상황 인식 모바일 컴퓨팅을 위한 사운드 분류 시스템 설계

김주희*, 이석준*, 김인철**
*경기대학교 컴퓨터과학과 학부생
**경기대학교 컴퓨터과학과 교수
e-mail:jh.k@kyonggi.ac.kr

Design of a Sound Classification System for Context-Aware Mobile Computing

Joo-Hee Kim*, Seok-Jun Lee*, In-Cheol Kim**
*Undergraduate Course, Dept of Computer Science, Kyonggi University
**Faculty, Dept of Computer Science, Kyonggi University

요 약

본 논문에서는 스마트폰 사용자의 실시간 상황 인식을 위한 효과적인 사운드 분류 시스템을 제안한다. 이 시스템에서는 PCM 형태의 사운드 입력 데이터에 대한 전처리를 통해 고요한 사운드와 화이트 노이즈를 학습 및 분류 단계 이전에 미리 여과함으로써, 계산 자원의 불필요한 소모를 막을 수 있다. 또한 에너지 레벨이 낮아 신호의 패턴을 파악하기 어려운 사운드 데이터는 증폭함으로써, 이들에 대한 분류 성능을 향상시킬 수 있다. 또, 제안하는 사운드 분류 시스템에서는 HMM 분류 모델의 효율적인 학습과 적용을 위해 k-평균 군집화를 이용하여 특징 벡터들에 대한 차원 축소와 이산화를 수행하고, 그 결과를 모아 일정한 길이의 시계열 데이터를 구성하였다. 대학 연구동내 다양한 일상생활 상황들에서 수집한 8 가지 유형의 사운드 데이터 집합을 이용하여 성능 분석 실험을 수행하였고, 이를 통해 본 논문에서 제안하는 사운드 분류 시스템의 높은 성능을 확인할 수 있었다.

1. 서론

스마트폰 사용자의 실시간 상황 인식 기술은 헬스케어, 안전 지키미, 소셜 네트워크, 모바일 게임 등 매우 다양한 분야에 유용하게 활용될 수 있다. 실시간 상황 인식을 위한 센서 데이터는 카메라, 마이크로폰, 가속도 센서, 방향 센서, 근접 센서 등 다양한 종류의 스마트폰 내장 센서들을 이용해 얻을 수 있다. 특히 마이크로폰 센서를 이용해 수집할 수 있는 사운드 데이터는 카메라 영상이나 가속도 데이터 등에 비해 주변 환경의 조명 상태나 스마트폰의 위치 및 방향에 비교적 영향을 적게 받으면서도, 사용자가 있는 장소나 수행 활동을 추정해볼 수 있는 풍부한 단서들을 포함하고 있다. 하지만 일상생활 속에서 수집되는 사운드 데이터에는 많은 소음이 혼재되어 있으며, 때로는 의미 있는 사운드의 볼륨이 너무 낮아 상황 인식이 어려운 경우가 많다. 또, 오랜 시간 사용자의 실시간 상황을 추적하기 위해서는 사운드 데이터 처리에 소모되는 스마트폰 에너지 문제도 해결해야 할 숙제로 남아있다.

본 논문에서는 스마트폰 사용자의 실시간 상황 인식을 위한 효과적인 사운드 분류 시스템을 제안한다. 본 시스템에서는 일상생활 속에서 수집되는 사운드 데이터에 대한 실시간 분류 성능을 향상하기 위해, 사운드 필터링과 증폭을 위한 전처리를 수행한다. 사운드 필터링 단계에서는 분류 작업이 필요하지 않은 고요한 사운드(silence)와 화이트 노이즈(white noise)들을 걸러내고, 사운드 증폭 단계에서는 의미 있는 사운드 이벤트를 포함하고 있으나 볼륨이 너무 낮아 인식이 어려운 사운드들, 즉 에너지 레벨이 낮은 사운드들을 증폭한다. 이러한 사운드 필터링은 불필요한 분류 계산을 줄여줌으로써 스마트폰의 에너지 효율성을 높여줄 수 있고, 사운드 증폭은 에너지 레벨이 낮은 사운드의 분류 성능 향상에 도움을 줄 수 있다. 또한, 본 논문에서 제안하는 사운드 분류 시스템에서는 HMM (Hidden Markov Model) 분류 모델의 효율적인 학습과 적용을 위해 k-평균 군집화를 적용하여 특징벡터들에 대한 차원 축소와 이산화를 수행한다. 대학 연구동내 다양한 일상생활 상황들에서 수집한 8 가지 유형의 사운드 데이터 집합을 이용하여, 본 논문에서 제안하는 시스템의 성능 분석 실험을 전개하고 그 결과를 소개한다.

