논문 2008-45SP-6-17

SMV코덱의 음성/음악 분류 성능 향상을 위한 Support Vector Machine의 적용

(Analysis and Implementation of Speech/Music Classification for 3GPP2 SMV Codec Based on Support Vector Machine)

김 상 균*, 장 준 혁**

(Sang-Kyun Kim and Joon-Hyuk Chang)

요 약

본 논문에서는 support vector machine (SVM)을 이용하여 기존의 3GPP2 selectable mode vocoder (SMV)코덱의 음성/음악 분류 성능을 향상시키는 방법을 제시한다. SVM은 통계적 학습 이론으로 훈련 데이터 사이의 최적 분류 초평면을 찾아내 최적화된 이진 분류를 보여준다. SMV의 음성/음악 실시간 분류 알고리즘에서 사용된 특징벡터와 분류방법을 분석하고, 이를 기반으로 분류성능향상을 위해 통계적 학습 이론인 SVM을 도입한다. 구체적으로, SMV의 음성/음악 분류알고리즘에서 사용되어진 특징벡터만을 선택적으로 사용하여 효과적으로 SVM을 구성한 분류기법을 제시한다. SMV의 음성/음악 분류에 적용한 SVM의 성능 평가를 위해 SMV 원래의 분류알고리즘과 비교하였으며, 다양한 음악장르에 대해 시스템의 성능을 평가한 결과 SVM을 이용하였을 때 기존의 SMV의 방법보다 우수한 음성/음악 분류 성능을 보였다.

Abstract

In this paper, we propose a novel approach to improve the performance of speech/music classification for the selectable mode vocoder (SMV) of 3GPP2 using the support vector machine (SVM). The SVM makes it possible to build on an optimal hyperplane that is separated without the error where the distance between the closest vectors and the hyperplane is maximal. We first present an effective analysis of the features and the classification method adopted in the conventional SMV. And then feature vectors which are applied to the SVM are selected from relevant parameters of the SMV for the efficient speech/music classification. The performance of the proposed algorithm is evaluated under various conditions and yields better results compared with the conventional scheme of the SMV.

Keywords: Support Vector Machine(SVM), Selectable Mode Vocoder(SMV), Speech/Music Classification Algorithm

I. 서 론

최근 IT기술의 발달로 이동통신기기 내에서의 다양한 멀티미디어 서비스가 본격적으로 사용화 되기 시작하면서 제한된 주파수 대역에서 효율적인 통신환경을 구축하기 위한 연구가 활발히 진행되고 있다. 제한된 통신망을 효과적으로 사용하기 위하여 입력 음성 신호 특징에 따라서 선택적으로 프레임마다 4단계로 나누어 전송률을 결정해 부호화 하는 방식을 3GPP2 의 표준 코덱인 selectable mode vocoder (SMV)에서 사용하고 있다[1~2]. 입력 음성 신호의 종류에 의해 매 프레임마다 전송률을 적절히 부여하는 것이 이동통신기기에서의 통화음질을 결정짓는 중요한 과제이다. 특히 최근의 이동통신 환경은 음성 전달에만 국한 된 것이 아니라 음악,

* 학생회원, ** 정회원, 인하대학교 전자공학부
(Department of Electronics Engineering, Inha University)

※ 본 연구는 지식경제부 및 정보통신연구진흥원의 IT 핵심기술개발사업 [2008-F-045-01]과 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음 (IITA-2008-C1090-0804-0007).

접수일자: 2008년4월11일, 수정완료일: 2008년10월10일

사진, 영상 등과 같이 다양한 정보를 전송해야 하기 때문에 효과적으로 음성/음악 분류 하는 방법을 찾기 위한 연구가 활발히 진행 되고 있다[3~4].

본 논문에서는 실시간 음성/음악 분류기반으로 가변 전송률 알고리즘을 채택하고 있는 SMV의 기존 방법을 충실히 분석하고 이를 기반으로 음성/음악분류 성능을 향상시키기 위해 기존 SMV 인코딩부분의 전처리과정에서 자동적으로 추출되는 파라미터 중 통계적 학습 분류성능이 우수한 것들을 모아 별도의 계산과정 없이 특징벡터로 사용하였다. 엄선된 특징벡터들을 이용하여 support vector machine (SVM)기반의 음성/음악 분류 알고리즘을 제안하고 이것을 기존의 SMV 방법과 다양한 환경에서 비교하였다.

