MMSE Estimator 기반의 적응 콤 필터링을 이용한 잡음 제거*

박정식(KAIST), 오영환(KAIST)

<차 례>

<Abstract>

Noise Reduction Using MMSE Estimator-based Adaptive Comb Filtering

Jeong-Sik Park, Yung-Hwan Oh

This paper describes a speech enhancement scheme that leads to significant improvements in recognition performance when used in the ASR front-end. The proposed approach is based on adaptive comb filtering and an MMSE-related parameter estimator. While adaptive comb filtering reduces noise components remarkably, it is rarely effective in reducing non-stationary noises. Furthermore, due to the uniformly distributed frequency response of the comb-filter, it can cause serious distortion to clean speech signals. This paper proposes an improved comb-filter that adjusts its spectral magnitude to the original speech, based on the speech absence probability and the gain modification function. In addition, we introduce the modified comb filtering-based speech enhancement scheme for ASR in mobile environments. Evaluation experiments carried out using the Aurora 2 database demonstrate that the proposed method outperforms conventional adaptive comb filtering techniques in both clean and noisy environments.

* Keywords: Robust speech recognition, Adaptive comb filtering, Gain modification function, Mobile communication environment.

* 본 연구는 과학기술부의 지원을 받아 2006년도 국가지정연구실 사업을 통해 수행되었음.

1. 서 론

음성인식 시스템의 입력 음성에 포함되는 채널 잡음 및 가산 잡음을 효과적으로 처리함으로써 인식 성능을 향상시키는 다양한 연구들이 오래 전부터 수행되어 왔다. 최근 무선 통신 환경에서의 음성 인식 시스템, 즉 DSR(Distributed Speech Recognition) 시스템의 성능 향상을 위한 잡음 처리 기법이 활발히 연구되고 있으며 다양한 잡음 처리 기법을 DSR 시스템에 적용함으로써 높은 성능을 나타냈다[1][2]. 그러나 다양한 변이를 나타내는 불안정한 잡음(non-stationary noise)에 대한 처리는 여전히 과제로 남아 있다.

본 연구에서는 QCELP 음성 코더 기반인 CDMA 단말기의 ASR에서의 잡음 제거를 목적으로 유성음의 기본 주파수(또는 피치 주기)를 이용하는 적응 콤 필터링 기법을 적용한다. Frazier에 의해 처음 소개된 적응 콤 필터링은 계산 부담이 크지 않음에도 효과적으로 잡음을 제거하는 방법으로 알려져 있으며[3], 음질 개선 및 잡음 제거 기법으로 적용되어 왔다[4][5]. 그러나 심한 잡음으로 인해 피치 주기 측정이 어려운 음성이나 잡음의 변이가 심한 음성에 대해서는 적용의 어려움이 있다. 본 논문에서는 기존의 적응 콤 필터링의 문제점을 살펴보고 이를 개선함으로써 인식 성능을 향상시키는 방법을 소개하며, 개선된 콤 필터를 CDMA 단말기 및 원음 음성에 적용함으로써 제안한 방법의 유효성을 검증한다.

본 논문의 구성은 다음과 같다. 2장에서 적응 콤 필터링의 특성을 살펴본 후 개선점을 밝히고 3장에서 콤 필터의 개선 방법을 설명한다. 4장에서 실험 환경 및 결과를 제시하며 5장에서 결론을 맺는다.

2. 적응 콤 필터링의 특성

유성음의 파형은 기본 주파수에 따라 주기적이라는 특성에 기반을 둔 적응 콤 필터링은 음성 신호의 고조파 성분(harmonics)을 강조함으로써 본래의 음성 신호를 보존하고 고조파 성분들 사이의 주파수 대역의 에너지를 감소시킴으로써 잡음 신호를 제거하는 음질 개선 기법이다[3].

적응 콤 필터링은 다음과 같은 임펄스 응답 $h(n)$을 통해 처리된다.