※ 본 연구는 경기도의 경기도지역협력연구센터사업의 일환으로 수행하였음

2. 관련 연구

에너지 레벨이 낮은 사운드들의 분류 성능을 향상시키고자 노력한 선행연구에는 [1][2]가 있다. [1]의 연구에서는 전처리 단계에 입회 조절기(admission control)를 두고 에너지 레벨이 낮은 사운드 데이터 중 에너지 피크가 발생하는 스펙트럼을 추출해 낸다. 이러한 방법은 고요한 사운드와 에너지 레벨이 낮은 사운드 이벤트를 구분할 수 있기 때문에, 의미 있는 사운드를 놓치지 않을 수 있다.

뿐만 아니라 고요한 사운드와 화이트 노이즈의 필터링이 가능하기 때문에 계산 효율성 향상에도 기여한다. 하지만 걷기와 같은 낮은 에너지 레벨의 사운드 유형은 재현율(recall)이 낮은 것을 볼 때, 고요한 사운드와 구분이 쉽지 않음을 알 수 있다. 또한 특징 추출 단계에서도 에너지 레벨 변화에 견고한 특징들을 추출하여 분류 성능을 높였다. [2]의 연구에서는 낮은 에너지 레벨의 사운드 이벤트를 스펙트럼의 변화로부터 알아낸다. 스펙트럼의 변화를 감지하는 특징을 추출하여 변화가 적다면 사운드 이벤트가 없는 것으로 판단하여 여과한다.

모바일 기기의 제한적인 계산 자원 문제를 해결하고자 한 선행연구에는 [3][4]가 있다. [3]의 연구에서는 모바일 기기의 계산 자원을 사용하지 않고, 서버에서 계산을 진행하는 서버 모드 방식을 추가하여 계산 자원의 효율성을 높였다. [4]의 연구에서는 고요한 사운드가 지속되면 센싱 간격을 늘리고, 낮은 에너지 레벨의 사운드 데이터는 가용 정보가 부족하다고 판단하여 필터를 통해 버린다. 이러한 방법은 계산 자원의 효율성 향상에는 도움이 될 수 있으나, 의미 있는 사운드를 센싱하지 못하거나 의미 있는 사운드라도 에너지 레벨이 낮다면 버려질 수 있는 문제가 발생한다. 한편, [4]의 연구에서는 분류 단계에서도 음성과 비음성에 대해 계층적 분류를 사용함으로써 불필요한 계산을 줄여 자원 효율성을 높이려는 시도를 하였다.

