

논문 2011-48SC-5-5

SVM에 기반한 음악 장르 분류를 위한 특징벡터 정규화 방법

(Feature-Vector Normalization for SVM-based Music Genre Classification)

임 신 철*, 장 세 진**, 이 석 필**, 김 무 영***

(Shin-Cheol Lim, Sei-Jin Jang, Seok-Pil Lee, and Moo Young Kim)

요 약

본 논문에서는 Mel-Frequency Cepstral Coefficient (MFCC), Decorrelated Filter Bank (DFB), Octave-based Spectral Contrast (OSC), Zero-Crossing Rate (ZCR), 그리고 Spectral Contrast/Roll-Off를 복합 특징벡터로 결합하여 Support Vector Machine (SVM)을 이용한 음악 장르 분류 시스템을 설계하였다. 기존 방식에서는 전체 학습 데이터에 대한 특징벡터를 정규화를 한 후 SVM 모델을 생성하여 분류를 시행하였다. 본 논문에서는 비교 대상이 되는 한 쌍의 클래스에 대해서 One-Against-One (OAO) SVM으로 모델을 생성할 때 선택된 두 클래스의 특징벡터에 대해서만 정규화를 시행하는 방식을 제안한다. 기존 정규화 방식을 이용하면 단일 특징벡터로 OSC를 사용할 경우에는 60.8%, 복합 특징벡터를 모두 이용하는 경우에는 77.4%의 인식율을 얻을 수 있었다. 또한, 제안된 정규화 방식을 이용하면 OSC와 복합 특징벡터에 대해서 각각 8.2%와 3.3%의 추가적인 성능 향상을 얻을 수 있었다.

Abstract

In this paper, Mel-Frequency Cepstral Coefficient (MFCC), Decorrelated Filter Bank (DFB), Octave-based Spectral Contrast (OSC), Zero-Crossing Rate (ZCR), and Spectral Contrast/Roll-Off are combined as a set of multiple feature-vectors for the music genre classification system based on the Support Vector Machine (SVM) classifier. In the conventional system, feature vectors for the entire genre classes are normalized for the SVM model training and classification. However, in this paper, selected feature vectors that are compared based on the One-Against-One (OAO) SVM classifier are only used for normalization. Using OSC as a single feature-vector and the multiple feature-vectors, we obtain the genre classification rates of 60.8% and 77.4%, respectively, with the conventional normalization method. Using the proposed normalization method, we obtain the increased classification rates by 8.2% and 3.3% for OSC and the multiple feature-vectors, respectively.

Keywords : Music genre classification, MFCC, OSC, DFB, SVM

I. 서 론

현존하는 디지털 음원들은 대부분 수작업으로 직접

음악 장르 등에 대한 태그를 작성하여 관리하고 있다. 하지만, 음원 데이터가 방대해짐에 따라 모든 음원들을 수작업으로 분류하기에는 너무 많은 시간이 필요하다. 따라서, 작곡가, 가수, 장르 등 다양한 태그들에 대한 자동 분류 알고리즘이 필요하다. 본 논문은 그 중 장르에 대해서 자동 분류하는 알고리즘을 제안한다.

각 음원별 장르를 정확히 정의 내리기는 쉽지 않다. 그로 인해 장르를 정확히 분류하기에도 어려움이 따른다. 장르를 정의 내리기 어려운 이유는 나라와 사람마다 장르의 경계가 일맥상통하지 않을 뿐만 아니라 문

* 학생회원, *** 정회원, 세종대학교 정보통신공학과 (Department of Information and Communication Engineering, Sejong University)

** 정회원, 전자부품연구원 디지털미디어연구센터 (Korea Electronics Technology Institute)

※ 이 논문은 2011년도 정부(지식경제부)의 재원으로 산업원천기반구축사업의 지원을 받아 수행된 연구임(No. 10037244).

접수일자: 2011년4월22일, 수정완료일: 2011년8월11일

화, 가수, 시장에 따라서도 그 정의가 달라지기 때문이다^[1].

