Journal of the Korea Academia-Industrial cooperation Society
Vol. 13, No. 2 pp. 817-821, 2012

http://dx.doi.org/10.5762/KAIS.2012.13.2.817

# 환경음 인식을 위한 GMM의 혼합모델 개수 추정

한다정[1], 박아론[1], 백성준[1*]
[1]전남대학교 전자컴퓨터공학부

# Estimation of Optimal Mixture Number of GMM for Environmental Sounds Recognition

**Da-Jeong Han[1], Aaron Park[1] and Sung-June Baek[1*]**
**[1]Division of Electronic and Computer Engineering, Chonnam University**

**요 약** 본 논문에서는 환경음 인식에 GMM(Gaussain mixture model)을 이용할 때 MDL(minimum description length)와 BIC(Bayesian information criterion) 모델선택 기준을 이용하여 최적의 혼합모델 개수를 결정하는 방법에 대해 다루었다. 실험은 모두 9가지 종류의 환경음으로부터 12차 MFCC(mel-frequency cepstral coefficients) 특징 27747개를 추출하고 이를 GMM으로 분류하였다. 각 환경음 클래스의 최적 혼합모델 개수를 추정 하기위해 MDL과 BIC를 적용하고 그 결과를 고정 개수의 혼합모델을 사용한 경우와 비교하였다. 실험 결과에 따르면 혼합모델 선택 방법을 적용한 경우가 그렇지 않은 경우에 비해 거의 유사한 인식성능을 유지하면서 계산복잡도는 BIC와 MDL를 통해 각각 17.8%와 31.7%가 감소하는 것을 확인하였다. 이는 GMM을 이용한 환경음 인식에서 BIC와 MDL 적용을 통해 계산복잡도를 효과적으로 감소시킬 수 있음을 보여준다.

**Abstract** In this paper we applied the optimal mixture number estimation technique in GMM(Gaussian mixture model) using BIC(Bayesian information criterion) and MDL(minimum description length) as a model selection criterion for environmental sounds recognition. In the experiment, we extracted 12 MFCC(mel-frequency cepstral coefficients) features from 9 kinds of environmental sounds which amounts to 27747 data and classified them with GMM. As mentioned above, BIC and MDL is applied to estimate the optimal number of mixtures in each environmental sounds class. According to the experimental results, while the recognition performances are maintained, the computational complexity decreases by 17.8% with BIC and 31.7% with MDL. It shows that the computational complexity reduction by BIC and MDL is effective for environmental sounds recognition using GMM.

**Key Words :** Gaussian mixture model, BIC, MDL, Bayesian information criterion, environmental sounds recognition

## 1. 서론

사회에 대한 패러다임이 최근 스마트 사회로 변화함에 따라 인간 중심의 컴퓨팅 서비스에 대한 요구가 확대되고 있다[1]. 인간 중심의 컴퓨팅 서비스는 사용자가 특별한 지식 없이도 언제 어디서나 목적에 맞는 서비스를 이용할 수 있는 스마트 환경의 핵심 기술이다. 상황인식(context aware) 서비스는 이러한 환경에서 사용자가 처한 상황을 인식하여 능동적이고 유용한 서비스를 제공하는 인간중심의 컴퓨팅 서비스이다. 상황인식 기술은 사용자 중심의 설계, 개인화 서비스, 위치정보, 지식검색 등을 포함한 다양한 분야에서 현재 활용되고 있으며[2], 데스크톱과 웹을 이용한 유선 서비스에서 스마트기기와 무선 인터넷을 중심으로 한 모바일 서비스로 발전하고 있다.

본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임. (No. 2011-0009213)
*교신저자 : 백성준(tozero@jnu.ac.kr)
접수일 11년 12월 05일 수정일 12년 01월 25일 게재확정일 12년 02월 10일

이와 같은 발전 방향에 따라 모바일을 통한 상황정보 기반 서비스의 중요성이 부각되고 있으며 현재 다양한 분야에서 많은 연구가 행해지고 있다. 예를 들어, 사용자가 극장 안에서 영화를 감상 중이거나, 강의실에서 수업 중인 경우, 사용자의 위치정보를 바탕으로 공공장소에서 자동으로 에티켓모드로 전환하는 서비스를 제공한다거나 시각장애인에게 주변 환경을 소리지도의 형태로 전환하여 환경정보를 제공하는 등 인간중심 서비스를 제공할 수 있다. 또한 환경음 인식 기술을 이용한 환경정보와 위치정보시스템(GPS)의 정보를 결합하여 사용자가 원하는 정보를 보다 고급한 형태로 더욱 정확하게 제공할 수도 있을 것이다.