3. 사운드 분류 시스템

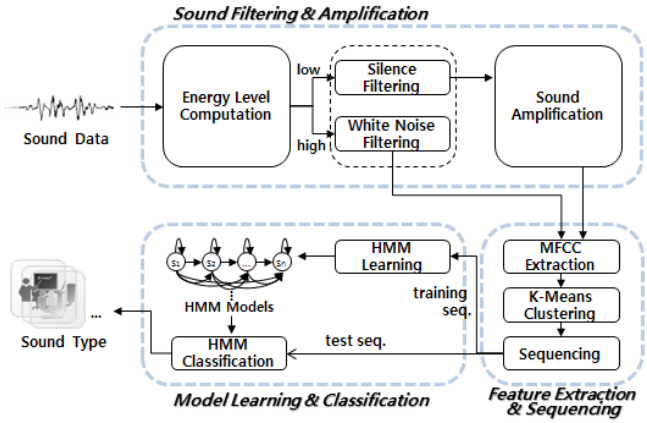
3.1 시스템 개요

본 논문에서는 스마트폰 사용자의 실시간 상황 인식을 위한 방법으로 효과적인 사운드 분류 시스템을 제안한다. 본 시스템에서 분류하고자 하는 사운드 유형은 <표 1>과 같다. 이들은 강의실, 연구실, 복도, 화장실 등에서 스마트폰 사용자가 강의듣기, 필기하기, 컴퓨터 작업하기, 물 마시기, 이야기하기, 걷기, 화장실 물 내리기, 손 씻기 등의 활동을 하는 상황에서 발생하는 전형적인 사운드 종류들이다.

<표 1> 사운드 종류와 발생 상황

Sound Type	Context	
	Activity	Location
S1	Lecture	Classroom
S2	Writing	Office
S3	Typing	Office
S4	Drinking	Office
S5	Talking	Classroom, Office
S6	Walking	Hallway, Office
S7	Flushing	Restroom
S8	Washing hands	Restroom

사운드 분류를 위한 전체 시스템 구성은 (그림 1)에서 보는 바와 같이 크게 사운드 입력 단계, 사운드 필터링 및 증폭 단계, 특징 추출 및 순차화 단계, 모델 학습 및 분류 단계로 이루어진다. 사운드 입력 단계에서는 각 상황에서 발생하는 사운드가 스마트폰의 마이크로폰 센서를 통하여 16kHz 16비트 스테레오 PCM 형태로 입력되며, 입력된 PCM 데이터는 30ms 길이의 프레임 단위로 나뉜다. 사운드 필터링 및 증폭 단계에서는 에너지 레벨에 따른 필터를 적용함으로써 입력 사운드에서 고요한 사운드와 화이트 노이즈 부분을 찾아서 버린다. 필터링 과정을 통과한 사운드 데이터 중 특히 에너지 레벨이 낮은 데이터는 사운드 증폭기를 통해 700% 증폭한다. 특징 추출 및 순차화 단계에서는 필터링 및 증폭 단계를 거친 사운드 데이터로부터 13차원의 MFCC로 구성된 특징 벡터들을 추출한 뒤, K-평균 군집화(K-means clustering)를 적용하여 각 특징 벡터를 소속 군집 번호로 변환한다.



(그림 1) 전체 시스템 구성

이 과정을 통해 13차원 MFCC 실수 벡터가 1차원 정수로 바뀌는 차원 축소(dimension reduction) 및 이산화(discretization)가 수행된다. 그리고 이렇게 변환된 입력 데이터들을 모아 시계열 학습을 위한 일정한 길이의 순차 데이터를 만드는 순차화 과정이 수행된다. 모델 학습 및 분류 단계에서는 훈련용 순차 데이터 집합으로부터 각 사운드 유형별 HMM 모델을 학습하고, 이 모델을 새로운 입력 사운드 데이터에 적용함으로써 해당 사운드의 유형을 자동으로 판별한다.

3.2 사운드 필터링 및 증폭

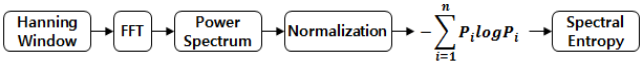
사운드 필터링 단계에서는 프레임 단위로 나뉜 입력 사운드 데이터에 대해 각각의 에너지 레벨을 계산하고, 해당 데이터의 에너지 레벨에 따라 고요한 사운드와 화이트 노이즈를 여과하기 위한 사운드 필터를 선택적으로 적용한다. 고요한 사운드(silence)는 특별한 소음조차도 없어 에너지 레벨이 매우 낮은 상태의 사운드를 말하고, 화이트 노이즈(white noise)는 의미 없는 소음들만 가득하여 에너지 레벨이 높은 상태의 사운드를 말한다. 여과 대상 사운드 데이터를 판별하기 위해 서로 다른 두 개의 척도, 제곱평균 제곱근(Root Means Square, RMS)과 스펙트럼 엔트로피(Spectral Entropy)를 사용한다.

