말소리와 음성과학 제2권 제1호 (2010)
pp. 77~85

잡음 환경에서의 음성 감정 인식을 위한 특징 벡터 처리

Feature Vector Processing for Speech Emotion Recognition in Noisy Environments

박 정 식[1] · 오 영 환[2]
Park, Jeongsik · Oh, Yunghwan

ABSTRACT

This paper proposes an efficient feature vector processing technique to guard the Speech Emotion Recognition (SER) system against a variety of noises. In the proposed approach, emotional feature vectors are extracted from speech processed by comb filtering. Then, these extracts are used in a robust model construction based on feature vector classification. We modify conventional comb filtering by using speech presence probability to minimize drawbacks due to incorrect pitch estimation under background noise conditions. The modified comb filtering can correctly enhance the harmonics, which is an important factor used in SER. Feature vector classification technique categorizes feature vectors into either discriminative vectors or non-discriminative vectors based on a log-likelihood criterion. This method can successfully select the discriminative vectors while preserving correct emotional characteristics. Thus, robust emotion models can be constructed by only using such discriminative vectors. On SER experiment using an emotional speech corpus contaminated by various noises, our approach exhibited superior performance to the baseline system.

Keywords: Speech emotion recognition, noisy environments, comb filtering, feature vector classification

1. 서론

인간의 삶의 질을 향상시키기 위한 목적으로 끝없이 진보해 온 인공지능 기술은 이제 인간과 기계 사이의 거리를 좁히기 위해 다양한 방법을 시도하고 있다. 휴대폰, 내비게이션과 같은 생활필수품은 손을 이용하는 인터페이스를 거쳐 음성으로 구동되는 형태로 진화하고 있으며, 굴러다니는 로봇 청소기의 모습은 향후 몇 십 년 내에 휴먼 로봇의 형태로 인간과 눈높이를 맞추게 될 것이다. 이처럼 인간-기계 인터페이스는 사용자 편의(user-friendly)를 향상시키는 것에서 나아가 사용자를 이해하는(user-comprehensive) 수준으로 발전하고 있다. 선진국을 중심으로 상용화되고 있는 무인 자동응답기를 이용한 예약 서비스의 경우 통화 음성으로부터 상대방의 감정 상태를 파악하는 것이 필수적이며, 간병 로봇, 가사 로봇 등 향후 사람의 곁에서 직접적으로 도움을 제공하는 서비스 로봇 역시 사용자의 음성 및 표정으로부터 마음 상태를 인지하는 것이 무엇보다 중요하다. 이처럼 인간-기계 인터페이스의 성능을 사용자를 이해하는 수준으로 향상시키기 위해서는 감정 인식 기술이 필수적이며, 특히 음성을 통한 감정 인식은 고가의 장비 없이 원거리에서도 사용자의 감정 인지가 가능하다는 장점을 지닌다.

지금까지 음성 감정 인식을 위한 다양한 연구가 지속적으로 수행되어 왔다. 감정 특성을 나타내는 파라미터 및 감정 분류에 유용한 다양한 식별 방법이 연구되었으며[1]-[3], 두 가지 또는 세 가지의 감정을 대상으로 한 감정 인식의 경우 70% 이상의 성능을 보였는데, 대부분의 연구에서 성능 평가에 사용된 음성 자료는 잡음이 없는 실험실 환경에서 녹음된 자료들이다. 음성 인식 또는 화자 인식 등의 연구 분야에서 밝혀진 바와 같이 음성 신호에 포함된 배경 잡음은 인식 성능의 저하를 유발하는

1) 한국과학기술원 dionpark@speech.kaist.ac.kr
2) 한국과학기술원 yhoh@speech.kaist.ac.kr, 교신저자
본 연구는 방위사업청과 국방과학연구소의 지원으로 수행되었습니다.

접수일자: 2009년 11월 1일
수정일자: 2009년 12월 9일
게재결정: 2010년 1월 16일

대표적인 요인이다. 특히 감정 인식의 경우 감정들 사이의 모호성이 존재하고 화자마다 감정을 표현하는 방식이 다르므로, 잡음으로 인한 특징 벡터의 변이는 더욱 심각한 인식률 저하를 야기할 수 있다.

잡음에 인한 음성 인식 성능 저하 문제를 해결하기 위해 주파수 차감법, MMSE-LSA 등 다양한 음질 개선 기법이 연구되었고 이들에 대한 성능 평가가 수행되었다[4]. 그러나 감정 인식과 음성 인식에서 가장 유용하게 사용되는 특징 파라미터의 종류가 각기 다르기 때문에, 명료성 및 음질 개선을 목표로 한 잡음 제거보다는 잡음에 의해 변이가 발생한 특징 파라미터를 감정 인식에 유용하도록 개선하는 방법이 연구될 필요가 있다.

본 연구에서는 잡음 환경에서의 감정 인식 성능을 향상시키기 위해 잡음 제거에 유용한 콤 필터링을 적용하여 특징 파라미터를 개선하고, 또한 특징 벡터 선별 기법을 적용하여 잡음 환경에 강인한 감정 모델을 구축하는 방법을 제안한다. 감정 특성을 가장 잘 나타내는 대표적인 특징 파라미터는 피치 주기(pitch period)이며, 콤 필터링은 음성의 기본 주파수, 즉 피치 주기를 사용하여 잡음을 제거하는 대표적인 방법이다. 콤 필터링에 의해 처리된 음성은 스펙트럼 상에서 고조파(harmonics) 성분이 강조된 특성을 보이므로 이 음성으로부터 추출된 피치 정보는 감정 인식에 유용하게 사용될 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 음성 감정 인식 방법에 대하여 기술하고 3장에서는 본 연구에서 제안한 특징 벡터 처리 방법을 소개하며 4장에서는 제안한 방법의 유효성을 검증하기 위해 수행한 감정 인식 실험 및 결과를 기술한다. 끝으로 5장에서 결론을 맺는다.

