 지루함, 행복, 슬픔, 역겨움)을 독일어로 발성한 파일들로 구성되어 있다. DB의 화자는 5명의 남성과 5명의 여성의 독일인 전문 배우들이다. 발성 내용은 녹음 전에 제공된 것으로 10개 종류의 대사로 녹음되었다. 각 대사는 평균적으로 9 어절 이상을 포함하는 한 문장으로 이뤄져있다. 발화 길이는 2~3초 정도이다. 본 논문에서는 7개의 감정 중 감정 인식 실험에서 주로 쓰이는 4개의 감정(화남, 행복, 슬픔, 중립)을 선정하였다[14][15]. 실험에 쓰인 총 파일개수는 339개이다.

KRN-DB는 6개의 감정(화남, 중립, 슬픔, 놀람, 행복, 지루함)을 한국어로 발성한 파일들로 구성되어 있다. DB의 화자는 15명의 남성과 15명의 여성의 한국 일반인이다. 발성 내용은 평균적으로 1~2 어절로 구성된 문장으로 이뤄져있다. 발화 길이는 평균적으로 1초 이내이다. 이 데이터베이스에서도 마찬가지로 4개의 감정(화남, 행복, 슬픔, 중립)을 선정하였다. 실험에 쓰인 총 파일개수는 6,058개이다.

eNTERFACE-DB는 6개의 감정(행복, 슬픔, 놀람, 화남, 두려움, 역겨움)을 영어로 발성한 파일들로 구성되어 있다. DB의 화자는 14개 서로 다른 국적의 34명의 남성과 8명의 여성 일반인들이다. 발성 내용은 짧은 이야기에 대한 여러 반응이 대본으로 주어진다. 대본에 주어진 대사는 평균적으로 6 어절을 포함하는 한 문장으로 이뤄져있다. 발화 길이는 2~3초 정도이다. 이 데이터베이스는 중립의 감정을 포함하지 않아 6개의 감정에 대해 실험을 진행하였고, 실험에 쓰인 파일개수는 1,233개이다.

4.2. 실험 결과

실험은 총 3가지 방식으로 구성된다. 실험 방식은 베이스라인(Baseline), Baseline+히스토그램 등화(HE) 실험, 제안 방법인 Baseline+HE+데이터 증강(DA) 실험 3가지이다. Baseline은 <그림 1>과 같은 구조를 가지고 Baseline+HE 실험은 <그림 1>의 특징 추출 블록 다음 부분에 히스토그램 등화 블록이 추가된 구조를 가진다. 제안 방법인 Baseline +HE+DA 실험은 <그림 2>와 같은 구조를 가진다.

각 데이터베이스는 화자 교차 검증법을 통해 실험을 진행하였고, 각 감정의 혼동 행렬을 통해 가중치 인식률(weighted accuracy; WA)[16]과 비가중치 인식률(unweighted accuracy; UWA)[16]로 결과를 나타낸다. 가중치 인식률은 전체 시스템의 인식률을 나타내는 것이고, 비가중치 인식률은 각 감정에 대한 인식률의 평균, 즉 혼동행렬의 주대각선의 평균을 나타낸 것이다. 그러므로 가중치 인식률과 비가중치 인식률의 차이가 크다면 각 감정에 대한 인식 성능의 편차가 크다는 것을 나타낸다. 반대로 가중치 인식률과 비가중치 인식률의 차이가 작다면 각 감정에 대한 인식 성능이 유사하다는 것을 나타낸다.

4.2.1. EMO-DB 실험 결과

발화 속도 기반 데이터 증강 기법에 사용할 발화 속도를 결정해야한다. Baseline에 발화 속도 기반 데이터 증강 기법을 적용하여 발화 속도 종류에 따른 인식 성능을 비교해보았다. 이 실험은 <그림 2>에서 히스토그램 등화 블록이 제거된 구조를 가진다.

표 2. 발화 속도에 따라 증강된 데이터베이스의 인식률
Table 2. Recognition rate of augmented databases according to speech rate

발화 속도 종류	발화 속도(배속)	인식률(%)
D1	1, 1±0.02	86.1
D2	1, 1±0.04	85.8
D3	1, 1±0.06	**86.4**
D4	1, 1±0.08	85.5
D5	1, 1±0.1	85.3
D6	1, 1±0.02, 1±0.04	**86.4**
D7	1, 1±0.02, 1±0.04, 1±0.06	85.8
D8	1, 1±0.02, 1±0.04, 1±0.06, 1±0.08	85.5
D9	1, 1±0.02, ..., 1±0.08, 1±0.1	85.5