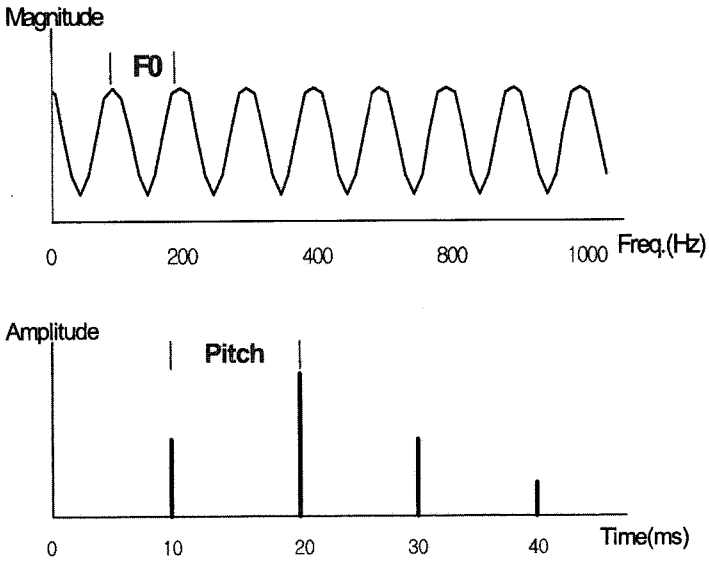
$$h(n) = \sum_{k=-L}^{L} a_k \times \delta(n - T_k) \tag{1}$$

$\delta(n)$은 단위 임펄스 함수이며 T_k는 피치 주기, 2L+1은 필터의 길이를 의미한다. a_k는 필터 계수로서 식 (2)와 같은 Hamming window 형태를 따른다.

$$a_k = \frac{0.54 + 0.46cos(2\pi k/2L+1)}{\sum_{k=-L}^{L}\{0.54 + 0.46cos(2\pi k/2L+1)\}} \quad (2)$$

이와 같은 콤 필터는 단위 프레임의 기본 주파수에 따라 일정한 피치 주기를 갖는 임펄스 응답 특성을 보인다. <그림 1>은 피치 주기가 10ms인 콤 필터의 임펄스 응답과 주파수 응답(기본 주파수 100Hz)을 나타낸 것이다. 시간 영역에서는 피치 주기마다 임펄스가 나타나며 주파수 영역에서는 기본 주파수의 배수, 즉 고조파마다 피크가 존재한다.

이 같은 콤 필터를 잡음 음성에 적용함으로써 고조파는 강조되고 고조파 사이에 존재하는 잡음 성분의 에너지가 감소되는 효과가 있다. 하지만 전체 스펙트럼 대역에서 주파수 응답이 일정하게 반복되는 콤 필터를 심한 잡음 음성에 적용하면 오히려 잡음 성분이 강조되고 음성이 존재하는 주파수 대역의 에너지가 감소하는 문제가 발생할 수 있다. 반면 깨끗한 음성에 적용하였을 경우 신호의 왜곡이 발생하며 이는 인식 성능 저하의 요인이 된다. 깨끗한 음성을 대상으로 콤 필터링을 처리한 음성이 그렇지 않은 음성에 비해 인식 성능이 저하되는 실험 결과가 이를 입증한다[4]. 콤 필터링은 피치 주기를 갖는 유성음에 대해서만 적용되어야 하므로, 필터링의 선행 과정인 유/무성음 선별 시 발생하는 오류 또한 왜곡을 일으킨다.



<그림 1> 콤 필터의 주파수 응답(上)과 임펄스 응답(下)

3. 적응 콤 필터링의 개선

본 장에서는 Gain Modification Function (GMF)를 사용하여 콤 필터에 의해 발생하는 왜곡을 주파수 영역에서 보상하는 방법을 제안한다. GMF는 효과적인 잡음 처리 기법으로 알려진 MMSE-LSA (Minimum Mean Squared Error-Log Spectral Amplitude) 추정 기법에서 잡음이 첨가되기 전의 음성을 복원하기 위해 사용하는 estimator로서 주파수 대역에서 음성과 잡음 성분의 비율을 나타내는 값이다. 음성 성분이 존재하지 않을 확률(Speech Absence Probability (SAP))과 priori SNR 및 posteriori SNR에 의해 계산되며, SAP가 작고 SNR이 큰 주파수 영역일수록 GMF는 큰 값을 갖는다[6]. 본 논문에서는 각 주파수 대역마다 잡음과 음성 성분의 분포 정도를 정량적으로 분석하는 GMF를 콤 필터에 반영함으로써 콤 필터의 주파수 응답 특성을 보다 유연하게 변화시키는 방법을 제안한다.