2. 음성 감정 인식

음성 감정 인식은 감정 특성을 잘 표현하는 특징 파라미터 및 감정 분류에 유용한 식별 방법이 중점적으로 연구되고 있다. 피치, 에너지, 지속 길이, MFCC 등 비교적 짧은 구간에서 추출된 음향 특징 파라미터들이 감정 인식 시스템에 주로 사용되며, 이 중 피치 정보는 감정 정보를 표현하는데 효과적인 특징 파라미터로 알려졌다[2]. 인식 단계에서 활용되는 식별 방법으로는 음성 인식 및 화자 인식 등에서 사용되는 Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Support Vector Machine (SVM), Artificial Neural Network (ANN) 등이 적용되었으며, 이 중 GMM 기반의 식별 방법이 피치나 MFCC와 같은 단구간 특징 파라미터에 적합하다는 연구결과가 있었다[2],[5]. 이 같은 연구 내용을 기반으로 본 연구에서는 단구간 특징 파라미터 및 GMM 기반의 감정 인식 시스템을 사용한다.

2.1 GMM 기반의 음성 감정 인식 시스템

GMM 기반의 음성 감정 인식은 각 감정에 따른 GMM을 구축하는 훈련 단계와 GMM을 이용하여 입력 음성을 인식하는 단계로 이루어진다. 훈련 단계는 <그림 1>과 같이 감정별로 분류된 음성 자료로부터 특징 벡터를 추출한 다음 Expectation-Maximization (EM) 알고리즘을 통해 GMM을 구축하는 과정이며, 각 GMM의 파라미터, 즉 가우시언 분포의 평균과 분산이 계산된다. 인식 단계에서는 입력된 음성에 대해 GMM 파라미터를 이용하여 로그-우도를 계산하는데, 가령 입력 음성으로부터 추출한 특징 벡터열 $X=\{\overrightarrow{x_1},\ldots,\overrightarrow{x_T}\}$이 주어졌을 때, 각 GMM ($\lambda_i$; 감정의 수가 E개인 경우, $i=1,\ldots,E$)에 대한 로그-우도는 아래와 같은 식을 통해 계산된다.

$$\log P(X|\lambda_i)=\sum_{t=1}^{T}\log P(\overrightarrow{x_t}|\lambda_i)\ \ (i=1,\ldots,E) \tag{1}$$

식 (1)로부터 계산된 E개의 로그-우도 중 가장 큰 값을 나타내는 감정 모델이 입력 음성의 감정으로 최종 결정된다.

2.2 음성 감정 인식에서의 잡음의 영향

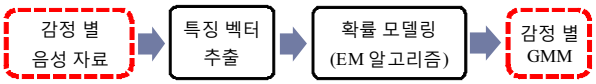

그림 1. GMM 기반의 음성 감정 인식을 위한 모델 학습 과정
Figure 1. Model training for GMM-based SER

음성 인식 등 다양한 분야에서 밝혀진 바와 같이 잡음으로 인한 음성 신호의 왜곡은 인식 성능을 저해하는 대표적인 요인이다. 감정 인식의 경우 인간의 판별력으로도 구분이 모호한, 감정 자체의 특성뿐만 아니라 화자 및 지역적 특성에 따라 감정을 표현하는 방식이 다르므로 잡음으로 인한 음성 신호의 왜곡은 더욱 심각한 인식률 저하를 야기할 수 있다. 특히 인식에 가장 효과적인 파라미터인 피치 주기는 잡음에 의해 쉽게 변질되는 특성을 보이기 때문에 잡음이 심한 음성의 경우 피치 주기를 비롯한 특징 파라미터의 정확도가 떨어진다.

잡음에 의한 감정 인식률의 저하를 개선하기 위해서는 주파수 차감법, MMSE-LSA 등의 음질 개선 기법을 적용하여 신호에 포함된 잡음 성분을 제거하는 것이 바람직하다. 그러나 스펙트럼 포락이나 켑스트럼 특성보다는 피치 정보, 에너지 등의 정보에 의해 인식 성능이 좌우되는 감정 인식의 경우 음질 개선뿐만 아니라 잡음에 의해 변이가 발생하는 특징 파라미터를 개선하는 것이 함께 고려되어야 한다. 또한 음질 개선 후에도 고유의 감정 특성이 훼손된 특징 벡터가 훈련 단계에 포함되지 않도록 처리하는 것 또한 인식 성능 향상에 기여할 것으로 판단된다. 본 연구에서는 이를 위해 기본 주파수 정보를 이용하여

잡음을 제거하는 콤 필터링 기법을 사용하고, 극심한 잡음에 의해 본래의 감정 특성이 훼손된 특징 벡터를 선별하기 위하여 화자 식별 분야에 사용된 바 있는 특징 벡터 선별 기법을 적용하는 방법을 제안한다.

3. 잡음 음성에서의 감정 인식을 위한 특징 벡터 처리

3.1 개선된 콤 필터링을 이용한 잡음 제거

유성음의 파형은 기본 주파수에 따라 주기적이라는 특성에 기반을 둔 콤 필터링은 음성 신호의 고조파(harmonics)에 해당하는 주파수 대역을 강조함으로써 고조파에 포함된 음성 성분을 보존하고, 고조파 사이의 주파수 대역의 에너지를 낮춤으로써 잡음 성분을 제거하는 음질 개선 기법이다[6]. 콤 필터링은 다음과 같은 단위 샘플 응답 $h(n)$을 통해 처리된다.

$$h(n)=\sum_{k=-L}^{L}\alpha_k\times\delta(n-kT) \tag{2}$$

$\delta(n)$은 단위 샘플 함수이며, T는 피치 주기, $2L+1$은 필터의 길이를 의미한다. α_k는 필터 계수로서 $\sum_{k=-L}^{L}\alpha_k=1$을 만족해야 하며 일반적으로 식 (3)과 같은 Hamming window 형태를 따른다.

$$\alpha_k=\frac{0.54+0.46\cos\left(2\pi k/(2L+1)\right)}{\sum_{k=-L}^{L}\left\{0.54+0.46\cos\left(2\pi k/(2L+1)\right)\right\}} \tag{3}$$

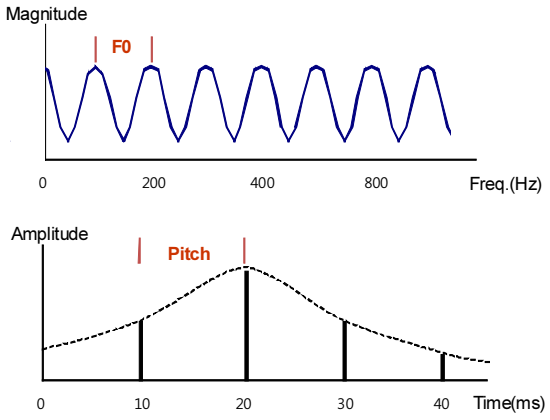

그림 2. 콤 필터의 주파수 응답(上) 및 임펄스 응답(下)
Figure 2. Frequency response (above) and impulse response (below) of a comb filter

