AURORA 잡음 처리 알고리즘을 이용한 전화망 환경에서의 강인한 음성 검출

서영주(ICU), 지미경(ICU), 김회린(ICU)

<차 례>

1. 서론
2. AURORA front-end 잡음 처리 기법을 이용한 음질 개선
 2.1. 배 경
 2.2. 잡음 처리 알고리즘의 구조
3. 잡음에 강인한 음성 검출기
 3.1. 주파수 영역의 전처리 과정
 3.2. 임계치 설정의 자동화
 3.3. 유용한 대역 선정 기법
4. 실험 및 성능 평가
 4.1. 음성 데이터베이스
 4.2. 실험 방법
 4.3. 성능 평가
5. 결 론

<Abstract>

Robust Speech Detection Using the AURORA Front-End Noise Reduction Algorithm under Telephone Channel Environments

Youngjoo Suh, Mikyong Ji, and Hoi-Rin Kim

This paper proposes a noise reduction-based speech detection method under telephone channel environments. We adopt the AURORA front-end noise reduction algorithm based on the two-stage mel-warped Wiener filter approach as a preprocessor for the frequency domain speech detector. The speech detector utilizes mel filter-bank based useful band energies as its feature parameters. The preprocessor firstly removes the adverse noise components on the incoming noisy speech signals and the speech detector at the next stage detects proper speech regions for the noise-reduced speech signals. Experimental results show that the proposed noise reduction-based speech detection method is very effective in improving not only the performance of the speech detector but also that of the subsequent speech recognizer.

* Keywords: Speech detection, AURORA noise reduction, Telephone channel

1. 서 론

정보기술의 급격한 발전으로 인하여 인간을 더욱 편리하게 하는 여러가지 응용 서비스들이 급격히 등장하고 있다. 더욱이 급속히 보급된 휴대전화와 이를 연결하는 무선망, 그리고 기존에 구축된 유선망으로 이루어지는 유무선 전화망은 이러한 서비스들의 도입에 더없이 좋은 환경을 제공하고 있어서, 전화망을 이용하는 서비스들의 등장은 앞으로 더욱 증가할 전망이다. 전화망에서는 음성을 주요 통신 수단으로 사용하기 때문에 이러한 서비스에 음성인식 기술을 적용하는 것은 여러모로 매력적으로 보인다. 실제로 전화망 환경에서 음성인식 기술을 이용한 응용 서비스들을 도입하려는 시도가 많이 이루어지고 있다. 그러나 음성인식 기술을 적용한 이러한 시도들이 성공적인 결과를 얻기 위해서는 먼저 음성인식 시스템의 성능이 어느 수준 이상에 도달하여 서비스를 이용하는 사용자들에게 음성인식 기능을 이용함으로써 더욱 편리해졌다는 느낌을 가지게 해야 한다. 아울러 여러 사용자들을 동시에 수용할 수 있도록 음성인식 시스템이 대용량의 처리 능력을 갖추어야 하며 빠른 실시간 응답 특성도 보여주어야 한다. 그러나 현재 개발된 대부분의 음성인식 시스템들은 실험실이나 주변 환경이 조용한 장소에서는 만족할만한 성능을 발휘하지만 채널의 왜곡과 주변 잡음이 존재하는 전화망과 같은 환경에서는 여전히 무시할 수 없는 성능 저하를 보이고 있다.

이러한 성능 저하 요인들 중의 하나로 입력된 음성에 대한 부정확한 음성 구간 검출을 들 수 있다[1]. 즉, 음성 구간 검출시 잡음이 포함된 비음성 부분을 과도하게 포함시키거나 음성 구간의 일부 또는 전부를 잃어버리는 것과 같은 음성 검출의 실패가 결과적으로 다음 단계인 음성인식에서 불완전한 음성 입력으로 귀결되어 음성인식 시스템의 성능 저하를 가져온다는 점이다. 이 음성 검출 문제점은 음성인식 시스템 자체의 제한된 인식 성능과 더불어 현재 전화망에서 제공되는 응용 서비스에 음성인식 기술의 도입을 제한하는 주요 요인으로 작용하고 있다.