3.1. 음성 존재 확률의 계산

GMF에 적용되는 SAP는 식 (3)의 $q(w)$ 함수에 의해 계산된다[7].

$$q_l(w) = \alpha q_{l-1}(w) + (1-\alpha) I_l(w) \tag{3}$$

$I_l(w)$는 l번째 프레임의 주파수 대역 ω의 posterior SNR (γ_w)로부터 이 대역에서의 음성 성분의 존재 여부를 결정하는 파라미터로 다음과 같이 0또는 1이 할당된다.

$$I_l(w) = \begin{matrix} 0\ (\gamma_w > \gamma_{TH}) \\ 1\ (\gamma_w < \gamma_{TH}) \end{matrix} \tag{4}$$

가령, γ_w이 γ_{TH}보다 큰 경우 $I_l(w)$=0이 되어 $q_l(w)$은 작아지며 이는 이 대역에 음성 성분이 존재할 확률이 높음을 뜻한다. 다시 말해, $q_l(w)$이 커질수록 대역 ω에서의 SAP는 큰 값을 나타낸다. α와 임계값 γ_{TH}은 기존의 방식([7])에서 고정된 상수(각각, 0.95, 0.8)를 사용하는데, 주파수 대역별 SNR은 입력 음성에 따라 변이가 크며 심한 잡음 음성의 경우 음성이 존재하는 구간에서 $I_l(w)$=1이 될 우려가 있으므로 본 연구에서는 아래 식과 같이 대역에 따라 임계값을 지속적으로 업데이트하였다.

$$\gamma_{TH}^{l}(w) = \sum_{k=l-\beta}^{l-1} \frac{\gamma_w^k}{\beta} \tag{5}$$

고정된 임계값 대신 β개의 프레임에서 측정된 γ_w의 평균을 반영함으로써 음성 대역 및 묵음 대역의 선별을 개선하였으며, 최소의 프레임 수는 실험에 의해 5개로 결정하였다. α 값의 경우 [7]에서 사용된 값(0.95)이 최적임을 실험적으로 확인하였다.

3.2. GMF에 의한 콤 필터의 개선

제안한 방법의 목표는 음성 존재 확률을 뜻하는 GMF를 통해 콤 필터의 주파수 응답을 원래 음성의 스펙트럼에 가깝도록 조정하는 것이다. GMF ($G_l(w)$)는 다음과 같은 식에 의해 계산된다[1].

$$G_l(w) = \frac{\Lambda_l(w)}{1+\Lambda_l(w)} \tag{6}$$

$\Lambda_l(w)$는 w에서의 음성 존재/비존재 값을 나타내는 비율이며, 식 (6)은 이 값이 클수록 음성이 존재할 확률이 높음을 뜻한다. GMF를 적용하여 콤 필터의 주파수 대역 w의 에너지($A(w)$)를 조정하는 방식은 다음과 같다.

$$\hat{A}_l(w) = G_l(w) \times A_l(w) \tag{7}$$

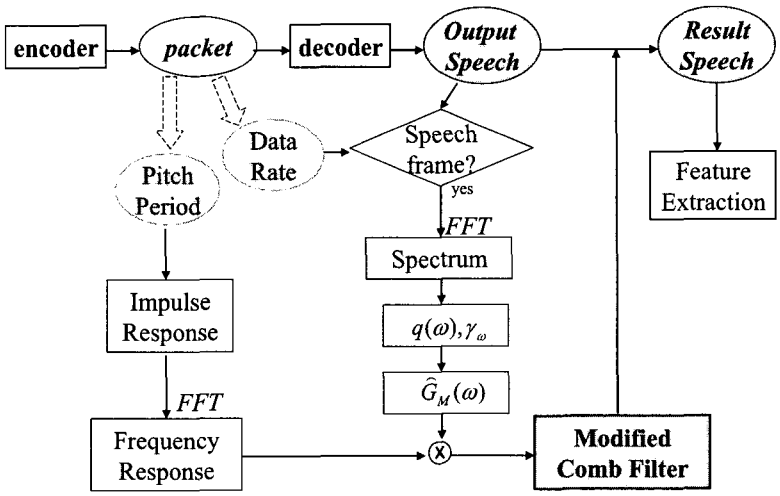
SAP의 값이 큰 대역, 즉 음성 존재 확률이 작은 대역에서의 GMF는 0에 가까워지며 식 (7)에 의해 해당 대역의 콤 필터 에너지가 감소된다. 반면, 음성 존재 확률이 큰 대역에서 GMF는 1에 가까워지므로 콤 필터의 에너지는 변함없이 지속된다. 그러나 0과 1사이의 범위를 갖는 GMF를 그대로 적용하는 경우 전체 스펙트럼 대역에서 에너지가 감소되어 신호 왜곡이 발생할 우려가 있으므로, $G_l(w)$ 대신 식 (8)과 같이 수정된 값을 반영한다.