<그림 2>는 100Hz의 기본 주파수(즉, 10ms의 피치 주기)를 갖는 신호로부터 생성된 콤 필터의 주파수 응답과 임펄스 응답을 나타낸 것이다. 그림에서 확인되는 바와 같이 기존의 콤 필터는 전체 스펙트럼 대역에서 주파수 응답이 기본 주파수에 따라 일정하게 반복되는 특성을 보이며 또한 필터 계수 α_k 역시 원래의 음원과 무관하게 window 함수에 의해 미리 정해진 값이 필터링에 사용되는 특성이 있다. 따라서 정확한 피치 주기를 측정하는데 제약이 있는 잡음이 심한 음성을 대상으로 기존의 콤 필터링을 적용하는 경우 다음과 같은 문제점이 있다. 즉, 주로 음성 성분이 포함된 고조파 대역의 에너지가 감소되고 반대로 잡음 성분이 포함된 주파수 대역(고조파들 사이의 대역)의 에너지가 증가되는 문제가 발생한다. 잘못 측정된 피치 주기에 기인한 이 같은 현상은 콤 필터링을 수행한 후 음성 신호의 왜곡을 야기할 수 있으며, 이 음성으로부터 추출한 피치 정보가 감정 인식의 파라미터로 사용된다면 심각한 오인식의 원인이 된다. 잡음이 심한 환경에서 발생하는 이 같은 문제점을 해결하기 위해 본 연구에서는 기존의 콤 필터의 주파수 응답 및 필터 계수 (α_k)를 개선하고자 한다.

전체 스펙트럼 대역에서 주파수 응답이 일정하게 반복되는 콤 필터링의 문제점을 해결하기 위해 Minimum Mean Squared Error-Log Spectral Amplitude (MMSE-LSA) 기법에서 사용된 추정값을 이용하는 방법이 연구된 바 있다[7]. 본 연구에서는 이 방법에서 사용된 각 주파수 대역별 음성 존재 확률을 이용하여 콤 필터의 주파수 응답을 변형한 후 이를 역 푸리에 변환시켜 얻은 임펄스 응답으로부터 새로운 필터 계수를 얻는다.

3.1.1 음성 존재 확률을 이용한 콤 필터의 변형

l-번째 프레임의 각 주파수 대역에서 추정된 음성 존재 확률을 $P_l(\omega)$라고 할 때 다음 식을 통해 콤 필터의 주파수 응답을 변형한다.

$$\hat{A}_l(\omega)=P_l(\omega)\times A_l(\omega) \tag{4}$$

$A_l(\omega)$는 l-번째 프레임에서 생성된 콤 필터의 주파수 응답 중 주파수 대역 ω에서의 스펙트럼 크기를 나타낸다. 식 (4)에 의해, 각 대역별 음성 존재 확률에 따라 기존의 콤 필터의 주파수 응답이 변형된다. 즉, 음성 존재 확률이 큰 대역일수록 $P_l(\omega)$는 1에 가까워지고 해당 대역에서 주파수 응답의 크기는 보존된다. 반면, 음성 존재 확률이 작은 대역에서는 $P_l(\omega)$의 값이 0에 가까워지므로 해당 대역의 주파수 응답의 크기가 줄어든다. 앞서 설명한 바와 같이, 잡음이 심한 음성의 경우 잘못 측정된 피치 정보에 의한 콤 필터링이 오히려 음성 신호의 왜곡을 야기할 수 있다. 본 연구에서는 음성 존재 확률을 이용하여 콤 필터의 주파수 응답을 조정함으로써 이 같은 왜곡 문제를 해결한다.

음성 존재 확률 계산을 위해 MMSE-LSA 기법에서 사용되는 Gain 함수(식 (5))를 적용한다.

$$G_l(\omega) = \frac{\Lambda_l(\omega)}{1+\Lambda_l(\omega)} \approx 1 - q_l(w) \quad (5)$$

$G_l(\omega)$는 l-번째 프레임의 ω 대역에서 계산된 Gain값으로, ω에서 측정된 음성 성분과 비음성 성분의 비율($\Lambda_l(\omega)$)에 의해 계산되며, 이는 다시 음성이 존재하지 않을 확률을 나타내는 $q_l(w)$에 의해 근사된다[8]. 즉, $1-q_l(w)$로 근사되는 Gain값은 ω 대역의 음성 존재 확률을 나타내며, 본 연구에서는 식 (4)의 $P_l(\omega)$를 계산하는데 이 값을 사용한다. $q_l(w)$는 사후 SNR로부터 계산되는 값으로 "참고문헌[7]"에 제시된 방법을 이용한다.

3.1.2 필터 계수의 개선

변형된 콤 필터의 주파수 응답($\hat{A}_l(\omega)$)을 역 푸리에 변환시켜 얻은 임펄스 응답으로부터 새로운 필터 계수($\widehat{\alpha_{l,k}}$)를 유도한다. 원래의 음원과 무관하게 window 함수에 의해 미리 정해진 값을 나타내는 기존의 필터 계수의 경우 모든 프레임에서 동일한 값이 적용되었던 것과 달리, $\widehat{\alpha_{l,k}}$는 l-번째 프레임에 포함된 음성과 잡음 성분의 특성을 반영함으로써, 잡음에 의해 잘못 측정된 피치 정보의 오류를 보정하는 효과가 있다.