이러한 새로운 서비스의 기본 기술로서 환경음 인식 기술은 기존 음성인식 기술을 토대로 그 성능을 개선하는 방향으로 발전하고 있다. 환경음은 무작위성과 높은 분산성 등의 특성을 가지고 있으므로 전처리, 특징추출, 분류방법을 적용할 때 이를 충분히 고려해야 한다. 기존의 연구에는 환경음 인식에 MFCC 특징을 사용하고 고정된 개수의 혼합모델(mixture)을 가진 GMM을 적용한 여러 결과가 있다[3-5]. 고정된 개수의 혼합모델로 데이터를 모델링 할 때 혼합모델의 개수가 증가할수록 인식성능이 향상되지만 그에 따라 계산비용이 크게 증가하는 단점이 있다. 각 클래스의 데이터 특성에 따라 서로 다른 최적 혼합모델 개수를 할당하여 각 클래스를 모델링함으로써 계산비용이 증가하는 단점을 개선할 수 있다. 이에 본 연구에서는 GMM분류에 혼합모델의 개수를 적절하게 결정하는 방법을 적용함으로써 계산복잡도(complexity)를 크게 감소시킬 수 있는 알고리듬을 제안하고자 한다. 혼합모델의 개수는 MDL과 BIC 모델선택 기준을 적용하여 최적의 혼합모델 개수를 추정한 경우와 모델선택 방법을 적용하지 않은 경우에 각각 인식성능과 계산복잡도를 비교하고 이를 분석하였다.

## 2. 실험방법

### 2.1 GMM

데이터로부터 추출된 특징들의 분류를 위해 GMM을 사용하였다. GMM은 주어진 데이터 집합의 분포를 여러 개의 가우시안 확률밀도함수로 모델링하고, 우도(likelihood)를 최대로 하는 클래스를 선택하는 방법이다[6]. 추출한 특징이 $D$차라고 할 때 특징벡터 $\boldsymbol{X}_k = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \cdots, \mathbf{x}_N\}$ 라고 하면 $M$개의 혼합모델을 가지는 가우시안 확률밀도 함수의 우도는 다음과 같다.

$$p(\mathbf{x}_t|\boldsymbol{\theta}) = \sum_{m=1}^{M} p_m N(\mathbf{x}_t|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \quad (1)$$

$$\boldsymbol{\theta} = \{w_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}_{m=1}^{M} \quad (2)$$

$$\sum_{m=1}^{M} w_m = 1, \; 0 \le w_m \le 1 \quad (3)$$

이때 $k$는 각 모델에 해당하는 클래스의 수를 나타낸다. $\boldsymbol{\theta}$는 GMM 파라미터로써, 한 혼합모델의 가중치(mixture weight) $w_m$, 평균 벡터(mean vector) $\boldsymbol{\mu}_m$, 공분산 행렬(covariance matrix) $\boldsymbol{\Sigma}_m$으로 구성된다. 본 실험에서는 최대 우도를 갖는 파라미터 $\boldsymbol{\theta}$를 추정하기위해 EM(expectation- maximization)알고리즘을 이용하였다.

### 2.2 혼합모델 선택 방법

혼합모델 선택은 주어진 데이터에 최적의 복잡도를 가지는 모델을 만드는 파라미터를 선택하여 혼합모델을 고르는 작업이다[7][8]. 즉, 혼합모델 선택이란 여러 후보 혼합모델에서 입력 데이터에 가장 적합한 혼합모델을 선택하는 것이다. $N$개의 샘플을 갖는 입력 데이터 집합 $X_N$에 대하여 이 데이터를 설명할 수 있는 $M_1, \cdots M_q$의 $q$개의 여러 혼합모델이 있을 때, 혼합모델 선택은 데이터 집합 $X_N$을 가장 잘 나타내는 $M_{opt}$을 선택해야 한다. 이때 혼합모델은 혼합성분의 개수를 바꿈으로써 생성할 수 있는데, 본 연구에서는 이를 1개부터 20개까지 변화시켰다.