$$\hat{G}_l(w) = \{1-(q_l(w)-\epsilon)\} \times G_l(w) \tag{8}$$

앞서 설명된 바와 같이 $q(w)$는 음성 대역($q(w)$<0.5) 또는 묵음 대역($q(w)$>0.5)임을 결정하는 확률 값이다. 따라서 식 (8)은 $q(w)$의 값이 ϵ보다 큰 주파수 대역에서는 1 미만의 GMF를 적용하여 필터의 스펙트럼 에너지를 감소시키며, ϵ보다 작은 주파수 대역에서는 1 이상의 값을 적용하여 에너지를 지속하거나 강조하는 의미를 갖는다. 실험에 의해 ϵ이 0.35일 때 최고의 성능을 보였다.

3.3. CDMA 단말기 ASR 전처리에서의 적용

개선된 콤 필터는 QCELP 음성 코더[1] 기반인 CDMA 단말기의 ASR 전처리에 유용하게 적용될 수 있다. QCELP 코더의 부호화/복호화 과정을 거친 출력 음성의 경우 코더에 의한 묵음 구간의 에너지 저하로 입력 음성에 비해 인식 성능이 향상되며, QCELP 코더에서 측정된 피치 주기를 출력 음성의 콤 필터링에 적용함으로써 더욱 개선된 성능을 보였다[4]. 개선된 콤 필터 역시 CDMA 단말기의 ASR 전처리에 효과적으로 적용될 수 있으며 그 과정은 <그림 2>와 같다. <그림 2>는 QCELP 코더의 출력 음성을 대상으로 개선된 콤 필터를 적용하여 잡음을 제거한 후 인식에 사용될 특징 파라미터를 추출하는 과정을 나타낸다. 개선된 콤 필터는 각 프레임의 피치 주기를 이용하여 3.2절에서 설명된 것과 동일한 과정에 의해 구성된다.



<그림 2> 개선된 콤 필터의 적용

이 과정에서 음성 코더에서 측정된 파라미터가 유용하게 사용된다. 3.1절에서 설명된 임계값의 지속적인 업데이트는 묵음 구간에 대해 적용할 필요가 없으므로, 음성 구간을 선별하기 위해 Data Rate 정보를 이용하였다[2]. 즉, Data Rate이 3이상인 프레임에 대해서만 임계값을 업데이트한다. 콤 필터에 필요한 프레임의 피치 주기 또한 코더의 패킷에서 직접 추출한 정보를 사용하였다. QCELP에서 측정되는 피치 주기는 음성 구간의 경우 한 프레임에 대해 최대 네 차례 업데이트되므로 잡음이 심한 음성에 대해서도 비교적 정확한 값을 얻을 수 있으며, 피치 측정에

1) 본 논문에서는 IS-96A (variable 8kbps QCELP) 코더를 사용하였다.
2) 'Data Rate' 정보는 speech activity에 따른 각 프레임의 음성/묵음 정보를 의미하며 1(묵음 프레임)부터 4(음성 프레임)의 정수로 표시된다[8].

소요되는 시간을 절감하는 장점이 있다. 특히 콤 필터링의 대상인, 복호화기에서 복원된 음성에 정확히 부합되는 값이므로 별도의 피치 측정 기법을 적용하는 것에 비해 보다 효과적인 필터링이 가능하다.