본 논문의 Ⅱ장과 Ⅲ장에서는 SMV 코덱에 대해서 간략하게 알아보고 SMV 음악 분류 방법과 특징 벡터를 소개한다. Ⅳ장에서는 SVM의 특징과 SVM을 적용한 음성/음악 분류 알고리즘을 제시한다. Ⅴ장에서는 다양한 환경에서 기존의 SMV 실험결과와 제안된 알고리즘의 실험결과를 비교하였으며 Ⅵ장에서 결론을 맺는다.

Ⅱ. SMV (Selectable Mode Vocoder)의 개요

SMV는 프레임 단위로 8.55 kbps, 4.0 kbps, 2.0 kbps, 0.8 kbps 총 4가지 전송률을 가지고 있으며 전송률은 각각 Rate 1 (full-rate), Rate 1/2 (half-rate), Rate 1/4 (quarter-rate), Rate 1/8 (eighth-rate)이다. 또한, SMV는 통신망의 상황에 따라 Mode 0 (premium mode), Mode 1 (standard mode), Mode 2 (economy mode), Mode 3 (capacity-saving mode) 총 4가지 모드에서 동작을 하여 평균 데이터율과 음질의 균형을 적절히 조절 할 수 있다[5].

SMV에 입력된 신호는 묵음 향상, 고 대역 통과 필터, 잡음 억제, 적응 틸트 필터 순으로 구성된 전처리 단계를 거친 후 프레임 수준 처리로 넘겨져서 선형 예측 분석, 개회로 피치 검출, 음성 변형, 프레임 클래스 분류 과정을 거친다. 선형 예측 분석과 개회로 피치 검출 과정으로부터 신호 분류에 필요한 파라미터들을 얻는다.

SMV는 8 kHz로 샘플링 된 입력 신호 160개 (20ms)를 한 프레임으로 묶고 입력 음성의 특성에 의해 묵음 또는 주변 잡음, 잡음 같은 무성음, 무성음, 시작 음, 불규칙적인 유성음, 규칙적인 유성음 이와 같이 6가지의 프레임 클래스로 나뉜다. 외부에서 입력된 모드와 프레임 클래스에 따라서 전송률이 결정되고 Rate 1과 Rate 1/2의 전송률일 경우에만 Type 0과 Type 1로 분류 된다. Type은 프레임 클래스가 일정한 유성음으로 분류되었을 경우 Type 1로 결정되어 Type 0 보다 고정 코드북에 더 많은 비트를 할당한다. 입력 신호가 음악으로 분류되면 전송률은 Rate 1로 주어지는 반면 그 외의 입력 신호들은 고정된 문턱 값에 의해 전송률이 선택 된다[6].

Ⅲ. SMV 음악 분류 방법과 특징벡터

SMV에서의 음악 분류 과정은 먼저 음성 검출기 (voice activity detection, VAD)에서 입력 신호가 음성과 묵음 또는 주변 잡음으로 나뉜 후 음성으로 판별된 경우에만 거치게 되며 음성/음악 분류에 사용되는 파라미터들은 다음과 같다.

1. 이동 평균 에너지 $\overline{E}$

$$\overline{E} = 0.75 \cdot \overline{E} + 0.25 \cdot E \tag{1}$$

E는 프레임 에너지 이다.

2. 잡음/묵음의 이동 평균 반사계수 $\overline{k_N}(i)$

$$\overline{k_N}(i) = 0.75 \cdot \overline{k_N}(i) + 0.25 \cdot k_1(i)$$
$$i = 1,...,10 \tag{2}$$

3. 부분적 잔류 에너지의 이동 평균 $\overline{E_N^{res}}$

$$\overline{E_N^{res}} = 0.9 \cdot \overline{E_N^{res}} + 0.1 \cdot E^{res} \tag{3}$$

$\overline{E_N^{res}}$는 $\overline{k_N}$에 따라서 값이 새로워진다.

4. 정규화 된 피치 상관도의 이동 평균 $\overline{corr_P}$

$$\overline{corr_P} = 0.8 \cdot \overline{corr_P} + 0.2 \cdot \left(\frac{1}{5} \cdot \sum_{i=1}^{i=5} corr_p^B(i) \right) \tag{4}$$

$corr_P^B(i)$는 이전 프레임의 피치 상관도이다.