BIC는 다량의 데이터가 있을 때 우도함수나 사전확률이 다변량 가우시안 분포로 근사된다는 점에서 유도되며 정의는 다음과 같다[9].

$$\log P(X|M) \approx -2\log(X|M, \hat{\theta}) + d\log N \quad (4)$$

식 (4)에서 $d$는 혼합모델에서 파라미터의 수, $N$은 데이터의 수이다. 첫 번째 항의 데이터의 우도 $\log P(X|M, \hat{\theta})$는 데이터를 가장 잘 설명할 수 있는 확률 모델을 찾도록 유도하는 성분이다. 두 번째 항인 $d\log N$은 혼합모델 내의 파라미터 개수에 대한 패널티 항으로 볼 수 있다. 따라서 BIC는 이러한 두 항에 상호 배타적인 특성이 서로 조화되는 타협점에서 최선의 모델이 구축되는 방안을 제시한다. 즉 최종적으로 데이터의 우도와 모델 파라미터의 개수가 조화를 이루는 BIC 최소값에서 혼합모델을 선택한다.

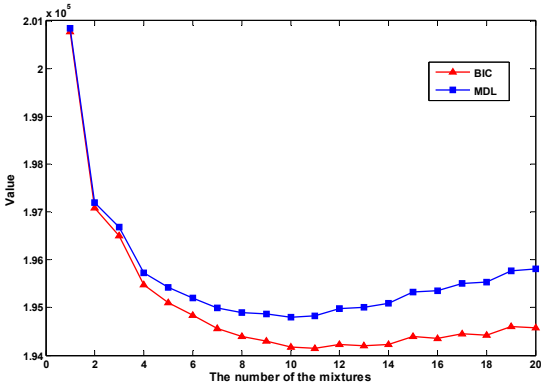
MDL는 우도함수와 데이터 수뿐만 아니라 데이터의 차원을 고려해서 혼합모델을 결정하는 기법이며 다음 두

가지 항목의 합을 최소화하는 모델을 선택한다. 첫째는 모델을 표현하는데 필요한 정보의 양과 둘째는 주어진 모델을 이용하여 입력 데이터를 표현하는데 필요한 정보의 양이다.

$$MDL = -2\log(X|M,\hat{\theta}) + d\log(ND) \quad (5)$$

이는 때때로 BIC와 같다고 가정되지만[7], MDL은 BIC와 다르게 패널티 항에 차원을 추가함으로써 패널티 항이 반영되는 비중이 커진다[10]. 이는 그림 1에서도 보듯이, BIC와 MDL의 패널티 항의 차이에 따라 결과가 달라지는 데서 확인할 수 있다.

먼저 혼합모델 선택방법을 적용하기 전에 기준 성능을 구하기 위해 GMM의 혼합모델 개수를 1부터 20까지 증가시키며 각 모델의 분류율을 구하였다. 그리고 1부터 20까지 각 혼합모델에 대한 BIC와 MDL값을 계산하여 최소값을 갖는 혼합모델의 개수를 각 클래스의 최적 혼합모델 개수로 추정하였다.




[그림 1] 버스내부 클래스에서 각 혼합모델에 따른 BIC와 MDL
[Fig. 1] BIC and MDL according to the mixtures in inside bus

이 실험의 한 예로 버스내부 클래스에서 1부터 20까지 각 혼합모델에 따른 BIC와 MDL값 변화를 그림 1에 보였다. 그림에서 BIC와 MDL의 최소값이 각각 혼합모델 11개와 10개에서 확인되었다. 따라서 이 경우 최적의 혼합모델 개수는 각각 11개와 10개로 결정된다.

## 3. 실험 결과

### 3.1 실험 데이터

환경음 실험 데이터는 모두 9가지 환경에서 휴대용 microphone을 이용하여 획득하였다. 각 환경은 서울 지하철 내부(Subway), 자동차 내부(Veghicle), 고속철도 내부(KTX), 시내버스 내부(Urban Bus), 실외 걷기(Outside Walking), 실외 뛰기(Outside Running), 상영 중인 극장 내부(Theater), 식당 내부(Restaurant), 수업중인 강의실 내부(Classroom)로 구성하였다. 환경음 데이터 각각의 길이는 대략 60분 전후이고, 샘플링 주파수는 8㎑, 양자화비트는 16bits이다. 구축한 환경음 데이터베이스의 정보는 표 1에 나타내었다.