따라서 본 연구에서는 음성 구간 검출에서의 해결책의 하나로서 다양한 잡음이 존재하는 전화망 환경에서 제공될 수 있는 음성인식 응용 서비스에 적용이 가능한 잡음에 강인한 음성 검출 알고리즘을 제안하고자 한다. 제안된 기술은 전처리 단계에서 AURORA project에서 표준안으로 채택된 front-end 잡음 처리 기법[2, 3]에 기반한 잡음 제거 알고리즘을 적용하여 음성에 부가된 잡음을 먼저 일정 수준 이하로 제거하고, 잡음이 제거된 입력 음성 신호에 대하여 멜 대역 에너지에 기반한 음성 검출 방법으로 음성 부분을 검출하는 구조를 취하고 있다.

본 논문의 전체 구성은 다음과 같다. 먼저, 1장의 서론에 이어 2장에서 AURORA front-end의 잡음 처리 기법을 적용한 음질 개선에 대해 살펴본다. 3장에서는 멜 대역 에너지를 이용한 잡음 환경에 강인한 음성 검출 알고리즘을 제안하

고 4장에서 실험에 사용한 음성 데이터베이스, 실험 및 평가 방안, 실험 결과에 대해서 기술한다. 마지막으로 5장에서 결론을 맺도록 한다.

2. AURORA front-end 잡음 처리 기법을 이용한 음질 개선

2.1. 배경

ETSI (European Telecommunications Standards Institute)의 AURORA 워크그룹에서는 DSR (Distributed Speech Recognition)이라는 새로운 개념의 음성인식 적용 모델을 제안하였다[2]. DSR은 이동망 환경에서 이동 단말기를 통한 음성인식 응용 방법으로서 이동 단말기에서는 입력된 음성신호로부터 음성인식에 사용되는 특징 파라미터를 추출하고 이들을 채널 부호화한 다음 데이터 채널을 통해 음성인식기 서버로 전송하고, 인식기 서버에서는 수신된 특징 파라미터로부터 음성인식을 수행하는 구조를 취하고 있다. 이에 따라서 DSR을 위해 제안된 AURORA front-end의 잡음 처리 기법은 이동 단말기의 입력 음성에 부가되는 가산성 잡음을 효과적으로 제거하는 것을 목적으로 하고 있다. 내부 구조를 보면, 이러한 환경에서의 잡음 제거를 위해 2단계 mel-warped Wiener filtering 과정을 기본으로 채택하고 있으며 추가적으로 SWP (SNR-dependent Waveform Processing) 과정을 두고 있다[4]. 이 중에서 SWP 과정은 음성인식에 사용되는 특징 파라미터의 잡음에 대한 강인성을 향상시키는 데는 적합하지만 개선된 음성 신호를 복원해서 사용해야 하는 음성 검출과 같은 분야에서는 적용하기에 부적합하므로 배제하였다.

한편 AURORA front-end에서의 잡음 처리 방법은 이동 단말기에 입력되는 음성에 부가된 잡음을 제거하기 위해서 개발되었지만 본 연구에서의 잡음 제거는 전화망을 통과한 전화 음성을 대상으로 하고 있다. 따라서 기존의 AURORA front-end에서의 잡음 처리 기법을 전화 음성에 수정 없이 그대로 적용할 경우 서로 간의 환경 차이로 인한 성능 저하 문제를 야기할 가능성이 있다. 이러한 적용 환경 차이에서 기인하는 문제점을 해소하기 위한 방안으로서 전화망 환경에서 기존 AURORA front-end 잡음 처리 알고리즘의 파라미터 최적화 작업도 수행하였다.

2.2. 잡음 처리 알고리즘의 구조

2.2.1. Mel-warped Wiener filter

음성 신호에 잡음이 부가되었을 경우를 식 (1)과 같이 나타낼 수 있다

$$y(i) = s(i) + n(i) \tag{1}$$

여기서 y(i), s(i)와 n(i)는 각각 잡음이 부가된 음성, 양질의 음성, 및 잡음을 나타낸다. 위 식을 주파수 전력 스펙트럼으로 나타내면 다음 식과 같다.

$$P_y(f) = P_s(f) + P_n(f) \tag{2}$$

여기서 $P_y(f)$는 FFT(fast Fourier transform)를 기반으로 구해진 전력 스펙트럼의 연속적인 2 프레임에 대한 평균치로서 잡음이 내재된 음성신호의 추정된 전력 스펙트럼을 나타낸다.