0과 1 사이의 값을 갖는 $G_l(\omega)$을 식 (4)의 $P_l(\omega)$를 대신하여 콤 필터의 주파수 응답을 변형시킨 결과 모든 주파수 대역의 스펙트럼 크기는 감소하게 된다. 이 때 각 프레임마다 계산되는 $P_l(\omega)$의 값이 다르기 때문에 프레임마다 스펙트럼 크기가 감소하는 정도가 다르며, 이로 인해 프레임 간 에너지의 차이가 발생할 수 있다. 따라서 이 같은 에너지의 차이를 조정하는 과정이 필요하다. 본 연구에서는 이를 위해 모든 필터 계수의 합이 1 (즉, $\sum_{k=-L}^{L} \widehat{\alpha_{l,k}} = 1$)이 되도록 $\widehat{\alpha_{l,k}}$을 정규화 하는 과정을 최종적으로 수행한다. 이상의 과정을 통해 변형된 필터 계수 $\widetilde{\alpha_{l,k}}$를 식 (2)에 적용하면 다음과 같은 새로운 단위 샘플 응답 $\widehat{h(n)}$을 얻을 수 있다.

$$\widehat{h_l(n)} = \sum_{k=-L}^{L} \widetilde{\alpha_{l,k}} \times \delta_l(n-kT) \quad (6)$$

<그림 3>은 변형된 콤 필터의 주파수 응답 및 필터링 적용 후의 스펙트럼을 나타낸 것이다. 전체 스펙트럼 대역에서 기본 주파수에 따라 일정하게 반복되는 기존의 콤 필터(<그림 2>)와 달리 변형된 콤 필터는 음성 존재 확률에 따라 스펙트럼 특성이 재조정된 모습을 보이며, 이에 따라 음성 존재 확률이 높은 대역은 에너지가 상대적으로 강조되는 반면 음성이 존재하는 않는 대역의 에너지는 감쇄하는 결과를 나타낸다.

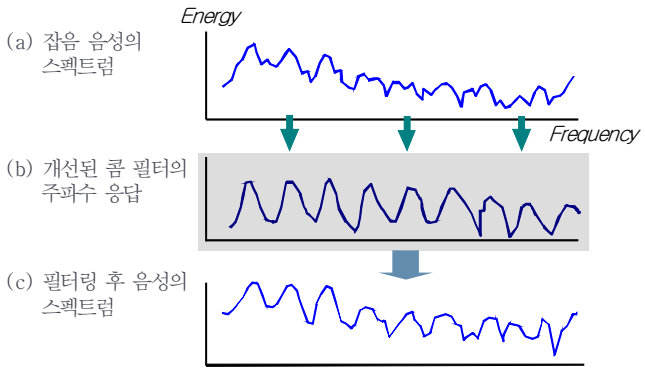

그림 3. 변형된 콤 필터의 주파수 응답(b) 및 필터링 전후의 스펙트럼(a,c)

Figure 3. Frequency response of modified comb filter (b) and spectrum of noisy speech (a) and de-noised speech (c)

3.2 특징 벡터 선별 기법을 이용한 프레임 선별

기존의 콤 필터링의 문제를 해결하기 위해 제안한 방법은 주파수 대역별로 추정한 음성 존재 확률을 이용하여 콤 필터의 주파수 응답을 원음성에 적합하게 조정하고 필터 계수를 개선하는 방법이며, 이 같은 필터링을 수행한 음성의 경우 피치 및 고조파 정보가 개선되는 결과를 얻는다. 이들 정보는 감정 인식에 유용한 특징 파라미터로써, 제안한 방법에 의해 개선된 특징 파라미터는 감정 인식 성능 향상에 크게 기여할 것으로 판단된다. 하지만, 콤 필터링은 피치 정보를 갖는 유성음에 대해서만 적용이 가능하므로 잡음 처리 후 무성음 프레임과 유성음 프레임 간의 에너지 차이가 발생할 수 있으며, 이는 에너지 정보가 유용하게 사용되는 감정 인식에서 중요하게 고려될 필요가 있다. 또한, 입력되는 잡음의 종류 및 크기(dB)가 지속적으로 바뀌는 경우, 콤 필터링 후에도 제거되지 않은 잡음 성분으로 인해 음성에 포함된 감정 정보가 훼손된 프레임이 존재할 수 있다. 본 연구에서는 이 같은 문제를 해결하기 위해 필터링 과정 후 감정 정보가 손실된 프레임을 선별한 뒤 해당 프레임을 인식 단계에서 제외하는 방법을 제안한다.

화자 식별 시스템의 성능 향상을 위해 특징 벡터 선별 기법이 연구된 바 있다[9]. 이 방법에 따르면, 화자들 사이에 발화 유사성이 존재하기 때문에 GMM 화자 모델 간 중첩이 발생하며, 훈련 데이터 가운데 중첩 영역에 포함되는 프레임들을 제거한 뒤 새로운 모델을 구축함으로써 모델 중첩에 의한 오인식 문제를 해결할 수 있다. 즉, 해당 화자의 특성이 뚜렷하게 나타나지 않는 프레임을 제외하고 그렇지 않은 프레임을 사용하여 새롭게 구축된 화자 모델은 특정 화자에 고유한 정보들로 표현된 모델이라는 사실에 기반한 방법이다. 이 방법을 감정 인식에 적용하면 심한 잡음으로 인해 감정 정보가 손실된 프레임을 선별하고 그렇지 않은 프레임, 즉 고유의 감정 정보를 지닌 프레임만으로 모델을 구축함으로써, 잡음에 강인한 감정 모델을 생

성하는데 기여할 것으로 판단된다. 따라서 본 연구에서는 기존의 특징 벡터 선별 기법을 감정 인식에 적용하는 방법을 제안한다.

먼저, 감정별로 분류된 훈련용 음성 자료에 대하여 각 프레임 별로 특징 파라미터를 추출함으로써 각 감정별 특징 벡터열 $X_e(=\{x_{e,1},...,x_{e,T}\})$을 얻는다. T는 e-번째 감정의 음성 자료에서 추출한 특징 벡터의 수이며, 감정 종류가 E가지인 경우, e는 1부터 E 사이의 값을 갖는다. 다음으로 각 감정의 특징 벡터를 이용하여 감정마다 GMM (λ_e; $e=1,\ldots,E$)을 구축한다. 모델 훈련에 사용된 각각의 특징 벡터 $x_{e,t}$를 각 GMM에 적용하여 로그-우도를 계산한 다음, $x_{e,t}$가 자신에 해당하는 감정 모델인 λ_e에서 최대 우도를 보이는 경우 이 벡터를 '비중첩 벡터 (non-overlapped vector)'로 분류한다. 반대로 해당 벡터의 감정이 아닌 다른 감정 모델에서 최대 우도를 보이는 경우 이 벡터를 '중첩 벡터 (overlapped vector)'로 분류한다. 비중첩 벡터는 본래의 감정 정보를 제대로 포함하고 있는 벡터를 뜻하며, 중첩 벡터는 다른 감정 정보의 특성을 지니고 있거나 심한 잡음에 의해 본래의 감정 특성이 훼손된 벡터를 의미한다.

특징 벡터 선별 과정을 요약하면 다음과 같다.