전통적인 자동 장르 분류 시스템은 Mel-Frequency Cepstral Coefficient (MFCC), Octave-based Spectral Contrast (OSC) 등 다양한 특징들을 이용하여 설계된다^[2~3]. 분류기로는 Gaussian Mixture Model (GMM)과 Support Vector Machine (SVM) 등으로 분류된다^[2~4]. 일반적으로 GMM보다 SVM이 높은 성능을 나타낸다. SVM은 각 특징벡터들의 dynamic range가 큰 경우, 커널함수가 각 클래스를 나누기 위해 많은 계산량을 요구한다. 따라서, 각 특징벡터는 정규화를 통해 dynamic range를 줄인 후, SVM으로 모델링한다. 이 방법은 속도를 개선하지만, 성능에는 큰 변화가 없다. 따라서, 본 논문에서는 성능 개선을 위한 새로운 특징벡터 정규화 방식을 제안하였다. 특징벡터로는 화자 인식과 음성 인식 등에서 MFCC^[5]와 함께 사용되는 Decorrelated Filter Bank (DFB)^[6]를 이용하였고, MFCC/DFB/OSC 등 다양한 특징벡터들을 결합하여 멀티 특징벡터를 구성하였다. SVM은 One-Against-One (OAO) 방식을 이용하였다.

본 논문의 구성은 다음과 같다. II장에서 제안한 시스템에 대하여 설명하고, III장에서는 제안한 시스템의 실험 결과를 보여준다. 마지막으로 IV장에서는 최종 결론을 서술하겠다.

II. 본 론

음악의 장르를 분류하기 위해선 그림 1과 같이 특징 추출 (feature extraction), 모델링 (modeling), 분류

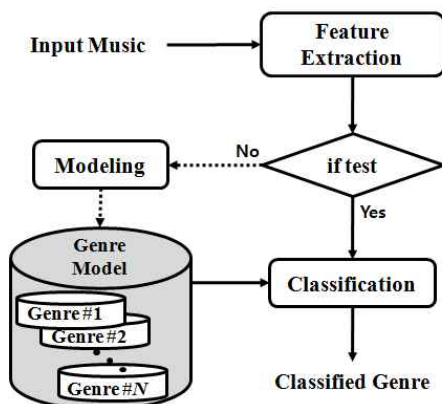


그림 1. 음악 장르 분류 시스템 블록도
Fig. 1. Blockdiagram of music genre classification system.

(classification) 과정이 필요하다.

1. 특징벡터 추출

가. MFCC와 DFB

DFB는 주로 화자인식에 사용되며, MFCC와 마찬가지로 mel-scale band-pass filter를 사용한다^[6]. DFB는 MFCC 추출 시 마지막 과정인 DCT를 대신하여, high-pass filter를 이용하는 점이 다르다. DFB는 먼저 B 개 밴드의 mel-scale band-pass filter를 이용하여 밴드별 스펙트럼의 가중치 합을 구한다. 밴드별 가중치 합에 log를 적용한 신호를 $S(k)$ 라 하면, K 차 DFB는 다음과 같이 구할 수 있다.

$$DFB(k) = S(k+1) - S(k), \quad 1 \leq k \leq K < B \quad (1)$$

나. OSC

OSC는 octave-scale band-pass filter를 통하여, 각 밴드내의 spectral contrast를 구한다. OSC는 각 밴드별 스펙트럼의 peak와 valley를 이용한다. 강한 peak는 음악의 harmonic 부분에, 강한 valley는 non-harmonic 부분과 연관된다. 따라서, OSC는 음악의 harmonic과 non-harmonic을 동시에 고려하는 특징벡터이다^[3].

본 논문에서는 22050Hz의 샘플링 주파수를 기준으로 octave-scale band-pass filter의 8개 밴드를 [0 - 100), [100 - 200), [200 - 400), [400 - 800), [800 - 1600), [1600 - 3200), [3200 - 6400), [6400 - 11025)로 설정하였다. 앞서 설계한 밴드를 이용하여 각 밴드의 구간을 나눈 후, i 번째 밴드에 속한 스펙트럼 $\{x_{i,1}, x_{i,2}, \dots, x_{i,L_i}\}$ 을 내림차순으로 정리한다. 여기서 L_i 은 i 번째 밴드에 속한 스펙트럼 bin의 갯수이다. 내림차순 정리된 스펙트럼을 $\{x'_{i,1}, x'_{i,2}, \dots, x'_{i,L_i}\}$ 라고 할 때, peak와 valley를 다음과 같은 식으로 구한다.