이와 같은 상황에서 Wiener filter의 전달함수는 다음과 같이 정해진다.

$$H(f) = \frac{\sqrt{\eta(f)}}{\sqrt{\eta(f)}+1} \approx \frac{\sqrt{P_s(f)}}{\sqrt{P_s(f)}+\sqrt{P_n(f)}} \tag{3}$$

$$\eta(f) = \frac{P_{den}(f)}{P_n(f)} \tag{4}$$

여기서 $P_n(f)$는 AURORA front-end 잡음 처리 기법에서 사용하는 음성 검출기로부터 검출된 비음성 구간에 대해 추정된 잡음의 전력 스펙트럼이며, 잡음이 제거된 음성신호의 전력 스펙트럼, $P_{den}(f)$는 잡음이 내재된 음성 신호의 전력 스펙트럼, $P_y(f)$에서 $P_n(f)$가 제거된 성분이다.

일반적인 Wiener filter 전달함수의 추정은 식 (3)을 구성하는 성분들의 추정 오차가 최소가 되도록 구해진다. 그러나 최근의 연구 결과에 의하면 음질 개선 측면에서는 청각 인지 영역에서의 오차 최소화가 더 효과적이라고 알려져 있다[5, 6, 7]. 또한 음성인식 측면에서도 대부분의 음성인식 전처리 단계에서의 특징 추출은 청각 인지 영역인 멜 주파수에 기반을 두고 있다[8]. 이 점에 근거하여 AURORA front-end 잡음 처리 알고리즘에서 사용되는 Wiener filter의 전달함수 유도는 다음과 같은 일련의 과정으로 이루어진다[2, 3]. 먼저 Wiener filter에 적용된 mel-warped 스펙트럼 변환 함수는 식 (5)와 같다.

$$f_{mel}(k) = k \times \frac{MEL\{f_{lin\_s}/2\}}{K_{FB}+1} \tag{5}$$

$$MEL\{f_{lin}\} = 2595 \times \log_{10}(1 + f_{lin}/700\} \tag{6}$$

여기서 k는 mel-warped 필터뱅크의 인덱스를 나타내고, K_{FB}는 전체 필터뱅크의 수를 나타낸다. 또한 f_{lin_s}는 샘플링 주파수이다. 이로부터 k 번째 필터뱅크의 중심 주파수는 식 (7)과 같이 계산된다.

$$f_{cnt}(k) = 700 \times \left[10^{\frac{f_{mel}(k)}{2595}} - 1 \right], \quad 1 \le k \le K_{FB} \tag{7}$$

위에서 설명한 mel-warped 변환 과정을 Wiener filter의 전달 함수에 적용한 후, IDCT (Inverse Discrete Cosine Transform)를 취하여 구한 Wiener filter의 임펄스 응답은 다음과 같다.

$$\begin{aligned} h_{WF}(i) &= \frac{1}{2}\int_0^{\pi} H(f(k))\cos(f(k)\times i)df(k) \\ &\approx \sum_{k=0}^{K_{FB}} H(f(k))\cos(\frac{2\pi \times i \times f_{cnt}(k)}{f_{lin\_s}}) \times (\frac{f_{cnt}(k+1) - f_{cnt}(k-1)}{f_{lin\_s}}), \\ & 0 \le k \le K_{FB}+1, 0 \le i \le K_{FB}+1 \end{aligned} \tag{8}$$

구해진 Wiener filter의 임펄스 응답으로부터 음의 시간 영역에 해당하는 계수들의 복원을 위해 다음과 같은 미러링 과정을 거친다.

$$h_{WF\_mirr}(i) = \begin{cases} h_{WF}(i), & 0 \le i \le K_{FB}+1 \\ h_{WF}(2(K_{FB}+1)+1-i), & K_{FB}+2 \le i \le 2(K_{FB}+1) \end{cases} \tag{9}$$

Mel-warped Wiener filter의 causal 임펄스 응답은 $h_{WF_mirr}(i)$로부터 다음 식과 같이 구해진다.

$$h_{WF\_caus}(i) = \begin{cases} h_{WF\_mirr}(i+K_{FB}+1), & i = 0, \cdots, K_{FB} \\ h_{WF\_mirr}(i-K_{FB}-1), & i = K_{FB}+1, \cdots, 2(K_{FB}+1) \end{cases} \tag{10}$$

다음 단계에서는 필터링에서의 계산량을 줄이기 위해서, causal 임펄스 응답, $h_{WF_caus}(i)$의 계수들 중에서 중요한 부분을 적당한 수만큼 떼어서 취한다.

$$h_{WF\_trunc}(i) = h_{WF\_caus}(i + K_{FB} + 1 - (FL-1)/2),\ i = 0, \cdots, FL-1 \quad (11)$$

여기서, FL은 필터의 tap 수를 나타내며 17로 정해진다.