5. 주기적 계수 $\overline{c_{pr}}$

$$\overline{c_{pr}} = \alpha \cdot \overline{c_{pr}} + (1-\alpha) \cdot c_{pr} \quad (5)$$

α는 c_{pr}에 따라 값을 바꿔주는 정해진 가중치이다.

6. 음악 연속 계수의 이동 평균 $\overline{c_M}$

$$\overline{c_M} = 0.9 \cdot \overline{c_M} + 0.1 \cdot c_M \quad (6)$$

SMV의 VAD에서는 식 (1)~(5)로 부터 나온 결과를 정해진 문턱 값과 비교하여 음성의 유무를 판단하며 Music Detection에서는 $\overline{c_{pr}} \geq 18$ 또는 $\overline{c_M} > 200$ 이면 음악으로 판단한다.

Ⅳ. SVM의 특징 및 제안된 알고리즘

SVM은 Vladimir Vapnik과 그의 AT&T Bell 연구소 팀이 개발한 식별 방법으로 최근 몇 년 동안에 이론적인 발전뿐만 아니라, 실제 구현되어 데이터 마이닝 분야는 문론 얼굴인식, 생체인식, 음성인식 등의 다양한 패턴인식 응용 분야에서 널리 사용되고 있다[7~8]. 또한, SVM은 기존의 학습 방법과 다르게 패턴을 고차원 특징 공간으로 사상시킬 수 있다는 점과 대역적으로 최적의 식별이 가능할 뿐만 아니라 알려지지 않은 확률 분포를 갖는 데이터에 대해 잘못 분류하는 확률을 최소화하는 구조적인 위험 최소화 (Structural Risk Minimization) 방법에 기초 하고 있다. 그리고 SVM은 선형적으로 분류 가능한 데이터에 대한 이진분류에 있어 두 개의 클래스를 분류할 수 있는 무수히 많은 초평면 (Hyperplane) 중 클래스의 가장 가까운 점들과 마진이 최대가 되는 최적 초평면을 구함으로써 높은 일반화 성능을 기대할 수 있다[9].

SVM의 학습능률을 높이기 위해서는 최적의 초평면을 구해야하므로 식 (9)의 제약 조건을 가지고, 식 (8)로 표현되는 마진의 역수가 최소가 되도록 하는 최적화 문제로 설정된다. 식 (9)은 식 (7)과 같이 두 가지 조건을 하나의 조건식으로 만든 것이다.

for $i = (1,...,N)$에 대하여

$$w^T x_i + b \geq 1, \text{ for } y_i = 1,$$

$$w^T x_i + b \leq -1, \text{ for } y_i = -1,$$

$$\rightarrow y_i(w^T x_i + b) - 1 \geq 0, \text{ for } i = (1,...,N) \quad (7)$$

$$\text{Minimize} : J(w) = \frac{\| w \|^2}{2} \quad (8)$$

$$\text{Subject to} : y_i(w^T x_i + b) - 1 \geq 0, \quad \text{for } i = (1,...,N) \quad (9)$$

초평면에 대한 단위 (Normal) 법선벡터 w와 중심에서 초평면까지의 거리 b만 주어진다면, 최적 분류 초평면을 알 수 있으므로 모든 데이터 점이 정확히 어느 클래스에 속하는지 판별할 수 있고 마진의 폭도 계산할 수 있다. 모든 데이터 점에 적합하고 가장 넓은 마진을 이루는 최적의 $\hat{w}$와 $\hat{b}$은 라그랑지안 최적화 (Lagrangian Optimization) 기법을 이용하여 목적식과 제약식을 결합한 후 라그랑제 승수 α_i를 포함하여 다음과 같은 식으로부터 구한다.

$$J(w,b,a) = \frac{1}{2} \| w \|^2 - \sum_{i=1}^{m} \alpha_i (y_i \cdot [w^T x_i + b] - 1),$$

$$(\alpha_i \geq 0, i = 1,2,...,N) \quad (10)$$

마진 최대화에 KKT (Karush-Kuhn-Tucker) 조건을 적용하여 식 (11), (12)에서 각각 최적 가중치 벡터 $\hat{w}$와 최적 바이어스 $\hat{b}$을 구한다.