RMS은 사운드의 볼륨(volume), 즉 에너지 레벨(energy level)을 측정하는 척도이며, (식 1)과 같이 계산한다.

$$f_{rms} = \sqrt{\frac{1}{T}\int_0^T [f(t)]^2\, dt} \qquad \text{(식 1)}$$

스펙트럼 엔트로피는 사운드의 스펙트럼 분포 패턴을 측정하는 척도이며, (그림 2)과 같이 구할 수 있다. 이 척도를 통해 측정하고자 하는 사운드 구간의 스펙트럼 피크 발생 정도를 알 수 있기 때문에, 분류하고자 하는 사운드 유형들과 플랫(flat)한 스펙트럼을 구분할 수 있다.

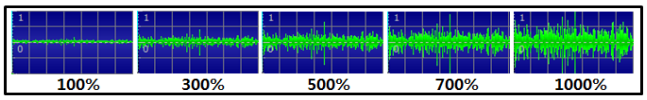
필터에는 고요한 사운드 필터와 화이트 노이즈 필터가



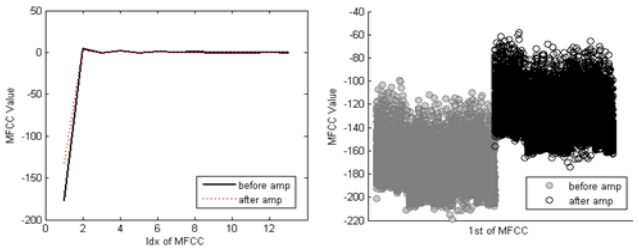
(그림 2) 스펙트럼 엔트로피 계산 방법

있다. 고요한 사운드 필터는 에너지 레벨이 낮은 사운드 데이터를 입력받아 고요한 사운드를 여과한다. 여과 방법은 RMS 임계값과 스펙트럼 엔트로피 임계값을 이용하여, 각 특징이 임계값보다 작을 경우 여과한다. 화이트 노이즈 필터는 에너지 레벨이 높은 사운드 데이터를 입력받아 화이트 노이즈를 여과한다. 여과 방법은 스펙트럼 엔트로피 임계값을 이용하여, 해당 특징 값이 임계값보다 작을 경우 여과한다.

에너지 레벨이 낮은 데이터가 필터를 통과하여 여과가 완료되면, 증폭기를 통해 증폭을 수행한다. 증폭을 수행하면 (그림 3)과 같이 패턴 변화가 미약한 왼쪽의 신호에서, 패턴의 변화가 명확한 오른쪽의 신호들로 각 증폭률에 따라 변환된다.



(그림 3) 증폭에 따른 사운드 신호 변화



(그림 4) 증폭에 따른 MFCC 변화

하지만 무한대로 증폭률을 높인다면 16비트로 제한되어 있는 진폭을 넘어서게 되어, 원래의 사운드 데이터가 왜곡되는 클리핑(clipping) 현상이 발생할 수 있다. 본 시스템에서는 사전 실험을 통해, 데이터가 왜곡되지 않는 범위 안에서 적정한 증폭률을 700%로 정하였다. 또한 증폭을 거친 데이터는 데이터로부터 추출되는 특징에도 영향을 끼친다. 증폭된 데이터로부터 추출된 13차원의 MFCC 특징 벡터는 (그림 4)에서 볼 수 있듯이 첫 번째 계수의 값만이 증폭전과 다르게 변화하는 것을 볼 수 있다.