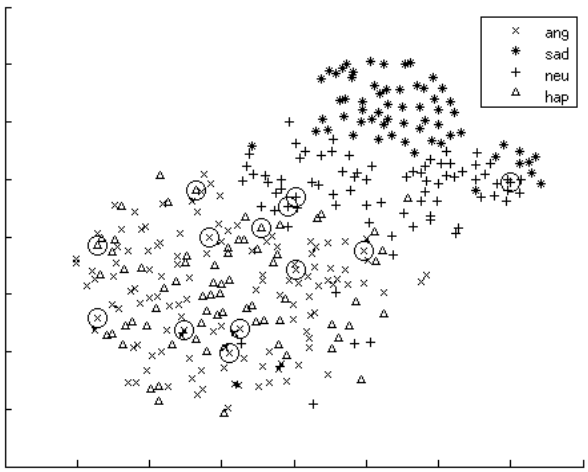

그림 6. 2차원 평면으로 투영된 EMO-DB 특징벡터의 분포
Figure 6. Distribution of EMO-DB feature vectors projected into 2-dimensional plane

<표 2>는 Baseline에 발화 속도 기반 데이터 증강 기법이 적용된 실험의 인식 결과이다. 발화 속도 종류 D1에서 D5로 갈수록 발화 속도 변환 폭이 0.02 배속만큼 커지고, D6~D9는 0.02배속 변환 폭으로 커지는 데이터가 중첩으로 쌓여 최대 11배까지 데이터 증강을 한다. <표 3>의 Baseline보다 모든 발화 속도 종류에서 향상된 성능을 보이고, 특히, D3과 D6 발화 속도 종류에서 가장 높은 성능을 보인다.

<그림 6>은 EMO-DB에서 모든 데이터들의 특징벡터를 2차원 평면에 투영한 것이다[17]. 그림에서 원으로 표시된 데이터들은 감정 모델에 대한 유사도가 낮은 데이터로, 감정 모델의 경계선에 있는 데이터들이다. 이 데이터들은 Baseline에서 오인식 되었지만, D3을 대상으로 한 DA 실험에서는 제대로 인식되었다. 발화속도 기반 DA는 감정 모델들의 경계선에서 오인식되는 데이터들을 더욱 잘 인식되도록 한다.

위의 실험에서 발화 속도 기반 DA를 통해 Baseline보다 감정 인식 성능을 개선시킬 수 있다는 것을 알 수 있었고, 각 감정 모델의 경계선을 가장 잘 표현하는 D3과 D6 발화 속도 종류에 대해 HE와 같이 적용하여 제안 방법 실험을 하였다.

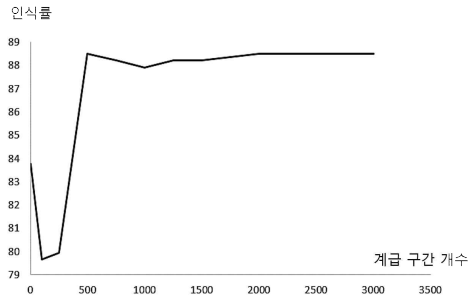

그림 7. 계급 구간에 따른 히스토그램 등화 실험 인식률
Figure 7. Recognition rate of HE experiment according to bin size

표 3. EMO-DB 실험 결과
Table 3. Experimental results in EMO-DB

데이터베이스 종류	실험	WA	UWA
EMO-DB	Baseline	83.8	84.2
	Baseline+HE	88.5	88.7
	Baseline+HE+DA	**88.8**	**88.3**

<표 3>은 Baseline, Baseline+HE, Baseline+HE+DA 3가지 실험에 대한 결과표이다. 히스토그램 등화 과정에서 계급 구간이 많을수록 조밀한 누적분포함수를 만들 수 있다. 또한, 조밀한 누적분포함수를 이용하여 더 정확한 새로운 특징 값을 추정할 수 있다. 본 논문에서는 <그림 7>과 같이 2,000개 이상의 히스토그램 계급 구간에서 일정한 성능을 보여, 2,000개의 히스토그램 계급 구간을 이용하였다. Baseline+ HE 실험에서는 88.5% 가중치 인식률과 88.7% 비가중치 인식률 결과로, Baseline 실험 결과보다 29.1% 상대적 가중치 인식률 개선 결과와 29.0% 상대적 비가중치 인식률 개선 결과를 얻었다. 제안 방법인 Baseline+HE+DA 실험 결과는 D3의 발화속도 종류에서 89.4% 가중치 인식률과 89.5% 비가중치 인식률 결과로 가장 높은 인식 성능을 보였다. Baseline 실험 결과보다 34.7% 상대적 가중치 인식률 개선과 33.8% 비가중치 인식률 개선 결과를 얻었다.