$$Peak_i = \log\left(\frac{1}{\alpha L_i} \sum_{j=1}^{\alpha L_i} x'_{i,j}\right) \quad (2)$$

$$Valley_i = \log\left(\frac{1}{\alpha L_i} \sum_{j=1}^{\alpha L_i} x'_{i,L_i-j+1}\right) \quad (3)$$

α 는 상수로서 0.02로 설정하여 사용하였다. Spectral contrast는 각 밴드의 peak와 valley의 차로 구할 수 있다. i 번째 밴드의 spectral contrast는 다음과 같이 구한다.

$$SC_i = Peak_i - Valley_i \quad (4)$$

본 논문에서는 i 번째 밴드에서 $\{SC_i, Valley_i\}$ 를 특징으로 추출하며, 따라서 총 16차의 OSC 특징벡터를 추출하였다.

다. Additional Feature Vectors (AFV)

가와 나에서 설명한 MFCC, DFB, OSC 이외에도 아래와 같이 다양한 특징벡터를 추출하여 장르 인식에 사용하였다.

(1) Zero-Crossing Rate (ZCR)

ZCR은 Voice Activity Detection 등에 많이 사용되어 왔다. ZCR은 신호의 부호가 바뀌는 양을 나타내는 특징으로 다음과 같이 구한다.

$$ZCR = \frac{1}{N} \sum_{n=2}^N Sign(n), \quad (5)$$

$$Sign(n) = \begin{cases} 1, & x(n) * x(n-1) < 0 \\ 0, & otherwise \end{cases}$$

$x(n)$ 과 N 은 각각 입력신호와 총 샘플수를 나타낸다.

(2) Spectral Centroid (SC)

SC는 spectral magnitude의 centroid를 나타내는 특징으로 다음과 같이 구한다.

$$C = \frac{\sum_{k=1}^K M[k] \times k}{\sum_{k=1}^K M[k]} \quad (6)$$

$M[k]$ 와 K 은 각각 magnitude와 magnitude의 총 샘플수를 나타낸다.

(3) Spectral Roll-Off (SRO)

SRO는 magnitude가 α 만큼 집중되어있는 frequency f 를 찾는다. α 는 0.85로 설정하였다.

$$\sum_{k=1}^f M[k] = \alpha \times \sum_{k=1}^K M[k] \quad (7)$$

라. Texture window

가-다에서는 analysis window를 이용하여 MFCC,

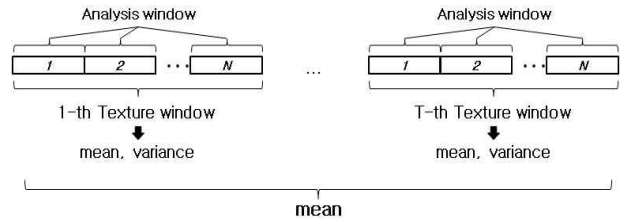


그림 2. Texture window를 사용한 전체 음악에 대한 특징벡터 추출방법

Fig. 2. Feature extraction from a whole music piece using texture windows.

DFB, OSC, AFV 등의 특징벡터를 추출하였다. 음악은 short-time 특징벡터의 시리즈로 표현될 수 있으므로, 음악의 장르를 분류하기 위해서 그림 2와 같이 N 개의 analysis window씩을 묶어서 T 개의 texture window로 정의한다. 그 후, texture window별로 특징벡터의 평균과 분산을 구한다. Analysis window마다 P 차 특징벡터를 추출한다면, texture window는 $2 * P$ 차 특징벡터를 추출하게 된다^[2]. 본 논문에서는 특징벡터 차수를 줄이기 위해서, texture window에서 구한 $2 * P$ 차 특징벡터들의 평균값을 최종적인 특징벡터로 이용하였다.

2. 모델링 및 분류

음악 장르 분류를 위해서 추출된 특징벡터를 SVM을 이용하여 모델링하고 분류를 수행하였다^[7]. SVM은 마진을 이용하여 각 클래스를 분류한다. 여기서 마진은 분류할 데이터와 커널 함수가 만날 때까지 확장한 폭을 이야기한다. 이 마진을 최대화하는 최적 분류 초평면은 다음과 같이 구한다.

$$f(x) = \sum_{i=1}^n \alpha_i K(x_i, x) + b \quad (8)$$

여기서 n , α_i , $K(x_i, x)$ 와 b 는 각각 특징벡터의 차수, 라그랑제 상수, 커널 함수, 그리고 바이어스를 나타낸다. 본 논문에서는 선형, 다항식, 방사 기저 함수 등 다양한 커널 함수들 중에서 선형 커널을 사용하였다. 일반적인 SVM은 두 개 클래스에 대해서만 사용이 가능하다. 따라서, 멀티 클래스 방법을 이용하여 분류할 수 있는데, OAO 방법과 One-Against-All (OAA) 방법이 있다. 본 논문에서는 OAO 방법으로 장르를 분류하였다. 따라서, $n(n-1)/2$ 의 모델이 필요하다.

