한국음향학회지 제27권 제6호 pp. 319~324 (2008)

변별적 가중치 학습을 이용한 3GPP2 SVM의 실시간 음성/음악 분류 성능 향상

Enhancement of Speech/Music Classification for 3GPP2 SMV Codec Employing Discriminative Weight Training

강 상 익*, 장 준 혁*, 이 성 로**
(Sang-Ick Kang*, Joon-Hyuk Chang*, Seong-Ro Lee**)

*인하대학교 전자공학부, **목포대학교 정보공학부
(접수일자: 2008년 6월 5일; 채택일자: 2008년 8월 14일)

본 논문에서는 변별적 가중치 학습 (discriminative weight training) 기반의 3GPP2 Selectable Mode Vocoder (SMV) 실시간 음성/음악 분류 성능을 향상 시키는 방법을 제안한다. SMV의 음성/음악 실시간 분류 알고리즘에서 사용된 특징벡터와 분류방법을 분석하고, 이를 기반으로 분류성능향상을 위해 MCE (minimum classification error)방법을 도입하여, 각 특징벡터별로 다른 가중치를 적용하는 음성/음악 결정법 (decision rule)을 제시한다. 구체적으로, SMV의 음성/음악 분류알고리즘에서 사용되어진 특징벡터만을 선택적으로 사용하여 가중치를 적용한 값을 기하 평균한 값을 문턱값과 비교하는 실시간 분류기법이 제시되었다. SMV의 음성/음악 분류에 제안한 방법의 성능 평가를 위해 SMV 원래의 분류알고리즘과 비교하였으며, 다양한 음악장르에 대해 시스템의 성능을 평가한 결과 가중치를 적용하였을 때 기존의 SMV의 방법보다 우수한 음성/음악 분류 성능을 보였다.

핵심용어: 음성/음악 분류, Minimum classification error, 변별적 가중치 학습, Selectable mode vocoder
투고분야: 음성처리 분야 (2)

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 discriminative weight training which is based on the minimum classification error (MCE) algorithm. We first present an effective analysis of the features and the classification method adopted in the conventional SMV. And then proposed the speech/music decision rule is expressed as the geometric mean of optimally weighted features which are selected from the SMV. The performance of the proposed algorithm is evaluated under various conditions and yields better results compared with the conventional scheme of the SMV.

Keywords: Speech/music classification, Minimum classification error, Discriminative weight training, Selectable mode vocoder
ASK subject classification: Speech Signal Processing (2)

I. 서 론

최근 이동통신의 발전으로 무선통신기기를 이용한 멀티미디어 서비스가 보편화 되면서 제한적인 주파수 대역에서 효과적으로 음성을 전송하는 연구가 지속적으로 이루어지고 있다. 현재 제한된 통신망을 효율적으로 사용하기 위해 가변적인 전송률을 갖는 다양한 음성 코덱이 개발 되었다 [1][2]. 실제로 입력 음성신호의 유형에 따라서 다른 비트를 할당하는 것은 바로 최종 음성의 음질에 영향을 미치기 때문에 정확한 신호분류기술의 핵심기술로서 다루어지고 있다. 특히, 단순히 음성통신을 다루던 것에서 벗어나 음악신호를 이동통신망을 통해 효과적으로 전송하기 위한 음성/음악 분류의 중요성이 증가하여 관련된 연구가 활발히 진행되고 있다 [3-5].

본 논문에서는 실시간 음성/음악 분류기법을 기반으로 가변 전송률 알고리즘을 채택하고 있는 ETSI의 3GPP2 표준코덱인 Selected Mode Vocoder (SMV)의 기존 방법을 분석하고 이를 기반으로 음성/음악 분류성능을 향상시키기 위한 기법을 제안한다. 구체적으로, 기존의 SMV의

책임저자: 장 준 혁 (changjh@inha.ac.kr)
402-751 인천시 남구 용현동 253 인하대학교 전자공학부
(전화: 032-860-7423; 팩스: 032-868-3654)

인코딩부분의 전처리과정에서 자동적으로 추출되는 파라미터 중 통계적인 분류특성이 우수한 것들을 모아 특징벡터로 사용하였다. 변별적 가중치 학습 (discriminative weight training)을 이용하여 도출된 최적화된 가중치를 선정된 특징벡터에 적용 후 기하 평균한 결정식으로 음성/음악 분류하는 알고리즘을 개발하고 이것을 기존의 SMV 방법과 다양한 환경에서 비교하였다.