3.3 특징 추출 및 순차화

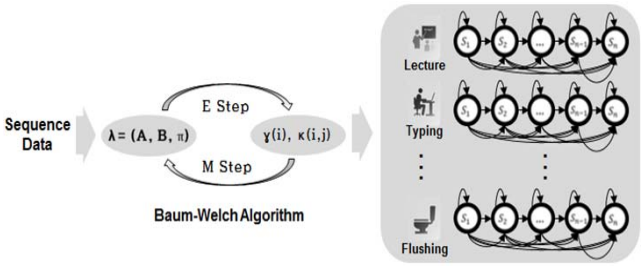
본 논문에서 특징으로 사용하는 MFCC는 사람의 청각이 저주파에서는 민감하고 고주파에서는 둔감한 특성, 즉 멜(mel) 스케일을 따르는 청각적 특성을 반영한 캡스트럼 계수이다. MFCC는 사운드 신호 처리에 뛰어난 성능을 보이기 때문에, 사운드 분류에 가장 널리 사용되는 특징 중 하나이다. MFCC는 (그림 5)와 같은 과정을 거쳐 추출할 수 있다.

(그림 5) MFCC 추출 과정

추출된 13차원의 MFCC 특징 벡터는 사운드 유형별로 라벨링한 후, 순차화시켜 HMM의 입력 데이터로 사용할 수 있다. 하지만 13차원의 실수 벡터는 HMM 학습 알고리즘의 입력 데이터로 사용하기에는 계산 부담이 크다. 이러한 이유로 13차원의 실수 벡터를 1차원 정수로 변환하는데, 변환에는 K-평균 군집화를 사용한다. K-평균 군집화는 주어진 수치 데이터를 K개의 군집으로 묶는 알고리즘으로, 각 군집과 거리 차이의 분산을 최소화하는 방식이다. 13차원의 각 실수 벡터는 K-평균 군집화를 통해 가장 가까운 군집의 색인 번호로 변환된다. 이렇게 1차원 정수로 변환된 특징 벡터들은 HMM 모델 학습을 위해, 150개의 특징 벡터가 하나의 시퀀스로 연결되어 HMM 학습 알고리즘의 입력 데이터가 된다.

3.4 모델 학습 및 분류

사운드 분류를 위해 사용되는 HMM 모델은 은닉 상태들 간의 상태 전이, 은닉 상태와 관측들 간의 상호 의존성을 확률적으로 잘 표현할 수 있다. 이와 같은 HMM 모델은 (그림 6)과 같은 과정을 거쳐 학습된다.



(그림 6) 사운드 유형별 HMM 모델 학습

HMM 모델 학습을 위해 결정해야 하는 상태 수와 관측 수는 사전 실험을 통해 각각 8개와 20개로 정하였다. 또한 모델 구조는 상태 전이가 왼쪽에서 오른쪽으로만 발생하는 좌-우 모델을 사용하였다. 각 사운드 유형별 HMM 모델 학습을 위해서는 Baum-Welch 알고리즘을 적용하였고, 이와 같은 방법으로 분류하고자 하는 사운드 유형별로 1개씩, 총 8개의 HMM 모델을 학습하였다. 분류하고자 하는 새로운 테스트 사운드가 발생하면 각 사운드별로 학습된 HMM 모델들에 입력으로 주어지고, 각 HMM 모델별로 테스트 데이터에 대한 로그 우도(log likelihood)가 계산된다. 그리고 이 때 최대 로그 우도를 갖는 HMM 모델에 따라 테스트 사운드의 유형이 자동 판별된다.

4. 실험 및 평가

본 논문에서 제안하는 사운드 분류 시스템의 성능을 분석하기 위해, 안드로이드 스마트폰을 이용한 사운드 분류 시스템을 구현하였다. 실험을 위한 사운드 데이터는 스마트폰 내장 마이크로폰 센서를 이용하여 각 상황별로 5초씩 15분 분량을 수집하였다. 수집된 샘플의 개수는 상황별로 총 180개이고, 그 중 훈련 데이터로 90개, 테스트 데이터로 90개를 사용하였다.