- $x_{e,t}$: 감정 e의 GMM (λ_e)을 구축하는데 사용된 t-번째 특징 벡터 ($e=1,...,E,\ t=1,...,T$)

i) $\hat{i}=\arg\max_i log\,P(x_{e,t}|\lambda_i)$, $(i=1,...,E)$

ii) 만약 $\hat{i}$가 e라면, $x_{e,t}$를 비중첩 벡터로 분류

iii) $\hat{i}$가 e가 아니라면, $x_{e,t}$를 중첩 벡터로 본류

특징 벡터 선별이 완료되면 비중첩 벡터로 분류된 벡터들을 대상으로 새로운 GMM $\hat{\lambda}_e$을 구축하고, 중첩 벡터들을 대상으로 '중첩 모델 (overlap GMM)'을 구축한다. <그림 4>는 특징 벡터 선별 기법에 의한 감정 모델 구축 과정을 나타낸 것이다.

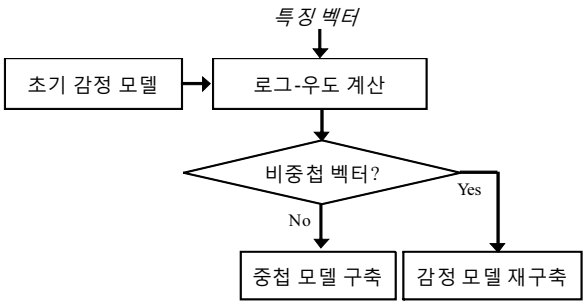

그림 4. 특징 벡터 선별 기법 기반의 감정 모델 구축 과정
Figure 4. Emotion model training based on feature vector classification

3.3 특징 벡터 선별 기반의 감정 인식

특징 벡터 선별 과정을 통해 각 감정마다 구축된 두 개의 모델(즉, 개선된 GMM과 중첩 모델)을 이용하여 입력 음성에 대한 인식 과정을 수행한다. 입력 음성 역시 심한 잡음에 의해 감정 정보가 손실된 프레임이 존재할 수 있으며 이 같은 프레임이 인식에 사용된다면 오인식을 야기할 수 있으므로, 입력 음성에 대해서도 특징 벡터 선별 과정이 필요하다. 본 연구에서는 이 과정에서 중첩 모델을 이용하며, 중첩 모델은 일종의 garbage 모델로서 기능한다. 즉, 입력 음성의 각 특징 벡터를 감정별 GMM $\hat{\lambda}_e$와 중첩 모델에 적용하여 로그-우도를 계산한 다음 중첩 모델에서 최대 우도를 보이는 경우 해당 특징 벡터를 '중첩 벡터'로 분류하여 인식 과정에서는 제외시킨다. 반대로 감정별 GMM에서 최대 우도를 보이는 벡터를 대상으로 다음과 같은 과정을 통해 최종 인식 결과($\hat{e}$)를 얻는다.

$$\hat{e}=\arg\max_e log\,P(D|\hat{\lambda}_e),\ e=1,...,E \tag{7}$$

$$\log P(D|\hat{\lambda}_e)=\frac{1}{F}\sum_{f=1}^{F}\log P(d_f|\hat{\lambda}_e) \tag{8}$$

D는 입력 음성의 특징 벡터 중 비중첩 벡터로 분류된 벡터들의 집합이며, F는 이 집합에 속한 벡터의 수를 나타낸다. <그림 5>는 잡음 음성을 대상으로 개선된 콤 필터링을 적용한 후 추출한 특징 벡터에 대하여 훈련 단계(<그림 4>)에서 재구축된 감정 모델과 중첩 모델을 이용하여 감정 인식을 수행하는 과정을 나타낸 것이다.

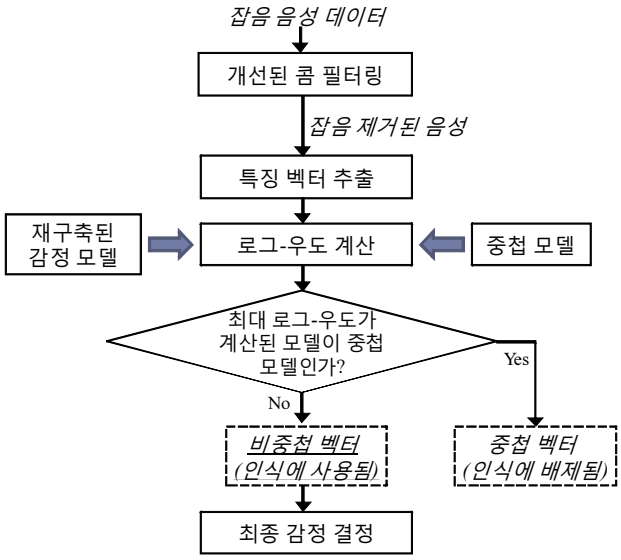

그림 5. 제안한 특징 벡터 처리 기반의 감정 인식 과정
Figure 5. Emotion recognition based on proposed feature vector processing

4. 실험 및 결과

본 연구에서 제안한 특징 벡터 처리 기법의 유효성을 검증하기 위하여 잡음 자료 및 감정 음성 자료를 사용하여 감정 인식 실험을 수행하였다. 제안한 콤 필터링의 성능을 평가하기 위해 기존의 콤 필터링 및 주파수 차감법([10])을 적용한 음성을 사용하여 감정 인식 성능을 비교하였다. 또한 특징 벡터 선별 기법을 추가로 적용하여 프레임 선별이 인식 성능 향상에 미치는 영향을 조사하였다.

다양한 잡음 제거 기법 가운데 주파수 차감법을 성능 비교의 대상으로 선정한 이유는 다음과 같다. 첫째, 묵음 구간을 검출한 후 잡음 성분을 추정할 필요가 없는 콤 필터링과 달리 묵음 구간 검출 및 잡음 성분 추정 과정이 요구되는 대표적인 잡음 제거 기법으로서 주파수 차감법을 선정하였다. 둘째, 감정 인식에 가장 유용하게 사용되는 고조파 정보 또는 기본 주파수(즉, 피치 주기)를 잡음 음성으로부터 정확히 추정하는 것은 한계가 있으며, 본 연구는 잡음 음성에서 고조파 정보를 개선하기 위한 방법으로서 콤 필터링의 개선을 제안하였다. 따라서 기본 주파수의 추정과 무관한 기존의 잡음 제거 기법 중 가장 널리 알려진 주파수 차감법을 성능 평가에 사용하였다.