본 논문의 II 장에서는 기존의 SMV 코덱에서의 음성/음악 분류 방법과 특징 벡터에 대해서 알아본다. III 장에서는 변별적 가중치 학습을 이용한 음성/음악 분류 알고리즘을 제시한다. IV 장에서 다양한 실험 환경에서 제안된 알고리즘과 기존의 알고리즘의 결과를 비교 검토한 뒤 V 장에서 결론을 맺는다.

II. SMV (Selectable Mode Vocoder) 음성/음악 검출 알고리즘

SMV 음성부호화기는 3GPP2의 표준화된 저전송율 음성 코덱이다 [6][7]. SMV는 8 kHz로 샘플링 된 입력 신호를 160개의 샘플 (20 ms)마다 프레임의 모드와 전송률을 결정하게 되는데 [8] 4가지 전송률과 4개의 동작모드를 갖고 있으며 이러한 다양한 평균 전송률을 갖는 특성 때문에 CDMA 시스템의 성능과 음질간의 관계에서 선택적으로 성능을 조절 할 수 있다 [9]. 입력된 신호는 전처리 과정을 거치게 되며 묵음 증가, 고대역 통과 필터, 노이즈 억제, 적응 틸트 보상 등을 통해서 백그라운드 노이즈를 제거한다. 프레임 처리 과정에는 선형 예측 분석, 개회로 (open-loop) 피치 검출, 신호 수정 및 분류 등이 포함된다.

선형 예측 분석과 개회로 피치 검출을 통해서 음성과 음악 분류에 대한 파라미터들이 추출되는데 신호 분류를 통해서 프레임은 잡음, 묵음, 무성음, 비정상적 유성음, 정상적 유성음중 한 개로 분류 된다. 프레임의 전송률은 현재 프레임의 분류된 종류와 통신 상태에 따라서 정해지는 모드를 바탕으로 전송률 결정 알고리즘 (Rate Determination Algorithm, RDA)에 의해 결정된다. 프레임이 음악으로 판별된 경우 프레임의 전송률은 모드에 상관없이 Rate 1 (8.55 kbps)로 결정되어 최고의 전송률을 할당하며 그 외의 경우에는 정해진 한계 값에 의해서 결정하여 인코딩 한다 [8].

SMV는 음성 검출 (Voice Activity Detection, VAD)에서 무음과 음성으로 분류 한 뒤 음악 분류를 한다. 그림 1은 SMV 음악 분류 알고리즘의 블록을 나타내는데 VAD, 개회로 피치 검출, 선형 예측 부호화 (Linear Prediction Coding, LPC)에서 추출된 특징 벡터를 입력 받고, 각 특징 벡터의 이동 평균 값을 음악 분류 알고리즘의 특징벡터로 이용한다. 이러한 특징 벡터들은 경험적으로 설정되어 문턱 값과 비교하여 음성/음악 분류를 한다. 구체적으로, 음악 분류 알고리즘에 입력되는 특징 벡터에 대해서 다음 세부 절에서 살펴본다.

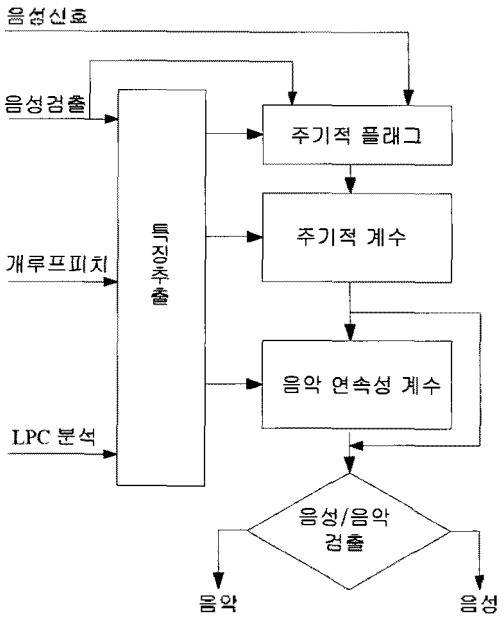

그림 1. 기존의 SMV 음성/음악 분류 알고리즘 블록도
Fig. 1. Block diagram of the speech/music classification algorithm of the SMV.

