Journal of KIIT. Vol. 6, No. 6, pp. 63-70, Jun. 30, 2018. pISSN 1598-8619, eISSN 2093-7571 63
http://dx.doi.org/10.14801/jkiit.2018.16.6.63

SVM 기반의 오디오 이벤트 검출 성능 분석

정석환*, 정용주**

Performance Analysis of Audio Event Detection Based on SVM

Suk-Hwan Chung*, Yong-Joo Chung**

요 약

본 논문에서는 오디오 이벤트 검출을 위한 SVM 기법에 대하여 심층적 고찰을 진행하였다. 이를 위해서 SVM에 대한 입력특징의 차원과 입력 형태를 달리하며 성능변화를 관찰하였다. 또한, 커널 함수와 슬랙변수의 파라미터들이 인식 성능에 미치는 영향을 관찰하였으며 이를 통하여 오디오 이벤트 검출을 위한 SVM 분류기에 대한 최적의 조건을 도출하고자 하였다. DCASE 챌린지 2016의 Task 3에 사용된 오디오데이터를 이용한 분류 실험결과, MFCC 특징을 80ms 구간동안 평균한 후에 SVM에 입력하며, MFCC 특징의 차원을 13차로 하고, 슬랙변수와 커널함수 파라미터값들이 각각 $C=1, \gamma=0.00769$일 경우, 최고의 성능을 나타냄을 알 수 있었다. 이러한 최적화를 통해서, SVM은 기존의 GMM에 비해서 F-score 값에서 상당한 성능 향상을 보임을 알 수 있었다.

Abstract

In this paper, we performed deep investigation on the SVM method for audio event detection. We observed performance variation by varying the dimension and form of the input features of the SVM. In addition, the effects on the recognition performance of the parameters of the kernel function and the slack variables were observed, through which we tried to find the optimum conditions of the SVM classifiers for the audio event detection. From the classification results using the audio data in Task 3 of DCASE Challenge 2016, we found that the best performance is obtained when MFCC features are input to the SVM after averaging for 80ms, the MFCC feature dimension 13 and the parameters of the slack variables and kernel function are $C=1, \gamma=0.00769$, respectively. Through these optimization, we could find that SVM showed significant performance improvement in F-score compared with the conventional GMM.

Keywords
audio event detection, MFCC, GMM, SVM

* 계명대학교 전자공학과
- ORCID: https://orcid.org/0000-0002-1898-4947
** 계명대학교 전자공학과(교신저자)
- ORCID: https://orcid.org/0000-0002-0060-1178

· Received: Apr. 02, 2018, Revised: May 09, 2018, Accepted: May 12, 2018
· Corresponding Author: Yong-Joo Chung
Dept. of Electronics Engineering Keimyung University, 704-701
Shindang-dong, Dalseo-gu, Daegu-si, 1000, Republic of Korea,
Tel.: +82-53-580-5925, Email: yjjung@kmu.ac.kr

Ⅰ. 서 론

오디오 이벤트 검출은 패턴 인식 기술을 기반으로 하며, 보안 감시나 멀티미디어 컨텐츠 검색과 인덱싱 그리고 헬스케어 및 자율주행차 등의 다양한 분야에서 활용 가능하여 최근 많은 연구자들의 관심을 받고 있다[1][2].

영상이나 음성인식 분야와 마찬가지로 오디오 이벤트 검출 연구에서도 공동의 데이터베이스를 가지는 것은 매우 중요한다. 이는 이 분야의 연구자들에게 데이터를 개별적으로 구축해야 하는 부담을 덜어 줄 뿐만 아니라, 개발된 알고리듬들의 성능을 상호 비교하게 해 준다는 면에서 매우 필요하다. 이러한 요구에 기반하여 오디오 신호 분류를 위한 챌린지(Challenge) 프로그램인 DCASE(Detection and Classification of Acoustic Scenes and Events)가 2013년 과 2016년에 걸쳐서 개최되었다[3][4]. 여기서는 크게 2가지의 인식 카테고리가 다루어지는데, 그 하나는 음향 장면 분류(Acoustic Scene Classification)이고 또 다른 하나가 오디오 이벤트 검출(Audio Event Detection)이다 . 음향 장면 분류는 소리가 발생하는 전체적인 환경을 분류하며, 오디오 이벤트 검출에서는 특정한 오디오 클래스의 발생 유무와 시간 정보를 함께 인식한다. 본 논문에서는 이중에서도 오디오 이벤트 검출에 대해서 논의하고자 하며, DCASE 2016에서 제시된 공용 오디오데이터 베이스를 활용하여 인식 결과를 제시하고자 한다.