표 4. EMO-DB에서 Baseline+HE+DA 실험 결과의 혼동 행렬
Table 4. Confusion matrix of Baseline+HE+DA experimental results in EMO-DB

실제 \ 예측	화남	슬픔	중립	행복
화남	89.8	0	0	10.2
슬픔	0	96.8	3.2	0
중립	1.3	3.8	91.1	3.8
행복	18.3	0	1.4	80.3

<표 4>는 Baseline+HE+DA 실험 결과에 대한 혼동 행렬이다. 혼동 행렬에서 각 감정에 따른 인식 성능을 볼 수 있었다. 화남의 오류는 행복에서 나타나고, 행복의 오류는 화남에서 나타는 것으로 보아, 화남과 행복 모델 사이의 인식 오류가 크다. 화자

간의 특성을 줄여 줌으로써 성능을 개선함과 동시에 화남과 행복 감정 모델 사이의 오류를 줄일 수 있다면 감정 인식 시스템의 성능을 더욱 높일 수 있을 것이다.

4.2.2. 다른 데이터베이스 실험 결과

데이터베이스에 따른 실험 성능을 비교해보기 위해 KRN-DB와 eNTERFACE-DB에 대해 추가 실험을 하였다. 앞의 실험과 같은 IS09 384차 특징 벡터와 동일한 설정의 SVM을 이용하였다. 히스토그램 등화는 2,000개의 히스토그램 계급 구간을 이용하였고, 발화 속도 기반 데이터 증강은 <표 2>의 D3 발화 속도를 이용하였다. 데이터베이스를 제외한 실험 조건은 모두 동일하게 설정하였다.

표 5. KRN-DB, eNTERFACE -DB 실험 결과
Table 5. Experimental results in KRN-DB, eNTERFACE-DB

데이터베이스 종류	실험	WA	UWA
KRN-DB	Baseline	61.4	61.5
	Baseline+HE	69.5	69.5
	Baseline+HE+DA	**70.6**	**70.7**
eNTERFACE-DB	Baseline	59.9	59.8
	Baseline+HE	69.6	69.6
	Baseline+HE+DA	**71.1**	**71.1**

<표 5>는 KRN-DB와 eNTERFACE-DB에서 Baseline, Baseline+HE, Baseline+HE+DA 3가지 실험에 대한 결과표이다. KRN-DB에서 Baseline+HE 실험 결과는 Baseline 실험 결과보다 21.0% 상대적 가중치 인식률 개선과 20.8%의 상대적 비가중치 인식률 개선 결과를 얻었다. 제안 방법인 Baseline+HE+DA 실험 결과는 Baseline 실험 결과 보다 23.7% 상대적 가중치 인식률 개선과 23.8% 비가중치 인식률 개선 결과를 얻었다.

eNTERFACE-DB에서 Baseline+HE 실험 결과는 Base- line 실험 결과보다 24.2% 상대적 가중치 인식률 개선과 24.2% 상대적 비가중치 인식률 개선 결과를 얻었다. 제안 방법인 Baseline+HE+DA 실험 실험 결과는 Baseline 실험 결과보다 28.1% 상대적 가중치 인식률 개선과 28.1% 상대적 비가중치 인식률 개선 결과를 얻었다.

표 6. KRN-DB에서 Baseline+HE+DA 실험 결과의 혼동 행렬
Table 6. Confusion matrix of Baseline+HE+DA experimental results in KRN-DB

실제 \ 예측	화남	슬픔	중립	행복
화남	78.0	1.1	10.1	10.8
슬픔	0.1	82.2	10.7	7.0
중립	10.4	11.8	61.5	16.3
행복	13.1	7.3	18.7	60.9

발화 속도 기반 데이터 증강을 통한 성능 개선은 독일어 데이터베이스와 다르게, 위의 두 데이터베이스에서는 성능이 많이 개선되지 않았다. EMO-DB에서는 화자가 10명이라 감정 모델에 여러 발화 속도가 고려될 수 없었고, 다른 두 DB는 화자 수가 30, 42명으로 상대적으로 많은 수의 화자로 구성되어 있어서 감정 모델에 여러 발화 속도가 고려되어 효과가 낮은 것으로 생각된다.

<표 6>는 KRN-DB의 Baseline+HE+DA 실험 결과에 대한 혼동 행렬이다. KRN-DB에서는 화남, 중립, 행복 모델 사이의 감정 인식 오류가 크다. EMO-DB와 마찬가지로 화자간의 특성을 줄여 줌으로써 성능을 개선함과 동시에 혼동 행렬에 나타난 특정 감정 모델 사이의 오류를 줄일 수 있다면 감정 인식 시스템의 성능을 더욱 높일 수 있을 것이다.