실험은 크게 세 가지로 진행하였다. 첫 번째 실험은 고요한 사운드 필터의 성능을 분석하기 위한 목적으로 수행하였다. 이를 위해 필터의 중요한 매개변수인 스펙트럼 엔트로피의 임계값을 변경해가면서, 에너지 레벨은 낮으나 유의미한 사운드와 고요한 사운드를 얼마나 정확히 구분하여 여과할 수 있는 지 비교 분석하였다.

<표 2> 스펙트럼 엔트로피 임계값에 따른 고요한 사운드 필터 정확도

	Energy Low-Level Sound	Silence
Energy Low-Level Sound	**96.37%** (292/303)	**20.79%** (63/303)
Silence	**3.63%** (11/303)	**79.21%** (240/303)

(a) Threshold = 0.9

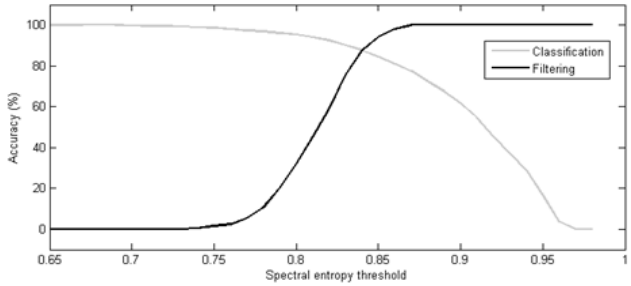
	Energy Low-Level Sound	Silence
Energy Low-Level Sound	**95.05%** (288/303)	**8.58%** (26/303)
Silence	**4.95%** (15/303)	**91.42%** (277/303)

(b) Threshold = 0.91

	Energy Low-Level Sound	Silence
Energy Low-Level Sound	**93.07%** (282/303)	**5.94%** (18/303)
Silence	**6.93%** (21/303)	**94.06%** (285/303)

(c) Threshold = 0.92

<표 2>의 실험 결과에서 볼 수 있듯이, 스펙트럼 엔트로피의 임계값이 소폭 증가할수록 전체적으로 고요한 사운드로 판정한 경우의 수는 점차 증가하고, 반대로 판정한 경우의 수는 감소한 것을 알 수 있다. 또, 임계값이 0.9~0.92로 증가할수록 고요한 사운드의 올바른 판정율은 79.21%~94.06%로 증가하였으나, 반대로 유의미한 사운드의 올바른 판정율은 96.37%~93.07%로 감소하였다. 따라서 본 실험에서는 종합적 관점에서 임계값이 0.91일 때, 고요한 사운드와 유의미한 사운드에 대한 올바른 판정율이 가장 높은 것으로 판단할 수 있다.

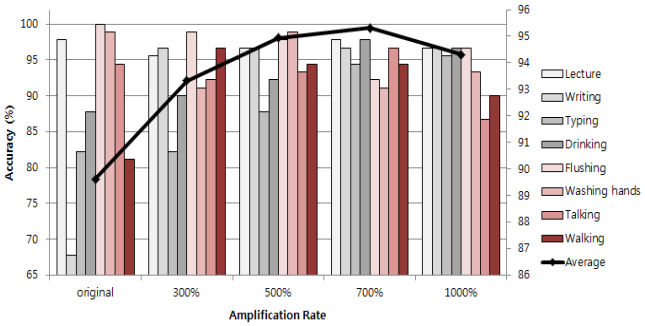


(그림 7) 스펙트럼 엔트로피 임계값에 따른 분류와 필터의 정확도

두 번째 실험은 화이트 노이즈 필터의 성능을 분석하기 위한 목적으로 수행하였다. 이를 위해 스펙트럼 엔트로피의 임계값을 변경해가면서, 화이트 노이즈 필터의 정확도와 사운드 분류기의 정확도를 함께 비교 분석하였다.