GMM(Gaussian Mixture Model)은 DCASE 2016 프로그램에서 음향 장면 분류와 오디오 이벤트 검출 모두에서 베이스라인 인식기로 사용되었다. 두 카테고리 모두에서 프레임 뭉치(Bag-of-frames) 기법에 기반한 GMM이 사용되었으며[5], 오디오 신호에 대한 특징 벡터로는 음성인식에서 많이 사용되는 MFCC(Mel- Frequency Cepstral Coefficients)가 적용되었다.

SVM(Support Vector Machine)은 오디오 이벤트 검출이나 음향 장면 분류에 대한 인식에서 GMM과 유사한 성능을 보임이 알려져 있다[6][7]. 오디오 신호 특징으로서는 GMM과 동일하게 MFCC을 사용하지만, 매 프레임의 오디오 특징을 독립적으로 취급하는 GMM과 다르게 SVM에서는 특징의 시간 구간별 평균을 구하거나 최대나 최소값을 이용하는 방법이 사용되기도 한다[8][9]. 또한, 비선형의 커널(Kernel) 함수를 사용함으로서 선형적으로 분리 가능하지 않은 데이터들의 분류가 가능하게 한다. SVM에서 일반적으로 가우시안 RBF(Radial Basis Function) 커널함수를 많이 사용하는데 가우시안 RBF에서 사용되는 하이퍼파라미터(Hyperparameter)인 γ 값을 조절함으로서 개별의 학습 샘플들이 결정경계에 미치는 영향을 조절 할 수 있다. 즉, 작은 γ 값은 결정 경계를 보다 부드럽게 하고 큰 γ 값은 결정 경계를 학습샘플에 크게 의존하게 함으로서 결정 경계가 보다 심하게 요동치게 한다. 따라서 동일한 학습데이터를 이용하더라도 γ 값에 따라서 결정 경계가 달라지게 됨으로써 인식 성능이 크게 달라질 수 있다. 또한, SVM에서는 학습시에 정규화(Regularization)를 위해서 하이퍼파라미터 C를 사용하는데 C 값을 크게 할 경우 결정 경계가 학습데이터의 오류를 최소화 하려는 경향이 생기고 C 값을 작게 하는 경우 결정 경계가 학습데이터에 덜 의존하게 되어 결과적으로 오버피팅 이 줄어들게 하는 역할을 한다.

이와 같이 GMM과 달리 SVM의 경우에는 인식 성능에 미치는 다양한 변수가 존재하며 이들이 오디오 이벤트 검출기의 성능에 미치는 영향에 대한 분석이 필요하다고 판단된다. 따라서 본 연구에서는 SVM 인식기의 하이퍼파라미터 값을 조절하고, 입력 음성특징의 형태를 변화시켜 다양한 오디오 이벤트 검출 성능을 검토함으로서 최적의 SVM 인식기를 제안하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 오디오 이벤트 검출을 위한 특징 추출 방법과 GMM과 SVM분류기에 대한 소개를 하며 3장에서는 본 연구에서 수행한 다양한 분류 실험결과를 제시하고, 비교 분석 하였으며 4장에서 결론을 맺고 향후 연구과제 등에 대해서 소개한다.