식 (11)과 같이 구해진 임펄스 응답에 Hanning window로 스무딩하여 최종적인 Wiener filter의 임펄스 응답을 다음과 같이 구한다.

$$h_{WF\_w}(i) = (0.5 - 0.5\cos(2\pi(i+0.5)/FL))h_{WF\_trunc}(i), 0 \le i \le FL-1 \quad (12)$$

위의 과정으로 구해진 임펄스 응답을 가진 Wiener filter에 잡음이 부가된 입력 음성 신호를 통과시켜 잡음이 제거된 음성 신호를 식 (13)과 같이 얻는다.

$$h_{WF\_w}(i) = \sum_{j=-(FL-1)/2}^{(FL-1)/2} h_{WF\_w}(j + (FL-1)/2)y(i-j), 0 \le j \le M-1 \quad (13)$$

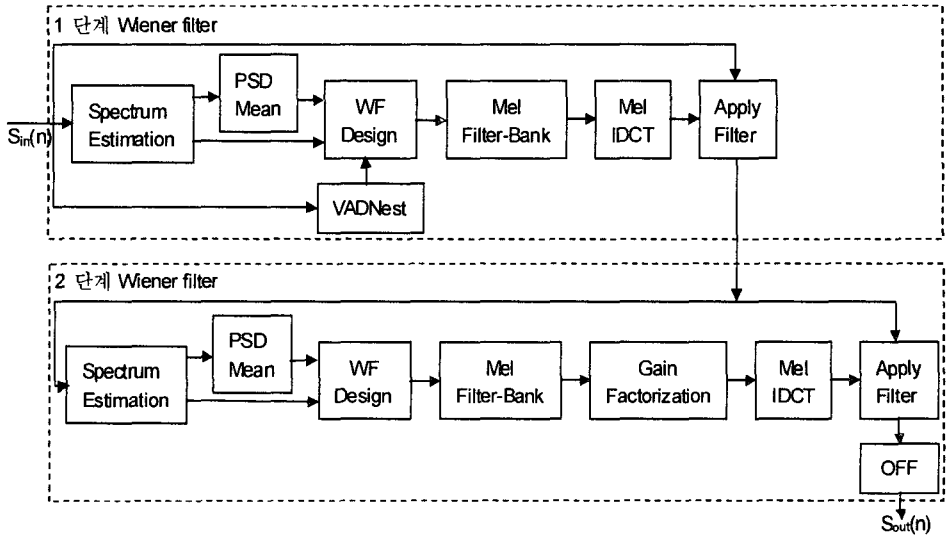
여기서, M은 프레임 이동 간격을 나타내며 80으로 정해진다.

2.2.2. 2단계 필터링

Mel-warped Wiener filter가 잘 동작하기 위해서는 잡음의 전력 스펙트럼 성분을 정확히 추정해야 한다. 이를 위해서 AURORA front-end에서는 1단계 필터링에서 시간 영역의 프레임 에너지에 기반하여 묵음과 음성 구간을 구분하는 간단한 음성 검출기(Voice Activity Detector)를 사용한다[2]. 그러나 이 음성 검출기는 초보적인 수준이어서 정확도 면에서 문제가 될 수 있다. 따라서 mel-warped 잡음 스펙트럼의 추정 신뢰성을 높이기 위해서 2단계의 적응 추정 기법을 사용한다[3]. 먼저 1단계의 Wiener filtering 과정에서는 VAD를 기반으로 추정된 잡음 스펙트럼의 추정 오차가 백색이라는 가정에서, mel-warped Wiener filtering이 부가 잡음을 제거하고 잡음 추정 오류에 의해 잔류하는 잡음을 백색화시킨다. 2단계에서는 1단계 필터링으로 백색화된 잔류 잡음에 대해 VAD를 사용하지 않고 사전에 정해진 초기 묵음 구간에 존재하는 잡음을 기반으로 하고 입력 신호의 전력 스펙트럼의 크기에 따라서 조정하는 방법으로 잔류 백색 잡음 성분을 추정하고 이를 제거한다.