2.1. LPC 분석에서 추출된 특징 벡터

1) 반사 계수 : i번째 반사 계수 (Reflection Coefficients, $k_1(i)$)가 Levinson-Durbin 알고리즘에 의해서 구해진다 [10].
2) Line spectral frequencies (LSF) : LSF는 인코딩 프레임의 마지막 1/4에 가중치를 갖고 구해진 10차 LPC 값을 변화한 값으로 $lsf(i)$로 나타낸다 [11].

2.2. VAD에서 추출된 특징 벡터

1) 프레임 에너지 : 프레임에 대한 에너지 값으로 최소 10 이상의 값을 갖는다.

$$E = \max\left(10, 10 \cdot \log_{10}\left(\frac{R_1(0)}{L_{LPC}}\right)\right) \quad (1)$$

여기서 $R_1(0)$값은 신호의 파워, L_{LPC} 값은 LPC 윈도우의 길이로 240을 갖는다.

2) VAD decision flag : VAD에서 추출된 파라미터들과 설정된 문턱값의 비교를 통해서 현재 프레임이 음성인지를 판단한다.

2.3 음악 분류 알고리즘 특징 백터

1) LSF의 이동평균 :

$$\overline{lsf}(1) = 0.75 \cdot \overline{lsf}(1) + 0.25 \cdot lsf_1(1) \quad (2)$$

여기서 $lsf_1(1)$는 LPC 분석을 통해서 구해진 10차의 lsf 값 중 첫 번째 값을 나타낸다 [7]. lsf_1은 3개의 LPC 윈도우 중에 3번째에 중점을 둔 특징벡터를 의미한다.

2) 에너지의 이동 평균 : 식(1)에 의해서 구해진 에너지의 이동 평균 값이다.

$$\overline{E} = 0.75 \cdot \overline{E} + 0.25 \cdot E. \quad (3)$$

3) 스펙트럼 차이 : 프레임의 반사 계수와 노이즈 프레임의 반사 계수의 차이에 의해서 구해진다.

$$SD_1 = \sum_{i=1}^{10} \left(k_1(i) - \overline{k_N}(i)\right)^2 \quad (4)$$

4) 정규화 된 피치 상관도의 이동 평균 :

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

여기서 $corr_p^B(i)$는 이전 프레임의 피치 상관도 이며, 여기서 i는 이전 프레임을 나타낸다.

5) 주기적 계수 (periodicity counter)의 이동 평균 : (2)~(5)의 특징 벡터와 설정된 문턱 값의 비교를 통해서 증가하거나 감소하고, $\overline{c_{pr}} \geq 18$ 이면 프레임을 음악으로 분류한다.

$$c_{pr} = \alpha \cdot \overline{c_{pr}} + (1-\alpha) \cdot c_{pr} \quad (6)$$

여기서 α 값은 고정된 가중치 값을 나타낸다.

6) 음악 연속성 계수 (music continuity counter)의 이동 평균 : (2)~(6)의 특징 벡터와 설정된 문턱 값의 비교를 통해서 증가하거나 감소하고, $\overline{c_M} \geq 200$이면 해당 프레임을 음악으로 분류한다.

$$c_M = 0.9 \cdot \overline{c_M} + 0.1 \cdot c_M. \quad (7)$$

따라서, 최종 SMV의 음성/음악 분류식을 다음과 같이 정의된다.

$$\begin{cases} \overline{c_M} > 200 \text{ or } \overline{c_{pr}} \geq 18 & : \text{음악} \\ else & : \text{음성} \end{cases} \quad (8)$$

III. 변별적 가중치 학습 기반 음성/음악 분류

기존의 SMV를 이용한 음성/음악 분류 시스템은 VAD에서 음성이라고 판별된 프레임에서 주기적 계수의 이동 평균과 음악 연속성 계수의 이동 평균을 이용하여 음성과 음악을 분류하게 된다. 본 논문에서는 기존의 SMV 코덱에서 추출되는 특징 백터 중 몇 가지에 대해 변별적가중치 학습을 이용하여 성능에 기여하는 정도에 따라 최적화된 가중치를 인가하는 효과적인 프레임별 음성/음악 분류 결정식을 제안하고 다음과 같이 정의한다.