$$\hat{w} = \sum_{i=1}^{N_S} \alpha_i y_i x_i \quad (11)$$

$$\hat{b} = \frac{1 - y_i \hat{w}^T x_i}{y_i} \rightarrow \hat{b} = 1 - \hat{w}^T x_i, \; y_i = 1 \quad (12)$$

최종적으로 임의 패턴 x가 주어질 때, 식 (11), (12)에서 구해진 $\hat{w}$과 $\hat{b}$을 사용하여 식 (13)의 판별함수에 의해 분류 결과가 계산되어 진다.

$$f(x) = sgn(\sum_{i=1}^{N} \hat{\alpha}_i y_i x_i^T x + \hat{b}) \quad (13)$$

한편, 우리가 접하는 대부분의 패턴들은 명확하게 선형분리가 되지 않는 경우가 대부분이며 음성 신호 또한 그러하기 때문에 비선형 변환함수를 이용하여 보다 고차원의 공간으로 사상 (Mapping)시키면 선형 분리가 가능한 조건이 된다. 사상된 공간에서도 원 공간에서의 거리 관계를 어느 정도는 보존 시킬 필요가 있기 때문에, 사상 함수를 이용하여 커널함수 (Kernel Function)

를 식 (14)과 같이 정의한다[10].

$$K(x,x') = \Phi(x)^T \Phi(x') \tag{14}$$

식별 함수와 최적화 문제에 $\Phi(x)$을 쓰지 않고 $K(\bullet)$로만 나타낼 수 있는데 이러한 계산 회피 방법을 커널트릭 (Kernel Trick)이라 하고 Φ가 존재할 수 있는 커널 함수가 주어진 경우에만 유용하며 표 1에서와 같이 주어지며 비선형 SVM의 최종판별 함수는 식 (15)과 같다.

$$f(x) = \sum \alpha_i y_i K(x_i, x_j) + b \tag{15}$$

본 논문은 SMV에서 자동적으로 추출되는 특징 벡터들 중에서 이진 분별이 가능한 이동 평균 에너지, 잡음/묵음의 이동 평균 반사계수, 부분적 잔류 에너지의 이동 평균, 정규화 된 피치 상관도의 이동 평균, 주기적 계수, 음악 연속 계수의 이동 평균 등 6개의 SMV 주요 특징 벡터만을 사용하여 음성/음악 분류를 하였다. 식 (14)에 사용된 커널 함수는 표 1의 커널 함수 중 radial basis function (RBF)를 사용하여 트레이닝을 하였다.

표 1. 커널 함수의 종류
Table 1. Type of Kernel function.

Kernel function	Type of Classifier
Polynomial	$K(x,x') = (x^T x' + 1)^p$
RBF	$K(x,x') = \exp(\frac{\| x - x' \|^2}{2\sigma^2})$
Sigmoid	$K(x,x') = \tanh(\beta_0 x^T x' + \beta_1)$

V. 실험 결과

본 논문에서는 SVM 기반의 강인한 음성/음악 분류 알고리즘 성능을 평가하기위해 기존 SMV 음성/음악 알고리즘을 receiver operating characteristics (ROC) 곡선과 음성/음악 검출확률 (P_d)로 표현하여 비교하였다.

본 실험을 위해서 사용된 음성 데이터베이스는 8 kHz로 샘플링 된 약 6 sec 정도의 깨끗한 음성으로 326명의 남자와 138명의 여자 화자에 의해서 각 10개의 파일이 발음된 TIMIT 데이터베이스가 사용되었다. 음악 데이터베이스는 CD로부터 여러 장르의 음악을 모바일 폰을 통해서 녹음하였고, 8 kHz로 다운 샘플링 되었으며, 5분 정도의 음악파일이 사용되었다. 제안된 음성/음악 분류 알고리즘의 모델은 음성 파일 4200개와 음악