4.1 실험 환경

본 연구에서는 LDC의 감정 음성 코퍼스('Emotional Prosody Speech and Transcripts')를 이용하여 감정 인식 실험을 수행하였다[11]. 이 음성 자료는 잡음이 없는 실험실에서 녹음된 것으로, 잡음 환경에서의 감정 인식 성능을 평가하기 위해 NoiseX-92에서 추출한 백색 잡음과 군중 잡음, 공장 잡음을 LDC의 감정 음성 자료에 추가하였다[12]. LDC 감정 음성 자료는 일곱 명의 전문 배우들이 날짜 및 숫자의 조합으로 구성된 어휘를 총 15개의 감정 종류별로 연기하여 녹음된 자료로, 한 화자 당 각 감정마다 평균적으로 20개의 어휘가 녹음되었다. 감정 인식 결과의 신뢰도를 향상시키기 위하여 7-fold cross-validation 기법을 사용하여 성능을 평가하였다. 즉, 한 화자의 음성 자료를 실험 자료로 사용하고 나머지 여섯 화자의 음성 자료를 학습 자료로 사용하였으며, 실험 자료로 사용되는 화자를 순차적으로 교체함으로써 총 7차례의 성능 평가를 시행하였다. 인식에 사용한 특징 파라미터는 피치 주기, 로그 에너지 및 영점 교차율 그리고 12차 MFCC이며, 40ms 단위의 프레임을 대상으로 특징 파라미터를 추출하였다. 지나치게 짧은 구간에서는 정확한 피치 주기를 측정하기 어려우므로 프레임의 길이를 40ms로 정하였으며, 이 같은 길이는 "참고문헌[13]" 등에서 사용된 바 있다. <표 1>은 실험에 사용한 감정의 종류를 나타낸 것이다. 부정적 감정을 대표하는 'anger'와 평상시 감정의 'neutral'은 2-클래스 감정 인식을 위해 사용되며, 긍정적 감정을 대표하는 'happy'가 포함되어 3-클래스를 구성한다. 또한 'boredom'과 'sadness' 감정이 추가되어 다섯 가지 감정을 인식하는데 사용되며, 이 같은 조합은 5-클래스 감정 인식에서 주로 사용되어 왔다[2],[5],[13]. GMM 기반의 식별 기법을 인식에 사용했으며, 제안한 특징 벡터 처리 기법의 유효성을 중점적으로 평가하기 위해 mixture 수는 1로 제한을 두고 감정 모델을 구축하였다.

표 1. 감정 인식 실험에 사용된 감정 종류
Table 1. Emotion types according to the number of classification categories

감정 수	감정 종류
2-클래스	anger, neutral
3-클래스	anger, neutral, happy
5-클래스	anger, neutral, happy, boredom, sadness

4.2 실험 결과

제안한 콤 필터링에 의한 잡음 제거 및 피치 주기의 개선에 대한 유효성을 평가하기 위하여, 특징 벡터 선별 기법을 적용하지 않은, 즉 잡음 처리의 결과 음성으로부터 모델 학습 및 인식 실험을 수행하였다. 잡음이 추가되지 않은 실험실 환경의 LDC 음성 자료를 사용하여 평가한 5-클래스 감정에 대한 인식률은 53.3%로, 동일한 음성 자료 및 식별 방법에 의해 평가된 기존의 연구 결과([5],[13])와 비슷하거나 다소 높은 성능을 보였다. 이와 반대로 잡음의 영향으로 인식률이 크게 저하되었음이 확인되었다. <그림 6>과 <그림 7>은 실험에 사용한 세 가지 잡음 환경(백색 잡음, 군중 잡음, 공장 잡음)에서 5-클래스 감정 자료를 대상으로 수행한 인식 실험의 결과이다. <그림 6>은 신호대 잡음비(Signal to Noise Ratio) 즉 SNR이 10dB인 경우, <그림 7>은 5dB인 경우의 결과를 나타낸다. 잡음 제거 기법을 수행하지 않은 음성(Baseline)의 성능이 가장 좋지 않았으며, 군중 잡음 및 공장 잡음 환경의 경우 주파수 차감법이 적용된 음성(Spectral Subtraction; SS), 기존의 콤 필터링이 적용된 음성(Conventional Comb Filtering; CCF), 그리고 개선된 콤 필터링이 적용된 음성(Advanced Comb Filtering; ACF) 순으로 인식률이 향상됨을 보였다. 하지만, 백색 잡음 환경의 결과에서는 주파수 차감법이 콤 필터링보다 높은 성능을 나타냈다. 주파수 차감법은 비음성 구간에서 추정한 각 주파수 대역의 잡음 에너지를 음성 구간에서 차감하는 기법으로, 전체 스펙트럼 대역에 잡음 에너지가 일정하게 분포한 백색 잡음을 제거하는 데 효과적이다. 반면, 군중 잡음이나 공장 잡음과 같은 비정적 잡음 환경에 대해서는 제안한 콤 필터링 기법을 적용한 후 성능이 크게 개선되었다. 주파수 차감법의 경우 묵음 구간에서 추정한 잡음 성분이 음성 구간에 포함된 잡음 성분과 상이할 때 잡음 성분이

정확하게 차감되지 않는 문제점이 발생하는 반면, 제안한 콤 필터링의 경우 음성 구간에서 직접 추정한 정보를 사용하므로 비정적 잡음에 대해 성능이 크게 향상된 것으로 파악된다. 두 종류의 잡음 수준(10dB과 5dB)에서 평가된 제안한 콤 필터링 기법의 평균 인식 성능은 37.1%를 보였으며, 이는 Baseline에 비해 5.3%, 그리고 SS와 CCF에 비해 각각 3.7%와 3.2% 성능이 개선되었음을 나타낸다.

<표 2>는 감정 수에 따른 인식률의 변화를 나타낸 것이다. 10dB SNR의 잡음 자료(백색, 군중, 공장 잡음)를 대상으로 실험한 결과이며 성능 비교의 대상으로 사용된 방법은 <그림 6>,<그림 7>에서와 동일하다. 감정의 종류가 증가할수록 인식률이 저하됨을 확인할 수 있다. 앞서 살펴본 5-클래스의 결과처럼 2-클래스 및 3-클래스의 경우 역시 잡음이 추가되지 않은 음성 자료의 성능과 비교했을 때 10dB 수준의 잡음 환경에서 각각 15%와 13%의 성능 저하를 나타냈다. 잡음이 포함된 음성(Baseline)은 주파수 차감법, 기존의 콤 필터링, 제안한 콤 필터링에 의해 평균적으로 각각 2%, 3%, 5.7%의 에러감소율 (Error rate reduction)을 보였다. 특히 표에서는 확인되지 않지만, 백색 잡음을 제외한 군중 잡음과 공장 잡음만을 대상으로 평가했을 경우 제안한 콤 필터링은 약 10%의 성능 개선을 나타냈다.