$$f(\boldsymbol{WF}) = \sum_{k=1}^{M} w_k f_k \underset{\text{speech}}{\overset{\text{music}}{\gtrless}} \eta \quad (9)$$

여기서 $\boldsymbol{W}$와 $\boldsymbol{F}$는 각각 $\{w_1, w_2, \cdots, w_M\}$, $\{f_1, f_2, \cdots, f_M\}$으로 기존의 SMV에서 도출한 특징백터 에너지 이동평균 $\overline{E}$, 정규화 된 피치 상관도의 이동평균 $\overline{corr_P}$, 음악 연속성 계수 c_M에 각각 다른 가중치 w_k를 적용하여 구하며 각 가중치는 다음의 조건을 만족한다.

$$\sum_{k=1}^{M} w_k = 1, \quad w_k \geq 0 \text{ for } k = 1, \cdots, M. \quad (10)$$

훈련할 데이터의 각각의 프레임에서 음악 $g_m(\cdot)$과 음성 $g_s(\cdot)$을 구분하는 두 개의 함수를 다음과 같이 정의한다.

$$g_m = f(\boldsymbol{WF}) - \theta \quad (11)$$

$$g_s = \theta - f(\boldsymbol{WF}), \tag{12}$$

여기서 θ는 음악과 음성을 구분하는 문턱값이며 T는 전치행렬이다. 제안된 연구에서는 최적화 알고리즘에 기반한 가중치를 구하기 위해 generalized probabilistic descent (GPD) 기반의 MCE 훈련을 적용하며 [12], 실제로 훈련 데이터 프레임의 분류 오류 D를 다음과 같이 정의한다.

$$D(t) = \begin{cases} -g_m(t) + g_s(t) \text{ if } g_m \textit{ is true class} \\ -g_s(t) + g_m(t) \text{ if } g_s \textit{ is true class} \end{cases} \tag{13}$$

여기서 식(13)이 음수인 값을 가질 때 올바른 분류가 되며 이를 기반으로 하는 손실함수 (loss function) L은 다음과 같이 sigmoid 함수 형태로 정의된다.

$$L = \frac{1}{1+\exp(-\beta D)}\ ,\ \beta > 0 \tag{14}$$

여기서 β는 sigmoid 함수의 기울기를 나타낸다. 최적화된 가중치를 구하기 위해선 손실함수가 최소가 되어야한다. MCE 훈련과정을 통해 가중치를 조정하는 과정에서 식(10)과 같은 제약조건 때문에 가중치 w를 $\tilde{w}$로 변환한다.

$$\tilde{\boldsymbol{W}} = \{\tilde{w_1}, \tilde{w_2}, \cdots, \tilde{w_k}\} \tag{15}$$

$$\tilde{w_k} = \log w_k . \tag{16}$$

가중치 $\tilde{w_k}$는 매 프레임마다 연속적으로 존재하는데, 각 주파수 가중치는 다음과 같은 식으로 갱신된다 [13].

$$\tilde{w_k}(t+1) = \tilde{w_k}(t) - \epsilon \frac{\partial L}{\partial \tilde{w_k}}\bigg|_{\tilde{w_k} = \tilde{w_k}(t)} \tag{17}$$

여기서 ϵ (> 0) 는 단조롭게 감소하는 구간의 크기이다. $\tilde{w_k}$를 갱신한 후에 아래의 식과 같이 w_k로 복원된다.

$$w_k = \frac{\exp(\tilde{w_k})}{\sum_{i=1}^{M} \exp(\tilde{w_i})} \tag{18}$$

식(18)에서 정규화 된 가중치를 사용했을 때 식(10)을 만족한다.

기존의 SMV의 결정 방법과 비교하여, 본 논문에서는 위에서 제시된 MCE 훈련방법을 이용해 구한 식(18)의 가중치를 SMV의 특징벡터에 적용한 결정식으로 최종적으로 음성/음악을 구분한다. 결론적으로 기존의 다른 SVM 음성/음악 분류 알고리즘과 달리 특징벡터가 성능에 기여하는 정도에 따라 가중치를 적용한 후 간단한 기하 평균 기반의 결정식을 도출하는 점이 주목할 만하다.