Ⅱ. 오디오 특징 추출과 인식 방법

2.1 특징 추출

오디오 신호에 대한 특징 추출 기법은 분류기에

따라서 달라지는데 GMM과 SVM의 경우에는 일반적으로 MFCC를 사용하게 된다[10]. 본 논문의 분류 실험에 사용된 DCASE 2016의 오디오 데이터는 44.1 KHz로 샘플링 되었으며, MFCC 추출을 위한 프레임의 길이는 40ms이며 50%의 건너뛰기 크기(Hop Size)를 가진다. 매 프레임의 오디오 샘플들은 프리-엠퍼시스(Pre-empahsis)와 해밍(Hamming) 윈도우를 거친 후 2048-포인트 FFT(Fast Fourier Transformation)을 통하여 주파수 영역으로 변환되게 된다. FFT의 결과로부터 얻게 되는 40차의 멜-스케일(Mel-scale)의 필터뱅크(Filterbank) 값을 로그 변환한 후 DCT(Discrete Cosine Transformation)를 적용함으로서 최종적으로 매 프레임 마다 20차의 MFCC가 얻어지게 된다(0차 계수 포함). 앞서 추출된 20차의 정적(Static) 계수의 차분과 차차분 계수를 포함하여 전체적으로 60차의 MFCC 계수가 사용되었다[4]. 차분과 차차분 MFCC 계수를 얻기 위한 윈도우의 길이는 9 프레임으로 하였다.

한편, 추출된 MFCC 계수는 GMM과 SVM에서 다소 다르게 활용되는 것이 일반적이다. 그림 1에는 GMM과 SVM 분류기에서 각각 추출된 MFCC 특징을 사용하는 방법에 대한 차이점이 나타나 있다. GMM의 경우 매 프레임에서 추출된 MFCC 계수들이 그대로 독립적으로 사용되는 반면, SVM에서는 슬라이딩 윈도우를 적용하여 얻어지는 이들의 평균값을 사용하는 것이 일반적이다.

본 연구에서는 SVM의 특징 값으로 그림 1에서처럼 평균값을 사용하였으며, 평균을 적용하지 않은 프레임 단위의 특징과의 성능비교를 실시하였다.

2.2 오디오 이벤트 검출 방법

2.2.1 GMM

GMM은 M개의 가우시안 확률 분포를 가중치를 곱한 후 합해서 만들어진 통계적 확률모델로 전통적으로 음성인식이나 오디오 이벤트 검출 분야에서 널리 사용되는 기법이다. GMM의 확률분포는 식(1)과 같이 표시된다.

$$b(\boldsymbol{O}_t) = \sum_{m=1}^{M} w_m N(\boldsymbol{O}_t|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{1}$$

$$N(\boldsymbol{O}_t|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = \frac{1}{\sqrt{(2\pi)^N|\boldsymbol{\Sigma}_m|}} e^{-\frac{1}{2}(\boldsymbol{O}_t-\boldsymbol{\mu}_m)\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{O}_t-\boldsymbol{\mu}_m)^T} \tag{2}$$

여기에서 $N(\boldsymbol{O}_t|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$은 N차 입력 벡터 $\boldsymbol{O}_t$에 대해 평균벡터 $\boldsymbol{\mu}_m$과 공분산행렬 $\boldsymbol{\Sigma}_m$을 가지는 단일 가우시안 확률 밀도 함수를 의미하며, ω_m은 각 단일가우시안 확률밀도 함수에 대한 가중치를 나타낸다.

본 논문에서 사용된 GMM 분류기는 DCASE 2016에서 제공된 베이스라인 시스템을 사용하였다. 앞 절에서 언급된 60차의 MFCC 중 0차 계수를 제외한 59차가 특징벡터로 사용되었으며, GMM은 이들의 모델링을 위해서 각 오디오 클래스 별로 16개의 가우시안 확률밀도 함수를 사용하였다(M=16).

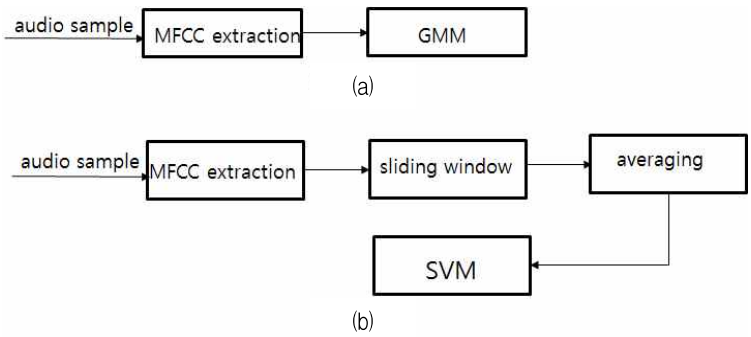

그림 1. GMM과 SVM에서의 오디오 특징 사용의 차이
Fig. 1. Difference in using audio features in GMM and SVM