실험 결과를 나타내는 (그림 7)에서 검은 실선은 화이트 노이즈 필터의 정확도를 나타내며, 회색 실선은 사운드 분류기의 정확도를 나타낸다. 이 그림을 통해, 스펙트럼 엔트로피의 임계값이 커질수록 화이트 노이즈 필터의 성능은 증가하지만, 반대로 사운드 분류기의 성능은 감소한 결과를 확인할 수 있다. 따라서 본 실험에서는 필터의 성능과 분류기의 성능이 서로 교차하는 점인 임계값 0.84일 때, 필터와 분류기 양쪽 면을 모두 고려한 최대 성능을 얻을 수 있었다.

세 번째 실험은 증폭률에 따라 사운드의 분류 성능이 어떻게 변화하는지를 분석하기 위한 목적으로 수행하였다. 이를 위해 이 실험에서는 입력 사운드 데이터의 증폭률을 변경해가면서, 사운드 유형별 분류 정확도와 평균 분류 정확도를 비교 분석하였다. (그림 8)의 실험결과에서 보는 바와 같이, 증폭률이 700%일 때 사운드 유형별 분류 정확도가 모두 90% 이상인 높은 성능을 보였다. 특히 증폭 이전의 분류 정확도 평균이 90% 미만이었던 것과 비교해볼 때, 700% 증폭을 수행한 후의 분류 정확도는 평균 6% 이상 향상되었음을 알 수 있다.



(그림 8) 증폭률에 따른 분류 정확도

5. 결론

본 논문에서는 스마트폰 사용자의 실시간 상황 인식을 위한 효과적인 사운드 분류 시스템을 제안하였다. 이 시스템에서는 입력된 PCM 형태의 사운드 데이터에서 고요한 사운드와 화이트 노이즈를 여과함으로써, 불필요한 계산 자원에 대한 소모를 감소시켰다. 또한 에너지 레벨이 낮아 신호 패턴을 파악하기 어려웠던 사운드 데이터를 증폭함으로써 분류 성능을 높이는 효과를 얻었다. 또, 효율적인 HMM 분류 모델의 학습과 적용을 위해 13차원의 MFCC 특징벡터들에 대한 차원 축소와 이산화를 수행하였다. 다양한 일상생활 상황들에서 수집한 사운드 데이터 집합을 이용한 실험을 통해, 본 논문에서 제안한 시스템의 높은 성능을 확인할 수 있었다.

참고 문헌

[1] Hong Lu, Wei Pan, Nicholas D. Lane, Tanzeem Choudhury, Andrew T. Campbell, "SoundSense: Scalable Sound Sensing for People-Centric Applications on Mobile Phones," Proc. of MobiSys-09, pp. 165-78, 2009.

[2] Lie Lu, Hao Jiang, Hongjiang Zhang, "A Robust Audio Classification and Segmentation Method," Proc. of ACM Multimedia, pp. 203-211, 2001.

[3] Mirco Rossi, Sebastian Feese, Oliver Amft, Nils Braune, Sandro Martis, Gerhard Troster, "AmbientSense: A Real-Time Ambient Sound Recognition System for Smartphones," Proc. of IEEE International Conf. on Pervasive Computing and Communications Workshops, pp. 230-235, 2013.

[4] Hong Lu, Jun Yang, Zhigang Liu, Nicholas D. Lane, Tanzeem Choudhury, Andrew T. Campbell, "The Jigsaw Continuous Sensing Engine for Mobile Phone Applications," Proc. of SenSys-10, pp. 71-84, 2010.

[5] L.Ma, D.J.Smith, B.P.Milner, "Context Awareness Using Environmental Noise Classification," Proc. of EuroSpeech-03, Geneva, pp. 2237-2340, 2003.

[6] Sachio Teramoto, Jun Noda, "O-MUSUBI: Ad-hoc Grouping System Enhanced by Ambient Sound-The Similarity based on Information Theoretical Features for Sound-Fields," Proc. of ThinkMind, ICONS, IARIA conf, Spain, The Eighth International Conference on Systems, pp. 52-58, 2013.