IV. 실험 결과

본 장에서는 SMV의 음성/ 음악 분류 알고리즘에 적용한 변별적 가중치 학습 기반의 실시간 음성/음악 분류 성능을 알아보기 위해, 기존의 SMV의 알고리즘과 비교하였다. 본 실험을 위해서 음성 데이터베이스로 8 kHz로 샘플링 된 약 6 sec 정도의 깨끗한 음성으로 326명의 남자와 138명의 여자 화자에 의해서 각 10개의 파일이 발음된 TIMIT 데이터베이스가 사용되었다 [14]. 음악 데이터베이스는 CD로부터 여러 장르의 음악을 모바일 폰을 통해서 녹음하였고, 8 kHz로 다운 샘플링 하여 사용하였으며, 5분 정도의 음악파일이 사용되었다. 제안된 음성/음악 분류 알고리즘의 가중치를 도출하기 위해 음성 파일 4200개와 음악 파일 60개 (메탈 12개, 재즈 12개, 블루스 12개, 힙합 12개, 클래식 12개)를 이용하여 가중치 학습을 하였다.

SMV와 제안된 알고리즘의 객관적인 성능을 평가하기 위해서 테스트 파일을 만들었으며 동일한 데이터에 의한 성능 향상을 피하기 위해서 트레이닝에 사용된 음성/음악 데이터는 테스트에 사용되지 않았다. 테스트 파일은 그림 2와 같은 형태로 5개 음성 파일 (6~12 초), 5개 음악 파일 (28~32 초), 10개 무음 (3~15 초)을 사용하여 만들었다. 다양한 음악 장르에 대한 음성/음악 분류 성능을 확인하기 위해서 테스트 파일의 음악을 2가지 형태로 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 시스템 실제 성능을 알아보기 위해서 테스트 파일의 20 ms 마다 실제로 결과를 비음성, 음성, 음악으로 수동으로 작성한 것과 비교하였다.

SMV과 제안된 알고리즘의 성능 평가를 위하여 혼합된 형태의 테스트 파일에 대해서 Receiver Operating Characteristic (ROC) 곡선을 그림 3에 도시하였다. ROC 곡선은 문턱값을 변화 시키면서 음성/음악에 대한 검출 확률 (P_d)을 구하였다. 여기서 P_d는 음악프레임 일 경우 음악으로 분류된 확률과 음성프레임 일 경우 음성으로 분류된

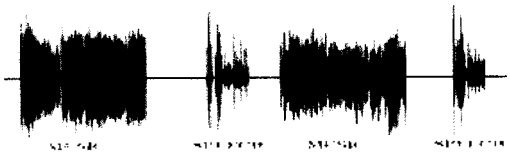

그림 2. 테스트 파일의 형태
Fig. 2. Shape of test file.

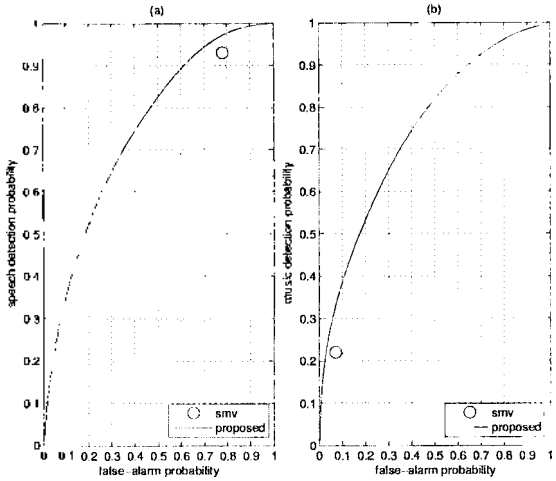

그림 3. ROC에 기반한 SMV과 제안한 방법의 성별인식 성능 비교 (a) 음성 (b) 음악
Fig. 3. Speech/audio identification performance of SMV and proposed method based on ROC. (a) speech (b) music.