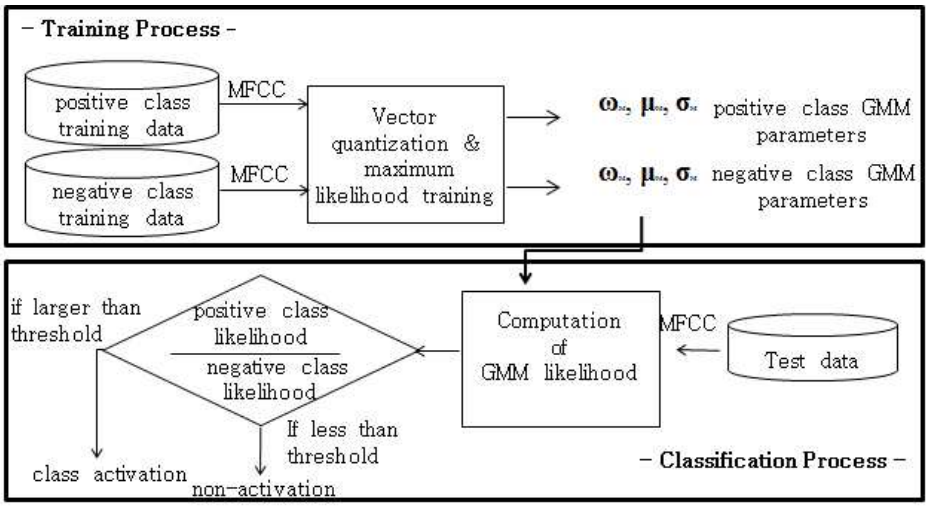

그림 2. 오디오 이벤트 검출을 위한 GMM의 학습과 인식과정
Fig. 2. Training and recognition process of GMM for audio event detection

학습과정에서는 GMM 모델의 가중치 ω_m, 평균 μ_m 그리고 공분산행렬의 대각원소값 σ_m 등의 파라미터 값을 구하기 위하여 59차의 MFCC 특징벡터에 대하여 벡터양자화를 하여 이들에 대한 초기값을 설정하도록 하였다.

그 후, 최대우도 기반의 학습과정을 통하여 최적의 파라미터 값을 추정하도록 하였다. 또한, 각 오디오 클래스 별로 바이너리 분류기를 구성하였으며 이를 위해서 각 클래스에 속하는 오디오 신호를 이용한 GMM 모델과 그 밖의 학습데이터를 이용한 GMM 모델을 각각 구성하였다. 인식과정에서는 두 개의 GMM 모델들로부터 얻어진 우도 값의 비교를 통해서 해당 오디오 클래스의 활성화 여부를 매 프레임마다 판단한다. 프레임마다의 활성화 출력은 스무딩 필터링을 거쳐서 최종적인 활성화 결과가 도출된다[4]. GMM을 이용한 학습과 인식과정이 그림 2에 요약되어 있다.

2.2.2 SVM

SVM은 두 개의 클래스 간의 경계를 최대화하는 패턴 학습기법을 사용하는 비확률적 분류 모델이다[10][11]. 그림 3에는 SVM의 작동 원리에 대한 설명이 나타나 있는데, 두 클래스 간의 결정 경계는 $w^Tx+b=0$로 표시되며 두 클래스의 훈련 샘플들의 가장자리를 나타내는 서포트벡터(Support Vector)들은 아래식과 같이 표시된다.

$$w^Tx+b=1 \text{ (positive 서포트벡터)} \quad (3)$$

$$w^Tx+b=-1 \text{ (negative 서포트벡터)} \quad (4)$$

SVM의 학습과정에서는 위에서 언급된 결정 경계와 서포트벡터 간의 거리를 최대화 하는 것을 목표로 하며, 이를 위한 목적함수를 수식적으로 표현하면 식 (5)과 같이 나타난다. 식 (5)의 최소화에서 식 (3)과 식 (4)는 전제 조건이 된다.