[표 1] 환경음 데이터베이스 구축 정보
[Table 1] Information of Environmental Sounds database

| 환경음 클래스 | 측정시간 | 측정횟수 | 비고 |
|---|---|---|---|
| 지하철 내부 | 20min | 7회 | 서울 7개호선 |
| 자동차 내부 | 20min | 4회 | 중형/소형 |
| KTX 내부 | 40min | 2회 | 서울-광주 |
| 버스 내부 | 25min | 4회 | 시내버스 |
| 실외 걷기 | 5min | 12회 | 남녀구분 |
| 실외 뛰기 | 5min | 12회 | 남녀구분 |
| 극장 내부 | 20min | 3회 | 시내 극장 |
| 식당 내부 | 20min | 3회 | 뷔페/일반식당 |
| 강의실 내부 | 20min | 3회 | 대학 강의실 |

데이터는 프레임의 길이가 $N$이고 다음프레임은 $N/2$만큼 이동하여 이전 프레임과 중첩하는 방식으로 분할하였다. 본 실험에서 사용한 프레임의 길이 $N$은 8000이다. 분할 시에는 원하지 않는 정보를 최소화하여 의미 있는 주파수 성분을 얻기 위해 Hamming 창을 이용하였으며, 분할된 데이터는 고주파 성분을 강조하여 주파수 특성을 평탄하게 하도록 다음과 같은 pre-emphasis를 적용하였다.

$$H(z) = 1 - 0.97z^{-1} \quad (6)$$

전처리 과정을 거친 데이터는 MFCC 분석으로 12차 특징을 추출하였다. 각 환경음 클래스의 데이터를 훈련그룹과 실험그룹으로 나누었는데 각각의 데이터 수는 2569개, 514개이다. 즉 9개의 클래스를 전부 고려하면, 23121개의 훈련데이터와 4526개의 실험데이터를 사용한 것이다.

### 3.2 혼합모델 선택 실험 결과

기준 성능을 구하기 위한 고정 개수 혼합모델의 경우에는 GMM을 이용하여 1부터 20까지 혼합모델의 수를

증가시키면서 10회 반복 실험하였다. 그 결과를 평균한 결과, 혼합모델의 개수가 20일 때 평균 인식률 87.59%로 가장 좋은 성능을 보였다. BIC와 MDL 모델선택 기준을 사용하여 각 환경음 클래스의 최적의 혼합모델 개수를 추정한 결과와 그에 따른 인식률을 표 2에 나타내었다.

**[표 2]** BIC와 MDL를 이용한 최적의 혼합모델 개수 추정 결과와 인식률

**[Table 2]** The number of optimal mixture using BIC and MDL and recognition accuracy

| 환경음 클래스 | 기존 | BIC 혼합모델 수 | MDL 혼합모델 수 |
|---|---|---|---|
| 지하철 내부 | 20 | 20 | 20 |
| 자동차 내부 | 20 | 19 | 15 |
| KTX 내부 | 20 | 19 | 14 |
| 버스 내부 | 20 | 11 | 10 |
| 실외 걷기 | 20 | 14 | 10 |
| 실외 뛰기 | 20 | 17 | 17 |
| 극장 내부 | 20 | 13 | 9 |
| 식당 내부 | 20 | 19 | 17 |
| 강의실 내부 | 20 | 16 | 11 |
| 총 혼합모델 수 | 180 | 148 | 123 |
| 인식률(%) | 87.59 | 87.51 | 86.73 |

표 2에서 알 수 있듯이 혼합모델 선택 방법을 적용하기 전 20개의 혼합모델을 사용한 경우보다 BIC와 MDL를 통해 추정된 최적의 혼합모델 개수가 감소하였음을 알 수 있다. 특히 MDL을 사용하였을 때 각 클래스별 혼합모델 개수가 크게 감소됨을 확인할 수 있다.

계산복잡도를 간단히 살피기 위해 총 혼합모델 수로 비교해보면, BIC의 경우는 혼합모델 선택 방법을 적용하기 전보다 계산량이 17.8% 감소하였고, MDL의 경우 31.7% 감소하였다. 인식률을 살펴보면 기존에 20개의 혼합모델을 사용한 경우와 BIC와 MDL를 적용하여 최적의 혼합모델을 사용한 경우 그 결과가 비슷하여 인식성능 차이가 거의 없음을 알 수 있다. 다만 상대적으로 더 적은 개수의 혼합 모델을 사용한 MDL의 경우가 BIC에 비해 조금 더 성능의 감소가 있음을 확인할 수 있으나, 이 또한 1% 미만의 성능 저하로 31.7%의 계산량 감소와 비교하면 그 성능 저하는 미미하다고 할 수 있다. 따라서 환경음 인식에, GMM에서 혼합모델 선택방법인 BIC와 MDL를 적용함으로써 인식성능은 유지하면서 계산복잡도를 줄일 수 있음을 확인할 수 있다.