정확한 분류를 위해 OAO SVM을 이용하여 모델을 생성하기 전에 특징벡터를 정규화 했다. 본 논문에서는

다음과 같이 min-max 정규화를 사용하였다.

$$\overline{x_{i,j|\tau}} = \frac{x_{i,j|\tau} - \min(\{x_{i,\lambda|\tau}\})}{\max(\{x_{i,\lambda|\tau}\}) - \min(\{x_{i,\lambda|\tau}\})} \quad (9)$$

where $\lambda = 1, 2, \dots, D(\tau)$ and $\tau = 1, 2, \dots, T$

여기서 $x_{i,j|\tau}$ 는 τ 번째 장르 클래스에 속한 j 번째 학습 음원 데이터에서 추출한 특징벡터의 i 번째 component이다. 그리고, $D(\tau)$ 와 T 는 각각 장르 클래스별 학습 음원 데이터의 개수와 비교할 장르 클래스의 개수이다.

기존 방식에서는 C 개의 전체 장르 클래스에 대하여 특징벡터를 정규화를 한 후, OAO SVM으로 모델링하고 분류를 한다. 하지만, 본 논문에서는 한 쌍의 클래스에 대해 OAO SVM을 모델링할 때 마다 정규화를 하고 분류를 하였다. 즉, 기존 방식은 $C(C-1)/2$ 의 OAO SVM 모델을 만들기 전에 식 (9)에 의해서 $T=C$ 로 두어 전체 클래스에 대하여 정규화를 하는 반면에, 본 논문에서는 $T=2$ 로 두어 비교 대상이 되는 각 쌍의 클래스에 대해서만 정규화를 하게 된다. OAO SVM 사용시 비교 대상이 되는 한 쌍의 장르 클래스에 대해서만 정규화를 수행함으로써, 특징 벡터의 dynamic range가 해당 장르 클래스에 적합하게 설정되어 다른 장르 클래스의 영향을 배제하고 좀 더 정확한 구분이 이루어질 수 있다.

III. 실험

제안 시스템의 성능을 알아보기 위해 공개된 음악 데이터베이스인 GTZAN을 사용하였다. GTZAN은 총 10개의 장르로 blues, classical, country, disco, hiphop, jazz, pop, metal, reggae, rock으로 구성 되어있다. 한 곡당 30초로 각 장르당 100곡이며, 16bit, 22050Hz, 모노, AU 파일 포맷형식이다^[8].

특징벡터는 90ms의 analysis window 별로 overlap

표 1. 정규화 방식에 따른 단일 특징벡터 별 성능
Table 1. Genre-classification accuracy of single feature vectors according to the normalization methods.

		MFCC	DFB	OSC
Accuracy (%)	기존 정규화 방식	62.4	61.8	60.8
	제안하는 정규화 방식	63.3	64.2	69.0
Feature Order		26	26	32

없이 MFCC와 DFB는 14개의 밴드로 각각 13차 특징벡터, OSC는 8개 밴드로 16차 특징벡터를 추출하였다. 그 후, 약 3초의 texture window (33개의 analysis window를 포함)에 대해서 1개의 analysis window만큼 쉬프트하면서 특징벡터를 추출하였다. 각 장르 데이터별 90%는 학습 데이터, 10%는 테스트 데이터로 사용하여 10-fold cross-validation을 하여 성능을 평가하였다.

표 1은 단일 특징벡터에 대하여 정규화 방식에 따른 장르분류 결과이다. 기존 방식은 전체 10개의 장르에 대해서 1번의 정규화를 수행하며, MFCC, DFB, OSC 순으로 장르 분류 성능이 좋았다. 하지만, 제안 방식은 비교하는 장르의 pair 별로 정규화를 실시하며 OSC, DFB, MFCC 순으로 성능이 좋았다. 기존 정규화 방식은 MFCC가 62.4%, 제안 정규화 방식은 OSC가 69.0%의 장르 인식율을 보였다. OSC의 경우 기존 정규화 방식을 사용한 경우에 비해서 제안 정규화 방식을 사용하면 8.2%의 성능 개선을 얻을 수 있었다.