제안한 콤 필터링에 의해 피치 주기가 효과적으로 개선되는지의 여부를 확인하기 위하여 피치 주기를 포함한 특징 파라미터의 종류를 변화시키며 인식 성능을 조사하였다. <그림 8>은 0dB, 5dB, 10dB의 잡음 수준에서 평가한 5-클래스 감정 인식의 평균 인식률로, 감정 인식에 사용한 특징 파라미터의 종류에 따른 인식률의 변화를 나타낸다. 특징 파라미터로 MFCC 정보만을 사용한 경우 인식률이 거의 변화하지 않는 반면, 피치 주기가 포함된 경우 제안한 방법에 의해 성능이 눈에 띄게 향상됨을 보였으며, 이 같은 결과는 제안한 콤 필터링에 의해 피치 정보가 효과적으로 개선되었음을 나타낸다.

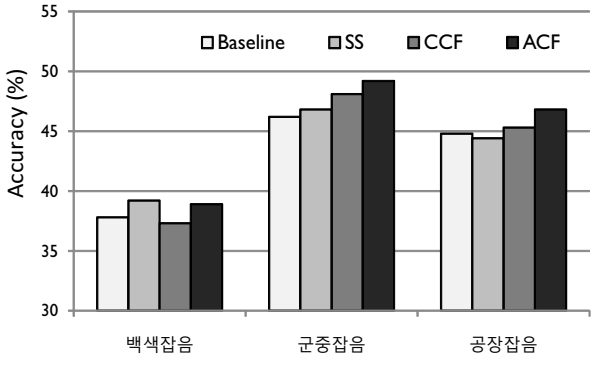

그림 6. 잡음 종류에 따른 인식률 변화
(5-클래스, 10dB 잡음환경)
Figure 6. Recognition accuracy according to the noise types (5-class, 10dB SNR)

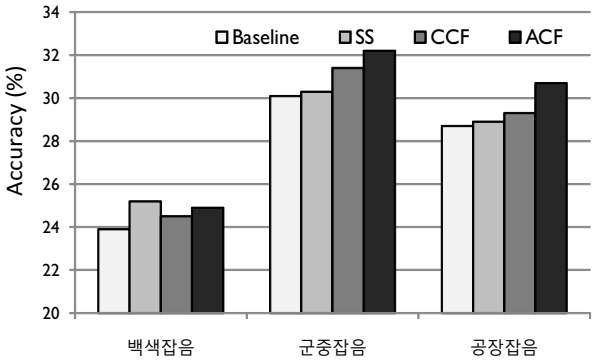

그림 7. 잡음 종류에 따른 인식률 변화
(5-클래스, 5dB 잡음환경)
Figure 7. Recognition accuracy according to the noise types (5-class, 5dB SNR)

표 2. 감정 수에 따른 인식률(%) 변화 (10dB 잡음환경)
Table 2. Recognition accuracy (%) according to the number of emotions (10dB SNR)

	2-클래스	3-클래스	5-클래스	평균	에러감소율
Baseline	78.6	57.6	42.9	59.7	-
SS	79.7	58.2	43.5	60.5	2.0
CCF	80.1	58.9	43.6	60.9	3.0
ACF	81.3	59.7	45.0	62.0	5.7

지금까지 살펴본 인식 성능을 통해 제안한 콤 필터링이 비정적 잡음 환경에서 감정 인식의 성능 개선에 효과적임을 확인하였다. 본 연구에서는 개선된 콤 필터링을 수행한 음성에 대하여 특징 벡터 선별 기법을 통해 강인한 감정 모델을 구축하는 방법을 함께 제안하였으며, 이와 관련된 실험 결과는 <표 3>과 같다. 세 가지 잡음 수준(clean, 10dB, 5dB)에 대하여 5-클래스의 감정 인식 실험을 수행하였으며, 성능 평가를 위해 가공되지 않은 본래 음성(Baseline), 개선된 콤 필터링을 적용한 음성(ACF), 그리고 이 음성에 대하여 특징 벡터 선별 기법(Feature Vector Classification; FVC)에 기반한 모델 학습을 적용한 후(ACF+FVC)의 인식 성능을 비교하였다. 실험 결과 실험실 환경(Clean)의 경우 콤 필터링을 적용한 후 인식 성능이 오히려 저하됨을 확인하였는데, 이는 콤 필터링에 의해 음성 신호에 왜곡이 발생하였기 때문인 것으로 판단된다. 하지만, 'ACF+FVC'의 경우 오히려 Baseline보다 성능이 향상되는 결과를 보였는데 이는 특징 벡터 선별 기법에 의해 모호한 감정 특성을 지니는 특징 벡터가 제외되고 고유 감정 특성을 지닌 벡터만으로 구축된 감정 모델의 결과에 기인한 것으로 판단된다. 10dB 및 5dB의 잡음 환경 역시 콤 필터링과 특징 벡터 선별 기법을 함께 적용함으로써 성능이 향상됨을 보였다. 잡음 환경(10dB 및 5dB)에서 평가한 평균 인식 성능은 baseline에 비해 5%의 에러감소율을 보였으며, 표에서는 확인되지 않으나 백색 잡음을 제외한 군

중 잡음과 공장 잡음만을 대상으로 평가한 경우 약 13%의 향상률을 보였다. 이 같은 결과는 특징 벡터 선별 기법이 잡음 환경에서의 감정 인식에 유용하게 적용될 수 있음을 뜻한다. 즉, 특징 벡터 선별 기법을 통해 심한 잡음에 의해 변이가 발생한 특징 벡터 및 모호한 감정 특성을 보이는 벡터를 효과적으로 제거함으로써 보다 강인한 감정 모델을 구축하는데 기여하는 것으로 판단된다.