4. 실험 및 결과

개선된 콤 필터의 유효성을 검증하기 위해 두 종류의 음성 데이터를 대상으로 인식 실험을 수행하였다. 첫 번째 데이터는 QCELP 코더를 통과한 음성으로써 묵음 구간의 잡음이 제거된 데이터이며, 다른 하나는 제안한 콤 필터 자체의 성능 평가를 위해 음성 코더의 영향을 받지 않은 순수한 원음 음성(raw speech data)을 대상으로 하였다.

4.1. 실험 환경

음성인식 평가는 Aurora2 DB 및 HTK를 기반으로 수행하였다[9]. 훈련 및 실험 자료는 각각 8,440개, 4,004개의 연결 숫자로 구성되었으며, 실험 자료는 잡음 종류 및 채널 상태에 따라 세 가지 set (Set A~C)으로 나뉜다. Set A에는 네 종류의 잡음(기차, 소음, 자동차, 전시회장)이 포함되어 있으며, Set B는 Set A와 다른 종류의 잡음(레스토랑, 거리, 공항, 기차역)으로 구성되었다. Set C는 채널의 영향을 평가하기 위한 실험 자료로, 훈련 자료에 사용되는 G712 대신 MIRS 채널이 사용되었다. 모든 자료에 대해 13차 MFCC (에너지 포함), 차분, 가속으로 구성된 39차 MFCC를 특징 파라미터로 사용하였으며, 각 단어에 대해 left-to-right, 16-state, 3-mixture의 연속 HMM 모델을 구성하였다.

4.2. 실험 결과

제안한 방법의 성능 평가를 위해 세 종류의 데이터, 즉 콤 필터링을 수행하지 않은 데이터("original"), 기존의 콤 필터링을 수행한 데이터("CCF") 및 개선된 콤 필터에 의해 처리된 데이터("MCF")를 사용하여 성능을 비교하였다. <표 1>은 음성 코더의 영향을 받지 않은 원음 음성에 대하여 인식 실험을 수행한 결과로써, SNR 0dB에서 20dB의 평균 오류율(Word Error Rate)을 나타낸다. Aurora2 DB의 모든 실험 set에 대하여 콤 필터에 의한 잡음 제거 효과를 확인하였다. "original" 음성에 비해, CCF와 MCF는 각각 4.98%와 9.09%의 WER 감소율을 보였으며, MCF는 CCF에 비해 4.3%의 감소율을 나타냈다.

CDMA 단말기의 ASR 전처리에 제안한 방법을 적용한 후의 성능 평가를 위해

QCELP 코더의 출력 음성을 대상으로 위와 동일한 실험을 수행하였다. <표 2>는 필터링을 수행하지 않은 출력 음성("decoded")과 기존의 필터 및 개선된 필터에 의해 처리된 데이터(각각, CCF, MCF)에 대한 실험 결과이다. <표 1>에 비해 전체적으로 WER이 감소하였으며 이는 QCELP 코더의 영향에 기인한다. 개선된 콤 필터는 CCF 및 "decoded" 음성에 비해 각각 5.84%와 11.87%의 성능 개선을 보였다.

실험실 환경("clean") 및 높은 SNR에서의 성능을 제시한 <표 3>은 제안한 방법이 기존의 콤 필터를 더욱 효과적으로 개선하였음을 입증한다. CCF는 "decoded" 음성 및 MCF에 비해 높은 WER을 보였으며 실험실 환경에서 MCF는 CCF에 비해 무려 45.4%의 WER 감소율을 나타냈다. 실험실 환경 및 높은 SNR에서 CCF의 성능이 저하된 이유는 콤 필터링에 의한 신호의 왜곡에 기인한 결과이며, <표 3>의 결과를 통해 MMSE estimator 기반의 콤 필터가 깨끗한 음성에서 발생하는 이 같은 왜곡 문제를 효과적으로 개선함을 확인하였다.