표 2는 MFCC, DFB, OSC, AFV를 복합 특징벡터 형태로 이용하여 장르 분류를 시도한 경우의 인식율을 나타내고 있다. 특징벡터의 차수는 단일 특징벡터보다 높다. 하지만, 전체적으로 단일 특징벡터를 가지고 실험한 경우보다 높은 성능을 보이는 것을 알 수 있었다. 특히, 같은 청각 모델 필터를 가지고 추출한 특징 MFCC와 DFB의 결합보다, octave 기준 청각 모델 필터로 추출한 OSC와 MFCC 또는 DFB의 결합이 보다 높은 성능을 보였다. 또한, DFB가 MFCC보다 다른 특징과 복합 특징으로 결합하였을 때, 향상된 성능을 보였다. 기존 정규화 방식에 비해서 제안 정규화 방식은 최소 2.4%에서 최대 4.0%의 성능 향상을 보였다. 또한, DFB+OSC+AFV를 복합 특징벡터로 결합한 방식은 모든 특징 MFCC+DFB+OSC+AFV를 이용한 경우보다 인식율은 0.9% 낮지만 특징벡터 차수는 90차에서 64차로 26차를 줄일 수 있음을 알 수 있다.

표 3은 DFB+OSC+AFV 특징벡터에 대해서 제안한 정규화 방식을 사용한 경우에 대한 각 장르별 분류 성공률의 confusion matrix를 나타낸다. 1-fold당 각 장르 별로 10개의 파일을 테스트하여, 총 100개의 파일을 분류한 결과이다. 행이 테스트 장르일 때, 각 열은 분류된 장르를 나타낸다. 그 결과 classical 장르에서 97%의 매우 뛰어난 인식 성능을 보였다. Classical 장르는 저주파에 많은 에너지가 분포하며, 보컬부분이 거의 없어 분류 성공률이 뛰어나다. Metal 장르의 경우엔 고주파

표 2. 정규화 방식에 따른 복합 특징벡터 별 성능

Table 2. Genre-classification accuracy of multiple feature vectors according to the normalization methods.

		MFCC+DFB	MFCC+OSC	DFB+OSC	MFCC+OSC+ AFV	DFB+OSC+ AFV	MFCC+DFB+ OSC	MFCC+DFB+ OSC+AFV
Accuracy (%)	기존 정규화 방식	66.4	75.0	75.7	75.2	76.0	77.6	77.4
	제안하는 정규화 방식	70.4	77.5	78.1	78.2	79.8	80.0	80.7
Feature Order		52	58	58	64	64	84	90

표 3. DFB+OSC+AFV 특징벡터와 제안된 정규화 방식을 이용한 장르별 인식율에 대한 confusion matrix

Table 3. Confusion matrix of the genre-classification accuracy using the DFB+OSC+AFV feature vectors and the proposed normalization method.

	bl	cl	co	di	hi	ja	me	po	re	ro
bl	86	0	4	0	0	1	2	0	4	3
cl	0	97	1	0	0	1	0	0	0	1
co	6	0	79	1	0	3	0	3	0	8
di	4	0	6	69	5	0	1	4	3	8
hi	2	1	0	6	76	0	2	3	9	1
ja	3	5	2	1	0	87	0	0	0	2
me	2	0	0	2	0	0	91	0	0	5
po	1	1	4	3	1	0	0	84	3	3
re	7	0	5	6	8	1	0	1	69	3
ro	6	1	11	8	2	1	5	1	5	60

에 다른 장르에 비해 많은 에너지가 있어서 91%로 분류가 잘 되었다. Classical 장르와 metal 장르를 제외한 나머지 장르는 주파수 전 대역에 고루 에너지가 분포하기 때문에 분류 성공률이 classical 장르와 metal 장르보다 저조하였다. 그 중 rock 장르는 가장 저조한 분류 성능을 보였고, 주로 blues, country, 그리고 disco 장르로 오인식 되었다.