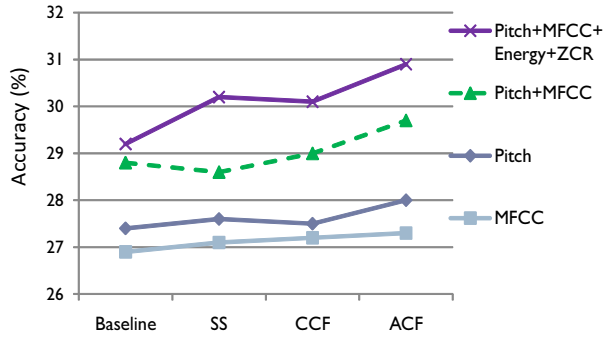

그림 8. 특징 파라미터에 따른 인식률 변화 (5-클래스, 0~10dB 잡음 환경)

Figure 8. Recognition accuracy according to feature parameters (5-class, 0~10dB SNR)

표 3. 특징 벡터 선별 기법에 의한 인식률(%) 변화 (5-클래스)

Table 3. Performance (%) of feature vector classification (5-class)

	Clean	10dB SNR	5dB SNR	평균	에러감소율
Baseline	53.3	42.9	27.6	41.3	-
ACF	52.9	45.0	29.3	42.4	1.9
ACF+FVC	54.3	45.8	31.2	43.8	4.3

5. 결론

본 연구에서는 잡음 환경에서 감정 인식 성능을 개선하기 위한 특징 벡터 처리 방법을 제안하였다. 제안한 방법은 콤 필터링을 기반으로 잡음 제거를 수행하며, 잡음 제거 후 추출한 특징 벡터를 대상으로 특징 벡터 선별 기법을 적용하여 잡음에 강인한 감정 모델을 구축한다. 심한 잡음으로 인해 잘못 측정된 피치에 의해 발생하는 콤 필터링의 문제점을 해결하기 위하여 음성 존재 확률을 이용하여 콤 필터의 주파수 응답 및 필터 계수를 개선하였다. 또한 잡음에 의해 감정 정보가 훼손된 프레임을 선별하기 위해 특징 벡터 선별 기법을 적용하였으며, 고유의 감정 특성을 지닌 특징 벡터로부터 감정 모델을 구축함으로써 기존의 감정 모델을 개선하였다. LDC의 감정 음성 자료 및 NoiseX-92의 잡음 자료를 이용하여 감정 인식 실험을 수행한 결과, 제안한 콤 필터링은 주파수 차감법 및 기존의 콤 필터링에 비하여 향상된 성능을 나타냈다. 또한 특징 벡터 선별 기법을 함께 적용한 결과 baseline에 비해 4.3%의 에러감소율을 보였다. 향후 연구에서는 문장 단위의 감정 음성 자료에 대하여 제안한 방법의 유효성을 검증하며, 특징 벡터 선별 기법을 HMM 기반의 시스템에 적용하고자 한다. 또한 위너 필터링, MMSE-LSA 등 다양한 잡음 제거 기법과의 성능 비교를 통해 제안한 방법의 유효성을 검증하고자 한다.

참고문헌

[1] Ververidis, D., Kotropoulos, C. (2006). “Emotional speech recognition: resources, features, and methods”, *Speech Communication*, Vol. 48, No. 9, pp. 1162-1181, Sep.

[2] Kwon, O., Chan, K., Hao, J., Lee, T. (2003). “Emotion recognition by speech signals”, *Proc. Int. Conf. on Eurospeech*, pp. 125-128.

[3] Tato, R., Santos, R., Kompe, R., Pardo, J. (2002). “Emotional space improves emotion recognition”, *Proc. Int. Conf. on ICSLP*, pp. 2029-2032.

[4] Jeong, Y.J. (2009). “Performance comparison of the speech enhancement methods for noisy speech recognition”, *Phonetics and Speech Sciences*, Vol. 1, No. 2, pp. 9-14. Jun.
(정용주, (2009). “잡음음성인식을 위한 음성개선 방식들의 성능 비교”, 말소리와 음성과학, 제1권, 제2호, pp. 9-14.)

[5] Huang, R., Ma, C. (2006). “Toward a speaker-independent real time affect detection system”, *Proc. Int. Conf. on Pattern Recognition (ICPR)*, pp. 1204-1207.

[6] Frazier, R.H., Samsam, S. (1976). “Enhancement of speech by adaptive filtering”, *Proc. Int. Conf. on ASSP*, pp. 251-253.

[7] Park, J.S., Oh, Y.H. (2006). “Noise reduction using MMSE estimator-based adaptive comb filtering”, *Malsori*, pp. 181-190, Dec.
(박정식, 오영환, (2006). “MMSE Estimator 기반의 적응 콤 필터링을 이용한 잡음 제거”, 말소리, pp. 181-190.)

[8] Malah, D., Cox, R.V., Accardi, A.J. (1999). “Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments”, *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 201-204, Mar.

[9] Kwon, S., Narayanan, S. (2007). “Robust speaker identification based on selective use of feature vectors”, *Pattern Recognition Letters*, Vol. 28, pp. 85-89.

[10] Boll, S.F. (1979). “Suppression of acoustic noise in speech using spectral subtraction”, *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 27, No. 2, pp. 113-120.

[11] Liberman, M., Davis, K., Grossman, M., Martey, N., Bell, J. (2002). *Emotional prosody speech and transcripts*, Linguistic

Data Consortium (LDC), University of Pennsylvania, PA, USA, Jul.

[12] Varga, A.P., Steenken, H.J.M, Tomlinson, M., Jones, D. (1992). “The NOISEX-92 study on the effect of additive noise on automatic speech recognition”, *Technical Report*, DRA Speech Research Unit.

[13] Sethu, V., Ambikairajah, E., Epps, J. (2007). “Speaker normalization for speech-based emotion detection”, *Proc. Int. Conf. on Digital Signal Processing (DSP)*, pp. 611-614.

• **박정식 (Park, Jeongsik)**
한국과학기술원 전산학과
대전시 유성구 구성동 373-1
Tel: 042-350-3556 Fax: 042-350-3556
Email: dionpark@speech.kaist.ac.kr
관심분야: 음성인식, 감정인식, 잡음처리
현재 전산학과 대학원 박사과정 재학중

• **오영환 (Oh, Yunghwan)** 교신저자
한국과학기술원 전산학과
대전시 유성구 구성동 373-1
Tel: 042-350-3516 Fax: 042-350-3556
Email: yhoh@speech.kaist.ac.kr
관심분야: 음성인식, 음성합성, 음성코딩
1985~현재 전산학과 교수