<표 1> "original" 음성과 기존의 필터 및 개선된 필터에 의해 처리된 음성 간 WER (%)

Data Set	Set A	Set B	Set C	Avg	WER Reduction
original	38.76	44.35	33.23	38.78	-
CCF	36.42	41.76	32.36	36.85	4.98
MCF	35.22	39.29	31.25	35.25	9.09

<표 2> "decoded" 음성과 기존의 필터 및 개선된 필터에 의해 처리된 음성 간 WER(%)

Data Set	Set A	Set B	Set C	Avg	WER Reduction
decoded	36.75	42.46	32.76	37.32	-
CCF	35.02	38.32	31.46	34.93	6.40
MCF	33.88	35.03	29.77	32.89	11.87

<표 3> 실험실 환경 및 높은 SNR에서의 성능 비교 (WER) (%)

SNR(dB)	clean	20 dB	15 dB	Avg	WER Reduction
decoded	1.68	4.83	17.07	7.86	-
CCF	4.05	6.06	13.94	8.02	-2.04
MCF	2.21	5.25	11.00	6.15	21.76

5. 결 론

본 논문에서는 MMSE estimator를 통해 기존의 콤 필터링에 의해 발생되는 신호 왜곡을 개선함과 동시에 보다 효과적으로 잡음을 제거하는 방법을 제안하였다. 제안한 방법은 각 주파수 대역의 음성 존재 확률을 기반으로 콤 필터의 주파수 응답을 원래 음성의 주파수 응답에 적합하도록 조정함으로써, 음성이 존재하지 않는 대역에서의 콤 필터의 에너지를 낮추고 그렇지 않은 대역에서의 에너지는 유지하거나 강조한다. 본 논문에서는 또한 개선된 콤 필터링을 CDMA 단말기의 ASR 전처리에 적용하는 과정에서 QCELP 코더에서 제공되는 음성 파라미터를 콤 필터링에 효과적으로 사용하는 방법을 제시하였다. Aurora2 DB를 사용하여 원래의 음성 및 복호된 음성을 대상으로 인식 실험을 수행한 결과, 개선된 콤 필터의 성능이 기존의 필터에 비해 월등히 향상되었으며 실험실 환경 및 높은 SNR에서 콤 필터링에 의해 발생하는 신호의 왜곡 문제 또한 개선되었음을 확인하였다. 향후, 제안한 방법을 CDMA 환경뿐만 아니라 잡음 제거가 필요한 다양한 분야에 적용하여 성능 향상을 확인할 필요가 있다.

참 고 문 헌

[1] H. K. Kim, R. C. Rose "Cepstrum-domain acoustic feature compensation based on decomposition of speech and noise for ASR in noisy environments", *IEEE Trans. on speech and audio processing*, Vol. 11, No. 5, pp. 435-446, 2003.

[2] J. Stadermann, G. Rigoll, "Flexible feature extraction and HMM design for a hybrid distributed speech recognition system in noisy environments", *Proc. ICASSP* Vol. 1, pp. 332-335, 2003.

[3] R. H. Frazier, S. Samsam, "Enhancement of speech by adaptive filtering", *Proc. IEEE Int. Conf. on ASSP*, pp. 251-253, 1976.

[4] 박정식, 정규준, 오영환, "적응 콤 필터링을 이용한 이동 통신 환경에서의 강인한 음성 인식", *말소리*, 46호, pp. 65-76, 2003.

[5] K. Yanagisawa, Y. Tanaka, "Applying comb filter to noise reduction of hearing aid", *Proc. IEEE Int. Conf. on SMC*, Vol. 6, pp. 352-357, 1999.

[6] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error log spectral amplitude estimator", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-33, pp. 443-445, 1985.

[7] D. Malah, R. V. Cox, A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments", *Proc. ICASSP*, Vol. 2, pp. 201-204, 1999.

[8] W. Gardner, "QCELP: A variable rate speech coder for CDMA digital cellular", *Speech and audio coding for wireless and network applications*, pp. 77-84, B. S. Atal et. al. (eds), Kluwer Academic Pub., 1993.

[9] H. G. Hirsch, D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", *Proc. ICSLP*, Vol. 4, pp. 29-32, 2000.

접수일자: 2006년 11월 14일
게재결정: 2006년 12월 19일

▶ 박정식(Jeong-Sik Park) : 교신저자
주소: 대전시 유성구 구성동 373-1
소속: 한국과학기술원 전자전산학부 전산학전공
전화: 042) 869-5556
E-mail: parkjs@kaist.ac.kr

▶ 오영환(Yung-Hwan Oh)
주소: 대전시 유성구 구성동 373-1
소속: 한국과학기술원 전자전산학부 전산학전공
전화: 042) 869-3516
E-mail: yhoh@speech.kaist.ac.kr