5. 결론

본 논문에서는 감정 인식을 위해 화자 간의 특성 차이를 줄이기 위한 방법을 제시했다. 화자의 발화 속도 차이로 인한 학습 모델의 오차를 줄이기 위해 데이터를 증강하였고, 화자 특징의 분포와 분포 범위 오차를 줄여주기 위해 히스토그램 등화를 적용했다.

히스토그램 등화는 모든 데이터베이스에서 인식 성능이 많이 개선되었지만, 발화 속도 기반 데이터 증강은 데이터베이스에 따라 성능 개선 정도가 달랐다. 발화 속도 기반 데이터 증강은 EMO-DB에서는 많은 성능 개선 효과를 보였고 다른 두 DB에 대해서는 적은 효과를 보였다.

히스토그램 등화를 이용했을 때, 독일어 데이터베이스와 다른 데이터베이스에서 평균적으로 25% 상대적 인식률 개선이 있었다. 발화 속도 기반 데이터 증강과 히스토그램 등화를 이용한 제안 방법을 이용하여 가장 높은 감정 인식 결과를 얻을 수 있었고, 독일어 데이터베이스와 다른 데이터베이스에서 평균적으로 28.5% 상대적 인식률 개선이 있었다. 제안된 방법을 통해 화자의 특성을 줄여 줌으로써 성능을 개선할 수 있음을 확인하였다.

참고문헌

[1] Sethu, V., Ambikairajah, E., & Epps, J. (2007). Speaker normalisation for speech-based emotion detection. *Proceedings of Digital Signal Processing* (pp. 611-614).

[2] Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio Augmentation for Speech Recognition. *Proceedings of INTERSPEECH* (pp. 3586-3589).

[3] Chiou, B. C., & Chen, C. P. (2014). Speech Emotion Recognition with Cross-lingual Databases. *Proceedings of INTERSPEECH* (pp. 558-561).

[4] Kwon, C., Song, S., Kim, J., Kim, K., & Jang, J. (2012). Extraction of Speech Features for Emotion Recognition. *Phonetics and Speech Sciences*, 4(2), 73-78. (권철홍·송승규·김종열·김근호·장준수 (2012). 감정 인식을 위한 음성 특징 도출. *말소리와 음*

성과학, 4(2), 73-78.)

[5] Han, K., Yu, D., & Tashev, I. (2014). Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. *Proceedings of INTERSPEECH* (pp. 223-227).

[6] Eyben, F., Wöllmer, M., & Schuller, B. (2009). OpenEAR－Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. *Proceedings of the Affective Computing and Intelligent Interaction* (pp. 1-6).

[7] Schuller, B., Steidl, S., & Batliner, A. (2009). The INTERSPEECH 2009 Emotion Challenge. *Proceedings of INTERSPEECH* (pp. 312-315).

[8] Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3), 273-297.

[9] Verhelst, W., & Roelands, M. (1993). An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. *Proceedings of International Conference Acoustics, Speech, and Signal Processing* (pp. 554-557).

[10] Bagwell, C., & Klauer, U. (2015). SoX - sound exchange. Retrieved from http://sox.sourceforge.net/ on November 25, 2016.

[11] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A Database of German Emotional Speech. *Proceedings of INTERSPEECH* (pp. 1517-1520).

[12] Jang, K., & Kwon, O. (2006). Speech Emotion Recognition for Affective Human-Robot Interaction. *Proceedings of International Conference on Speech and Computer* (pp. 419-422).

[13] Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The eNTERFACE'05 Audio-Visual Emotion Database. *Proceedings of International Conference Data Engineering Workshops* (pp. 1-8).

[14] Lee, J., & Tashev, I. (2015). High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition. *Proceedings of INTERSPEECH* (pp. 1537-1540).

[15] Jin, Q., Li, C., Chen, S., & Wu, H. (2015). Speech emotion recognition with acoustic and lexical features. *Proceedings of International Conference Acoustics, Speech, and Signal Processing* (pp. 4749-4753).

[16] Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37-63.

[17] Van der Maaten, L. (2014). Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research*, 15(1), 3221-3245.

• **허운행 (Heo, Woon-Haeng)**
충북대학교 제어로봇공학전공
충북 청주시 서원구 충대로1
Email: whheo@cbnu.ac.kr
관심분야: 감정인식, 음성인식
현재 제어로봇공학과 박사 재학 중

• **권오욱 (Kwon, Oh-Wook)** 교신저자
충북대학교 전자공학부
충북 청주시 서원구 충대로1
Tel: 043-261-3374
Email: owkwon@cbnu.ac.kr
관심분야: 음성인식, 화자인식, 감정인식, 음성신호처리
2003~ 현재 충북대학교 전자공학부 교수