$$\min_{w} \| w \|^2 \quad (5)$$

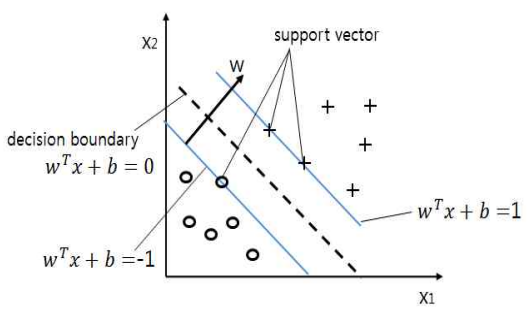

그림 3 SVM에서의 결정경계와 서포트벡터들
Fig. 3. Decision boundary and support vectors in SVM

한편, SVM에서는 슬랙변수 $\xi^{(i)}$을 통하여 학습데이터에 대해서 인식 오류를 허용함으로서 학습 알고리듬이 충분히 수렴할 수 있도록 한다. 슬랙변수가 목적함수에 끼치는 영향은 아래 식 (6)에서 보듯이 하이퍼파라미터 C에 의해서 조정되는데 C값이 크면 슬랙변수의 값들이 작아지도록 하여 학습데이터에 대한 인식오류에 대해서 엄격함이 강조되고 반면에 C값이 작으면 학습데이터에 대한 인식오류를 상대적으로 많이 허용하게 된다.

$$\min_{w,\xi^{(i)}} \frac{1}{2} \| w \|^2 + C\left(\sum_i \xi^{(i)}\right) \tag{6}$$

한편, SVM에서는 비선형 분류 문제를 해결하기 위해서 학습이나 인식과정에서 필요로 하는 데이터 샘플들 간의 내적(Dot Product)을 계산할시 커널함수를 대신 사용한다. 일반적으로 많이 사용되는 커널함수는 RBF이며 식 (7)에 나타나 있다.

$$k(x^{(i)},x^{(j)}) = \exp\left(-\gamma \| x^{(i)} - x^{(j)} \|^2\right) \tag{7}$$

식 (7)에서 γ 값을 증가시키면, 각 학습샘플이 결정 경계에 미치는 영향이 커지게 되며 반면에 γ 값을 감소시키면 각 학습샘플의 영향력이 작아져서 보다 부드러운 결정 경계가 형성된다.

인식과정에서는 앞선 학습단계에서 얻은 SVM의 하이퍼파라미터들을 사용하여 아래 식 (8)의 $d(x)$ 값이 양수일 경우 양(Positive)의 클래스로 인식하고 음수일 경우 부(Negative)의 클래스로 인식한다.

$$d(\boldsymbol{x}) = \sum_{i=1}^{I} \boldsymbol{\alpha}_i k(\boldsymbol{s}_i,\boldsymbol{x}) + b \tag{8}$$

여기서 $\boldsymbol{x}$는 입력 특징벡터이고, $\boldsymbol{\alpha}_i$와 $\boldsymbol{s}_i$, b는 각각 학습과정에서 얻은 가중치와 서포트벡터, 바이어스이며, I 는 서포트벡터의 수를 의미한다.

III. 실험 결과

3.1 데이터베이스

오디오 이벤트 검출을 위한 인식실험을 위해서 DCASE 2016의 Task 3에서 제공된 학습 및 테스트 데이터를 이용하였다[4]. 여기에는 두 종류의 일상적인 환경(실외와 실내)으로 부터의 다양한 오디오 소리들이 포함되어 있다. 실외 환경의 소리는 거주지 주변에서 발생하는 대표적인 7개의 오디오 클래스로 구성되어 있으며 실내 환경에서는 실내에서 발생하는 11개의 오디오 클래스로 이루어져 있다.