IV. 결 론

본 논문에서는 MFCC, DFB, OSC, AFV 등 다양한 특징벡터를 복합 특징벡터로 구성하고, OAO SVM으로 한 쌍의 장르 클래스 모델을 만들 때, 해당 클래스에 대해서만 정규화 하여 음악 장르 분류 시스템의 인식율을 향상시켰다. 제안 정규화 방식이 기존 정규화 방식보다 높은 성능 향상을 보였고, 단일 특징벡터의 경우 OSC에 대해서는 8.2 %, 복합 특징에 대해서는 3.3%의 성능 향상을 보였다.

MFCC, DFB, OSC, AFV 등 다양한 특징벡터를 복

합 특징벡터를 구성 시, 비교적 높은 차수를 가져도 성능이 크게 향상을 하지 않는 경우도 있었다. 이는 복합 특징을 구성하는 데 있어서 특징들의 조합도 중요하다는 것을 알 수 있다. 향후에는 특징벡터들의 조합에 대한 연구와 장르 분류에 보다 적합한 특징벡터 및 분류기를 연구할 예정이다.

참 고 문 헌

- [1] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: A survey," *IEEE Signal Process.*, vol. 23, no. 2, pp. 133-141, 2006.
- [2] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293-302, 2002.
- [3] D. N. Jiang, L. Lu, H. J. Zhang, J. H. Tao, and L. H. Cai, "Music type classification by spectral contrast feature," in *Proc. IEEE Int. Conf. Multimedia and Expo*, vol. 1, pp. 113-116, 2002.
- [4] C-H. Lee, J-L. Shih, K-M. Yu, and J-M Su, "Automatic music genre classification using modulation spectral contrast feature," in *Proc. IEEE Int. Conf. Multimedia and Expo*, pp.204-207, 2007.
- [5] 이성훈, 배건성, "Text-dependent한 화자인식 시스템에 관한 연구," *대한전자공학회, 대한전자공학회 학술대회 논문집*, 제7권, 제1호, 312-317쪽, 1994년 1월
- [6] J. Jung, K. Kim, and M. Y. Kim, "Noise robust speaker identification based on the advanced missing feature theory," *Electronics Letters*, vol. 46, no. 14, pp. 1027-1029, 2010.
- [7] 임정수, 송지현, 장준혁, "SVM의 미세조정을 통한 음성/음악분류 성능향상," *대한전자공학회 전자공학 학술대회 논문지-SP*, 제47권 SP편 제2호, 141-148쪽, 2011월 3월
- [8] GTZAN Genre Collection Database,

“http://marsyas.info/download/data_sets”

- [9] 임신철, 장세진, 이석필, 김무영, “Decorrelated Filter Bank를 이용한 음악 장르 분류 시스템,” *한국음향학회지*, 제30권, 제2호, 100-106쪽, 2011년 2월

저 자 소 개



임 신 철(정회원)
2011년 세종대학교 정보통신
공학과 학사 졸업
2011년~현재 세종대학교
정보통신공학과 석사과정
<주관심분야 : 화자인식, 음악정
보검색, 음질 향상>



장 세 진(정회원)
1995년 경북대학교
전자공학과 학사 졸업
1997년 경북대학교
전자공학과 석사 졸업
1997년~2002년 대우전자 전략
기술연구소, 전임연구원
2002년~현재 KETI 차세대음향산업지원센터,
센터장



이 석 필(정회원)
1990년 연세대학교
전기공학과 학사 졸업
1992년 연세대학교
전기공학과 석사 졸업
1997년 연세대학교 전기전자
공학과 박사 졸업
1997년~2002년 대우전자 영상 연구소,
선임연구원
2002년~현재 KETI 디지털미디어연구센터,
센터장



김 무 영(정회원)-교신저자
1993년 연세대학교
전자공학과 학사 졸업
1995년 연세대학교
전자공학과 석사 졸업
1995년~2000년 삼성종합기술원
전문연구원
2001년~2004년 Royal Institute of Technology
(KTH, 스웨덴) Dept. Signals, Sensors,
Systems, 박사 졸업
2004년~2005년 Royal Institute of Technology
(KTH, 스웨덴) Dept. Signals, Sensors,
Systems, PostDoc
2005년~2006년 Ericsson Research (스웨덴),
Senior Research Engineer
2006년~현재 세종대학교 정보통신공학과 부교수
<주관심분야 : 음악정보검색, 음성/오디오 신호처
리 및 코딩, 패턴인식, 정보이론.>