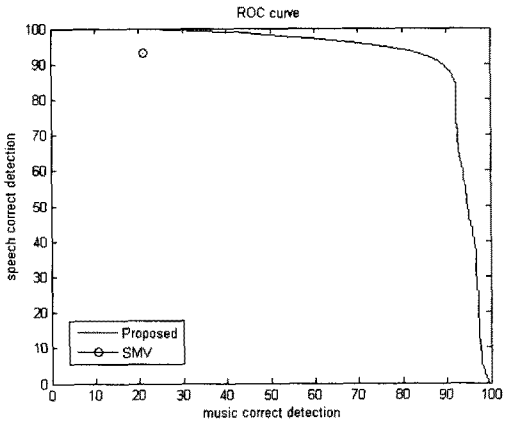

그림 1. SMV와 제안된 알고리즘의 음성/음악 분류에 대한 ROC 곡선
Fig. 1. ROC curve for the speech/audio classification using the SMV and the proposed method.

파일 60개 (메탈 12개, 재즈 12개, 블루스 12개, 힙합 12개, 클래식 12개)를 이용하여 트레이닝 하였다.

SMV와 제안된 알고리즘의 객관적인 성능을 평가하기 위해서 테스트 파일을 만들었다. 동일한 데이터에 의한 성능 향상을 피하기 위해서 트레이닝에 사용된 음성/음악 데이터는 테스트에 사용되지 않았다. 테스트 파일은 5개 음성 파일 (6~12 초), 5개 음악 파일 (28~32초), 10개 무음 (3~15 초)을 사용하여 만들었다.

다양한 음악 장르에 대한 음성/음악 분류 성능을 확인하기 위해서 테스트 파일의 음악을 2가지 형태로 각 장르별 (힙합, 메탈, 재즈, 블루스, 클래식)로 구성된 형태의 테스트 파일 60개, 음악 장르가 혼합된 형태의 테스트 파일 24개 총 84개의 테스트 파일을 만들었다. 두 시스템의 실제 성능을 알아보기 위해서 테스트 파일의 20 ms 마다 실제로 결과를 0 (무음), 1 (음성), 2 (음악)로 수동으로 작성한 것과 비교하였다[11].

그림 1은 기존 SMV의 음성/음악 분류 성능과 제안된 음성/음악 분류 성능을 ROC 곡선으로 표현하여 비교한 것이다. ROC곡선은 SVM에서 이진분류를 할 때 문턱 값을 -1~1까지 0.01씩 증가 시키면서 변화하는 음성/음악의 검출 확률 (P_d)을 나타낸 것이다.

표 2는 SMV와 제안된 SVM 기반의 알고리즘에서 음성/음악 검출 확률 (P_d)을 나타낸다. SVM에서는 문턱 값의 변화에 따라 음성 또는 음악의 P_d값을 조절할 수 있으므로 필요에 따라서 원하는 비율로 사용 할 수 있다. 우측에 P_e (Probability of Error)값은 음성과 음악에 대한 미검출 확률 ($1-P_d$)의 합이며 특히 메탈, 블루스, 힙합, 클래식, 혼합에서 뛰어난 성능을 보였다.

표 2. SMV와 제안된 알고리즘의 음성/음악 분류 성능 비교

Table 2. Comparison of speech/music detection probability P_d between the method of the SMV and the proposed technique.

TEST	Method	Music	Speech	Pe
Metal	SMV	0.22	0.91	0.44
	Proposed	0.90	0.92	0.09
Blues	SMV	0.15	0.90	0.43
	Proposed	0.90	0.90	0.10
Hiphop	SMV	0.28	0.90	0.37
	Proposed	0.66	0.90	0.18
Jazz	SMV	0.27	0.92	0.41
	Proposed	0.35	0.90	0.38
Classic	SMV	0.50	0.90	0.30
	Proposed	0.81	0.91	0.14
Mixed	SMV	0.21	0.93	0.43
	Proposed	0.72	0.90	0.19
R&B	SMV	0.23	0.90	0.43
	Proposed	0.77	0.90	0.17

추가적으로 트레이닝에 사용돼지 않은 음악장르 (R&B) 테스트파일 10개를 따로 실험하였으며 역시 뛰어난 성능을 보였다.

결론적으로 위의 ROC 곡선과 표 2에서 보는 것과 같이 본 논문에서 제안한 알고리즘이 기존의 SMV보다 음성/음악 분류 성능이 월등한 것을 알 수 있다. 따라서 입력 신호의 종류에 의해 매 프레임마다 전송률을 적절히 부여함으로서 이동통신기기에서의 음악, 사진, 영상 등과 같은 다양한 정보를 보다 효율적으로 전송하는 것뿐만 아니라 통화음질의 향상도 기대할 수 있다.