표 1에는 전체 18개 (7+11) 오디오 클래스의 종류와 발생빈도를 보여주고 있다. 표에서 보다시피 각 오디오 클래스 별로 발생빈도가 매우 다르다는 것을 알 수 있다. 실외 환경의 경우에는 “bird sing” 클래스의 발생빈도는 “object(bang)”의 경우보다 10배나 많다. 또한, 실내의 경우에는 “object impact”와 “dishes”가 전체 발생빈도의 40%를 차지하고 있다.

표 1. DCASE‘2016 Task 3의 오디오 이벤트 클래스의 종류와 발생빈도
Table 1. Audio event classes and their number of instances in the Task 3 of DCASE’2016

Residential		Home	
classes	# of instances	classes	# of instances
(object)banging	23	(object) rustling	60
bird singing	271	(object) snapping	57
car passing by	108	cupboard	40
children shouting	31	cutlery	76
people speaking	52	dishes	151
people walking	44	drawer	51
wind blowing	30	glass jingling	36
		object impact	250
		people walking	54
		washing dishes	84
		water tap running	47

오디오 이벤트 분류 결과를 나타내기 위한 준거(Metric)로는 DCASE 2016에서 이용되었던 F-score와 ER(Error Rate)를 사용하였으며, 1초 단위의 블록별로 준거를 계산하는 세그먼트(Segment) 기반 방식을 적용하였다[12].

3.2 인식 결과

표 2에는 SVM에서 오디오 특징의 평균 구간길이에 따른 성능변화를 보였으며 비교를 위해서

GMM의 인식결과와 SVM에서 특징에 대한 평균을 사용하지 않았을 경우도 함께 나타내었다. GMM과 SVM 모든 20ms 마다 발생하는 MFCC 특징을 사용하고 클래스별 독립적인 인식 모델을 만드는 바이너리 분류기라는 공통점이 있으나, GMM의 경우에는 20ms 마다의 MFCC에 대한 식 (1)의 로그-우도 값을 1초 길이동안 합산하여 분류 결정을 하는 구조인 반면에, SVM의 경우에는 20ms 마다의 MFCC의 값을 직접 SVM에 입력하거나 어느 구간 동안 평균한 다음 SVM에 입력하고 식 (8)의 결과 값에 따라서 분류를 수행하게 된다. 두 분류기 모두 매 프레임마다 분류 결정이 출력되고 이를 스무딩하여 최종적인 결과를 얻게 된다.

표 2의 결과에서 GMM은 60차원 MFCC 계수에서 c0를 뺀 59차를 활용하고 SVM의 경우에는 c0부터 c12 까자의 13차의 MFCC를 사용하였다.

표 2의 결과를 통해서 SVM은 입력 특징의 평균 길이가 80ms에서 F-스코어 값이 가장 좋은 것을 알 수 있다.

이를 통해서 SVM의 경우 입력 특징을 평균함으로서 성능향상이 발생하지만 평균길이의 적정 값을 찾는 것이 중요하다는 것을 알 수 있다. 한편, GMM과 비교해서 SVM의 F-스코어 값은 대체적으로 향상됨을 알 수 있다.

한편, SVM에서 사용되는 MFCC의 특징차원에 따라서 성능이 어떻게 변화하는지 조사하기 위해서 특징평균을 적용하지 않은 상태에서 MFCC의 특징차원을 달리 하면서 인식 성능의 변화를 살펴보았고 그 결과를 표 3에 나타내었다. 표 3에 나타난 그 결과로부터 SVM의 경우에는 GMM과는 다르게 특징차원이 증가한다고 성능이 향상되지 않으며 오히려 13차의 MFCC를 사용할 경우 가장 좋은 성능을 보임을 알 수 있었다.

표 2과 3에서는 SVM의 주요 파라미터인 γ와 C 값을 기본 값으로 고정해 놓고 인식실험을 하였는데 이들 값의 변화를 통해서 SVM의 성능에 어떤 변화가 있는 살펴보았다. 먼저, 표 3에는 식 (6)에서 슬랙변수의 영향을 조절하는 하이퍼파라미터 C값에 따른 인식성능의 변화를 나타내었다.

표 4의 결과에서 C=1일 경우에 가장 높은 F-score 값을 얻을 수 있음을 알 수 있으며 C 값이 0.1 에서 10 사이의 변화를 가져도 성능에는 큰 차이가 없음을 알 수 있었다.