## 4. 결론

본 논문에서는 GMM에 모델 선택기준인 BIC와 MDL를 이용하여 최적의 혼합모델 개수를 결정하고 이를 환경음 인식에서 적용하는 방법에 대해 연구하였다. 일반적으로 GMM은 고정된 개수의 혼합모델로 모든 클래스(Class)의 데이터를 모델링하는데, 혼합모델의 개수가 증가할수록 인식성능이 향상되지만 그에 따라 계산비용이 크게 증가한다. 따라서 모델링하고자 하는 각 클래스의 데이터가 그 성질에 따라 서로 다른 최적 혼합모델 개수를 지닌다는 점을 이용하면, 각 클래스에 서로 다른 혼합모델 개수를 적용함으로써 성능을 유지하면서 계산량을 감소시킬 수 있다. 이에 본 연구에서는 각 환경음 클래스의 최적 혼합모델 개수를 추정하기위해 MDL과 BIC를 적용하고 그 결과를 고정 개수의 혼합모델을 사용한 경우와 비교하였다. 실험 결과에 따르면 혼합모델 선택 방법을 적용한 경우가 그렇지 않은 경우에 비해 인식성능은 거의 떨어뜨리지 않으면서 계산복잡도를 BIC와 MDL의 경우 각각 17.8%와 31.7%가 감소시킬 수 있었다. 이는 GMM을 이용한 환경음 인식에 BIC와 MDL을 적용함으로써 계산복잡도를 효과적으로 감소시킬 수 있음을 보여준다. 추후에는 이 결과를 토대로 GMM에 적용될 수 있는 다른 훈련방법에도 최적의 혼합모델 개수를 적용하여 계산복잡도 감소와 환경음 인식성능 향상을 위한 연구를 진행할 예정이다.

## References

[1] National Information Society Agency Information Strategy Planning Division, “Paradigm shift in the era of smart vision and ICT strategy”, National Information Society Agency, 2010.

[2] Il-Young Hong, “Context-aware software, Now mind you should read beyond gesture,” Korea IT Industry Promotion Agency, 2008.

[3] Jun-Qyu Park, Seong-Joon Baek, “Improvement of Environmental Sounds Recognition by Post Processing”, the Korea Contents Society vol. 10, pp.31-39, 2010.

[4] S. Chu, S. Narayana, C.-C. J. Kuo, and M. J. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in Proc. ICME, 2006.

[5] S. Chu, S. Narayanan, and C.-C. Jay Kuo “Environmental Sound Recognition With Time-Frequency Audio Features,” IEEE Trans. on Audio, Speech, and Language Processing, Vol.17, No.6, pp.1-16, 2009.

[6] Richard O.Duda, Peter E.Hart, David G.Stork, Pattern Classification, John Wiley & Sons, 2001
[7] Burnham, Kenneth P, and David R. Anderson, Model selection and Multimodal Inference : A Practical Information-Theoretic Approach Seconded. New York : Springer-Verlag, 2002
[8] G. McLachlan., D. Peel., “Finite Mixture Models,” A wiley-interscience publication, 2000.
[9] S. S. Chen and P. S. Gopalkrishana, "Speaker, enviroment, and channel change detection and clustering via the Bayesian information criterion," Proceedings of the IEEE Interational Conference on vol.2, pp.645-648, 1998.
[10] J.Rissanen., “modeling by shortest data description,” Automatica, vol.14, pp.465-471, 1978.

**한 다 정**(Da-Jeong Han) **[준회원]**

- 2010년 2월 : 전남대학교 전자컴퓨터공학부 (공학사)
- 2010년 3월 ~ 현재 : 전남대학교 전자컴퓨터공학과 석사과정

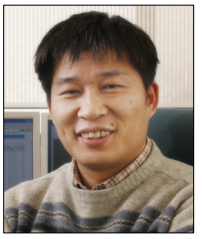
<관심분야>
디지털 신호처리, 패턴인식, 환경음 인식

**박 아 론**(Aaron Park) **[정회원]**

- 2006년 2월 : 전남대학교 전자컴퓨터정보통신공학부(공학사)
- 2008년 2월 : 전남대학교 전자공학과(공학석사)
- 2009년 8월 : 전남대학교 전자컴퓨터공학과(박사수료)

<관심분야>
디지털 신호처리, 패턴인식

**백 성 준**(Sung-June Baek) **[정회원]**



- 1989년 2월 : 서울대학교 전자공학과(공학사)
- 1992년 2월 : 서울대학교 전자공과 (공학석사)
- 1999년 2월 : 서울대학교 전자공학과(공학박사)
- 2002년 3월 ~ 현재 : 전남대학교 전자공학과 교수

<관심분야>
의료, 통신, 음성 관련 디지털 신호처리