확률을 뜻한다. SMV의 음성/음악 분류 알고리즘은 식(8)과 같이 주기적 계수와 음악 연속성 계수의 각각의 조건에 or 연산 형태이므로 문턱값을 변화시키면서 ROC 곡선을 그릴 수 없으므로 음성, 음악 각각의 성능을 점으로 표기하였다. 실험결과 음성과 음악 전체에서 제안된 음성/음악 분류 알고리즘 성능이 약 10% 우수함을 확인할 수 있었다.

V. 결 론

본 논문에서는 ETSI의 3GPP2 표준코덱인 SMV의 실시간 음성/음악 분류 성능을 향상시키기 위해 기존의 SMV 코딩 특징 벡터를 이용하여 각각의 특징벡터에 서로 다른 가중치를 적용하여 기하 평균한 값을 문턱값과 비교하여 음성/음악 분류하는 방법을 제안하였다. SMV의 음성/음악 분류 알고리즘에 사용되는 통계적 분류 특성이 우수한 특징 벡터에 가중치를 적용하였고, 기존의 SMV와 성능을 비교한 결과 제안된 방법을 이용하여 SMV의 분류 성능 향상한 기법이 기존의 SMV의 분류성능에 비해서 향상된 실시간 음성/음악 분류 성능을 보여 주었다.

감사의 글

본 연구는 지식경제부 및 정보통신연구진흥원의 IT핵심기술개발사업 [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, H.-Y. 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, 2, 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*, 2, 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, 3, 1325-1329, Dec. 2003.
5. 금지수, 임성길, 이현수, "스펙트럼 분석과 신경망을 이용한 음성/음악 분류", 한국음향학회지, 26(5), 207-213, Jul. 2007.
6. J. Makinen, P. Ojala, and H. Toukomaa, "Performance comparison of source controlled GSM AMR and SMV vocoders," Proc. International Symposium on Intelligent Signal Processing and Communication Systems, 51-154, Nov. 2004.
7. C. V. Goudar, P. Rabha, M. Deshpande, and A. Rao, "SMVLite: Reduced Complexity Selectable Mode Vocoder," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 1, 701-704, May 2006.
8. 3GPP2 Spec., "Selectable mode vocoder (SMV) service option for wideband spread spectrum communication systems," 3GPP2-C.S0030-0, v3.0, Jan. 2004.
9. S. Craig Greer, and A. Dejaco, "Standardization of the selectable mode vocoder," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2, 953-956, May 2001.
10. P. Vary and R. Martin, Digital Speech Transmission : enhancement, coding and error concealment, pp.182-187, 2006.
11. P. Kabal, R. Prakash and Ramachandran, "The computation of line spectral frequencies using Chebyshev polynomials," IEEE Trans. Acoustics, speech and signal processing, ASSP-34(6), 1419-1426, Dec. 1986.
12. B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," IEEE Trans. Speech Audio Processing, 5(3), 257-265, May 1997.
13. S.-I. Kang, Q.-H. Jo, J.-H. Chang, "Discriminative weight training for a statistical model-based voice activity detection," IEEE Signal Processing Letters, 15, 170-173, Feb. 2008.
14. W. M. Fisher, G. R. Doddington and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifi-

cations and status," Proc. DARPA Workshop Speech Recognition, pp.93-99, Feb. 1986.

저자 약력

•강 상 익 (Sang-Ick Kang)

2007년 2월: 인하대학교 전자공학과 학사
2007년 3월~현재: 인하대학교 전자공학과 석사과정

•장 준 혁 (Joon-Hyuk Chang)

1998년 2월: 경북대학교 전자공학과 학사
2000년 2월: 서울대학교 전기공학부 석사
2004년 2월: 서울대학교 전기컴퓨터공학부 박사
2000년 3월~2005년 4월: 인터넷더스 연구소장
2004년 5월~2005년 4월: 캘리포니아 주립대학, 산타바바라 (UCSB) 박사후연구원
2005년 5월~2005년 8월: 한국과학기술연구원 (KIST) 연구원
2005년 9월~현재: 인하대학교 전자전기공학부 조교수

•이 성 로 (Seong-Ro Lee)

1987년 2월: 고려대학교 전자공학과 졸업
1990년 2월: 한국과학기술원 전기및전자공학과 석사
1996년 8월: 한국과학기술원 전기및전자공학과 박사
2005년 3월~현재: 목포대학교 정보공학부 정보전공 학전공 부교수