표 2. SVM에서 특징 평균길이에 따른 인식성능의 비교 ($\gamma = 1/$특징차원, $C = 1$)
Table 2. Performance comparison in SVMs depending on the time-length of the feature averaging

	F-score(%)	ER	feature dimension	length of averaging
GMM	24.3	0.97	59	-
SVM	26.6	0.96	13	None
	31.4	1.14	13	60 ms
	32.3	1.21	13	80 ms
	25.5	0.97	13	100 ms

표 3. SVM에서 MFCC 특징차원에 따른 인식성능의 비교 ($\gamma = 1/$특징차원, $C = 1$)
Table 3. Performance comparison in SVMs depending on the dimension of MFCC

	F-score(%)	ER	feature dimension	length of averaging
SVM	26.6	0.96	13	None
	19.0	0.95	20	None
	13.3	0.97	40	None
	11.0	1.00	60	None

표 4. SVM에서 파라미터 C 값에 따른 인식성능의 비교 (특징차원=13, 특징 평균길이= 80*ms*)
Table 4. Performance comparison in SVM depending on the value of C

	F-score(%)	ER	γ(=1/13)	C
SVM	17.5	0.92	0.0769	0.01
	31.7	1.13	0.0769	0.1
	32.3	1.21	0.0769	1
	31.1	1.49	0.0769	10
	30.4	1.57	0.0769	100

표 5. SVM에서 파라미터 γ 값에 따른 인식성능의 비교 (특징차원=13, 특징 평균길이= 80*ms*)
Table 5. Performance comparison in SVM depending on the value of γ.

	F-score(%)	ER	γ(=1/13)	C
SVM	24.1	0.93	0.0000769	1
	31.0	0.96	0.000769	1
	34.1	1.04	0.00769	1
	32.3	1.21	0.0769	1

표 6. SVM($\gamma = 0.00769$, $C = 1$)과 GMM간의 인식 성능의 비교
Table 6. Performance comparison between SVM and GMM

	F-score(%)	ER	feature dimension	length of averaging
GMM	24.3	0.97	59	-
SVM	34.1	1.04	13	80ms

한편, 표 5에서는 γ의 변화가 인식성능에 미치는 영향을 나타내었다. C 값과 마찬가지로 γ의 경우에도 $0.000769 \le \gamma \le 0.0769$의 범위에서는 성능의 변화가 크게 나타나지 않고 안정적으로 유지되는 것을 볼 수 있었다.

위에서 진행된 다양한 실험결과 SVM의 경우에는 MFCC 특징의 평균길이를 80ms로 하고 특징의 차원을 13차로 하며, $C = 1, \gamma = 0.00769$일 경우 최적의 성능을 나타냄을 알 수 있었다. 표 6에는 이러한 실험 조건하에서의 SVM의 결과와 GMM 간의 성능비교를 나타내었으며 SVM이 GMM에 비해서 F-스코어에서 상당한 성능 우위를 보임을 알 수 있었다.

Ⅳ. 결 론

최근 들어 실생활에서의 다양한 응용 가능성으로 인하여 오디오 이벤트 검출 기법에 대한 많은 관심이 생겨나고 있다. GMM과 SVM은 비교적 적은 양의 학습데이터와 계산량 으로도 높은 성능을 보인다는 장점으로 인하여 오디오 이벤트 검출에서 널리 사용되고 있다.

GMM의 경우에 비해서 SVM의 경우에는 특징의 입력 방법이나 특징차원의 수 그리고 모델 파라미터 값 등이 인식성능에 미치는 영향이 크다고 알려져 있다. 그러나 이와 같은 다양한 실험조건들이 성능에 미치는 영향에 대한 연구결과가 오디오 이벤트 검출 분야에서는 지금까지 충분히 제시되지 않았다. 따라서 본 연구에서는 이러한 점에 착안하여 이러한 다양한 조건하에서의 SVM의 성능변화를 관찰하였다.