VI. 결 론

본 논문에서는 support vector machine (SVM)을 이용하여 기존의 3GPP2 selectable mode vocoder (SMV)의 음성/음악 분류 성능을 향상시키는 방법을 제시하였다. SMV의 음성/음악 분류알고리즘에서 사용되어진 특징벡터만을 선택적으로 사용하여 효과적인 SVM을 구성한 분류기법을 제시하였다. SMV의 음성/음악 분류에 적용한 SVM의 성능 평가를 위해 SMV 원래의 분류알고리즘과 비교하였으며, 다양한 음악장르에 대해 시스템의 성능을 평가한 결과 SVM을 이용하였을 때 기존의 SMV의 방법보다 우수한 음성/음악 분류 성능을 보였다.

감사의 글

본 연구는 지식경제부 및 정보통신연구진흥원의 IT 핵심기술개발사업 [2008-F-045-01]과 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음 (IITA-2008-C1090-0804-0007)

참 고 문 헌

[1] Y. Gao, E. Shlomot, A. Benyassine, J. Thyssen, Huan-yu Su, and C. Murgia, "The SMV Algorithm Selected by TIA and 3GPP2 for CDMA Applications," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 709-712, May 2001.

[2] 3GPP2 Spec., "Source-controlled variable-rate multimedia wideband speech codec (VMR-WB), service option 62 and 63 for spread spectrum systems," *3GPP2-C.S0052-A*, v.1.0, Apr. 2005.

[3] J. Saunders, "Real-time discrimination of broadcast speech/music," *Proc. IEEE International Conference on Acoustics, Speech, and Processing*, vol. 2, pp. 993-996, May 1996.

[4] W. Q. Wang, W. Gao, and D. W. Ying, "A fast and robust speech/music Discrimination Approach," *Proc. International Conference on Information, Communications and Signal Processing*, vol. 3, pp. 1325-1329, Dec. 2003.

[5] S. Craig Greer, and A. Dejaco, "Standardization of the selectable mode vocoder," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 953-956, May 2001

[6] 3GPP2 Spec., "Selectable Mode Vocoder (SMV) Service Option for Wideband Spread Spectrum Communication Systems," *3GPP2-C.S0030-0*, v3.0, Jan. 2004.

[7] X. Wang, J. Chen, P. Wang, Z. Huang, "Infrared Human Face Auto Locating Based on SVM and A Smart Thermal Biometrics System," *Proc. Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06)*, vol. 2, pp. 1066-1072, Oct. 2006.

[8] A. Ganapathiraju, J. E. Hamaker, J. Picone, "Applications of support vector machines to speech recognition," *IEEE Trans. Signal Processing*, vol. 52, pp. 2348-2355, Aug. 2004.

[9] V. N. Vapnik, "An overview of statistical

learning theory," *IEEE Trans. Neural Networks,* vol. 10, no. 5, pp. 988 -999, Sep. 1999.
[10] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods.* Cambridge Univ. Press, 2000.
[11] J. H. Song, K. H. Lee, J.-H. Chang, J. K. Kim, and N. S. Kim, "Analysis and Improvement of Speech/Music Classication for 3GPP2 SMV based on GMM," *Proc. IEEE Signal Processing Letters,* vol. 15, pp. 103-106, Jan. 2008.

저 자 소 개

김 상 균(학생회원)
2008년 인하대학교 전자공학과 학사 졸업.
2008년~현재 인하대학교 전자공학과 석사과정.
<주관심분야 : 디지털신호처리>

장 준 혁(정회원)
1998년 경북대학교 전자공학과 학사.
2000년 서울대학교 전기공학부 석사.
2004년 서울대학교 전기컴퓨터 박사.
2000년~2005년 (주)넷더스 연구소장
2004년~2005년 캘리포니아 주립대학, 산타바바라(UCSB) 박사후연구원
2005년 한국과학기술연구원(KIST) 연구원
2005년~현재 인하대학교 전자공학부 조교수
<주관심분야 : 음성 신호처리, 오디오 신호처리, 통신 신호처리, 휴먼/컴퓨터 이니터페이스>