인식실험결과 SVM의 입력특징은 80ms 구간 동안 평균을 취한 경우와 13차의 MFCC를 사용한 경우에 가장 좋은 성능을 보임을 알 수 있었다. 레귤러라이제이션 조절 파라미터인 C 값은 일반적으로 많이 사용하는 디폴트 값인 C=1에서 최적의 성능을 보이는 것을 알 수 있었으며 RBF 커널 함수에서 사용되는 γ의 경우에는 일반적으로 권장되는 디폴트 값인 $\gamma = 0.0769 (= 1/\text{특징차원})$에 비해서 더 작은 값인 $\gamma = 0.00769$에서 최적의 성능을 보임을 알 수 있었다. 그러나 상당한 범위의 γ 값에 대해서 인식성능이 크게 변동하지 않음을 알 수 있었으며, 또한 이러한 넓은 범위 γ 값에 대해서 SVM은 GMM에 비해서 일관된 성능 향상을 보임을 확인할 수 있었다.

본 논문의 실험결과 SVM의 입력으로 사용되는 MFCC의 경우 13차에서 최고의 성능을 보이며 차수가 증가할수록 성능의 저하가 발생하였다. 향후 연구에서는 PCA(Principal Component Analsyis) 나 LDA(Linear Discriminant Analysis) 기법을 통하여 MFCC 차수를 증가시키더라도 성능 향상이 일어나는 방안에 대해서 고찰하고자 한다. 또한, 학습이나 인식에 사용된 오디오데이터를 보다 확대하여 본 연구에서 얻어진 결과들의 신뢰성을 높이는 방법에 대해서도 향후 연구에서 진행할 예정이다.

References

[1] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains", Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pp. 6460-6464, Mar. 2016.

[2] S. Ntalampiras and I. Potamitis, "Detection of human activities in natural environments based on their acoustic emissions", Signal Processing Conference, 2012 20th European, pp. 1469-1473, Sep. 2012.

[3] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events", IEEE Trans. Multimedia, Vol. 17, No. 10, pp. 1733-1746, Oct. 2015.

[4] A. Mesaros, T. Heittola, and T. Virtanen, "TUT

Database for acoustic scene classification and sound event detection", in Signal Processing Conference, 2016 24th European, pp. 1128-1132 Sep. 2016.

[5] J. J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music", J. Acoust. Soc. America, Vol. 122, No. 2, pp. 881-891, Jul. 2007.

[6] S. H. Chung and Y. J. Chung, "A Comparison between Methods for Scream Detection Based on SVM and GMM", Journal of KIIT. Vol. 15, No. 3, pp. 65-71, Mar. 2017.

[7] J. H. Park, J. Y. Lim, J. Y. Yang, J. M. Kyung, and M. S. Hahn, "False Positive Movie Clip Decision in Black-box Using Car Door-Closing Sound Classification", IEIE, Vol. 37, No. 1, pp. 761-763, Jun. 2014.

[8] W. Huang, T. K. Chiew, H. Li, T. S. Kok, and J. Biswas, "Scream detection for home applications", Industrial Electronics and Applications (ICIEA), 2010 the 5th IEEE Conference on, No. 399, pp. 2115-2120, Jun. 2010.

[9] B. Lei and M. W. Mak, "Sound-event partitioning and feature normalization for robust sound-event detection", Digital Signal Processing (DSP), 2014 19th International Conference on, pp. 389-394, Aug. 2014.

[10] Support Vector Machines for Binary Classification, http://kr.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html?requestedDomain=www.mathworks.com. [Accessed: Sep. 02, 2016]

[11] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines", ACM Transactions on Intelligent Systems and Technology, Vol. 2, No. 3, pp. 1-27, Sep. 2011.

[12] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection", Applied Sciences, Vol. 6, No. 6, pp. 162-178, May 2016.

저자소개

정 석 환 (Suk-Hwan Chung)

2017년 3월 ~ 현재 : 계명대학교 전자공학과 대학원 석사과정
관심분야 : 머신러닝 및 오디오 분류

정 용 주 (Yong-Joo Chung)

1995년 8월 : 한국과학기술원 (공학박사)
1999년 3월 ~ 현재 : 계명대학교 전자공학과 교수
관심분야 : 오디오 분류, 음성인식