또한 2단계 필터링에서는 gain factorization 과정을 별도로 두고있다. 이 과정에서는 먼저 1단계 필터링한 출력 음성 신호의 신호대 잡음비를 추정하고 Wiener filter의 이득값을 추정된 신호대 잡음비에 반비례하도록 조정한다. 필터링 과정에서는 구해진 이득값을 필터링의 정도(aggression)를 결정하는데 사용하여 잡음이 심한 낮은 신호대 잡음비 환경에서는 필터의 이득값을 높이고 높은 신호대 잡음비

의 입력 음성에 대해서는 이득값을 낮추는 방향으로 필터링한다. <그림 1>은 이 절에서 설명한 2단계 mel-warped Wiener filter를 이용한 AURORA front-end 잡음 처리 과정을 나타낸다.



<그림 1> 2-stage mel-warped Wiener filter의 구조

3. 잡음에 강인한 음성 검출기

3.1. 주파수 영역의 전처리 과정

제안된 음성 검출 알고리즘에서 사용되는 주파수 영역에 기반한 음성 검출 파라미터를 추출하는 방법[10]은 Wu et al[11]가 제안한 주파수 영역 파라미터를 추출하는 방법과 유사하다. 인간의 귀의 인지 특성을 반영하기 위하여 비선형 왜곡된 주파수 스케일인 멜-스케일 주파수 영역에서 등 간격으로 배치된 필터뱅크를 구현한다.

필터뱅크는 250~3500Hz의 주파수 범위에서 모두 13개의 대역 통과 필터들로 구성된다. 이 주파수 범위는 모음과 무성음의 에너지 분포 특성을 참고로 선정되었다. 필터의 응답 함수 f(i,k) ($0 \le i \le 12$, $0 \le k \le 128$)는 멜 주파수 영역에서 삼각형의 응답 특성을 나타내는, i 번째 대역 필터의 k 번째 주파수 값에 대한 필터 가중치이다.

구현된 멜-스케일 필터뱅크를 구성하는 각각의 대역 필터로부터 출력 에너지를 다음 과정을 통해서 구한 다음 주파수 영역에서의 파라미터를 추출한다. 시간 영

역의 음성 신호를 xt(n,m)이라 하고 n번째 음성 프레임의 m번째 음성 샘플로 정의하면 이 신호의 주파수 스펙트럼은 DFT (discrete Fourier transform)에 의해 Xfreq(n,k)로 나타난다. 여기서, k는 DFT에서의 주파수 인덱스를 의미한다. 이 주파수 스펙트럼 Xfreq(n,k)에 멜 주파수 대역의 필터 가중치, f(i,k)를 곱한 다음 대역별로 합산한 필터뱅크 에너지, Xmel(n,i)를 식 (14)와 같이 구한다.

$$X_{mel}(n,i) = \sum_{k=0}^{M(i)-1} | X_{freq}(n,k) | f(i,k), \quad 0 \le i \le 12 \tag{14}$$

여기서, n은 프레임 번호를, i는 필터뱅크에서의 주파수 대역 번호를, k는 푸리에 변환에서의 주파수 스펙트럼의 인덱스를 나타내고 M(i)는 i번째 대역에 포함된 주파수 인덱스의 수를 나타낸다. 구해진 필터뱅크 에너지에서 임펄스성 잡음을 제거하기 위하여 3-포인터 메디안 필터를 적용하여 스무딩된 $\hat{X}_{mel}(n,i)$을 구한다. 이 과정은 다음과 같다.

$$\hat{X}_{mel}(n,i) = median(X_{mel}(n-1,i), X_{mel}(n,i), X_{mel}(n+1,i)) \tag{15}$$

배경 잡음의 영향을 제거하기 위하여, 비음성으로 간주된 초기 몇 프레임들로부터 추정된 배경 잡음을 구하고 이 값을 $\hat{X}_{mel}(n,i)$에 감산하여 잡음 성분이 보상된 음성 신호의 필터뱅크 에너지, $\overline{X}_{mel}(n,i)$를 식 (16), (17)과 같이 얻는다

$$\hat{X}_{mel\_sil}(i) = \frac{1}{S}\sum_{n=0}^{S-1} \hat{X}_{mel}(n,i) \tag{16}$$

$$\overline{X}_{mel}(n,i) = \hat{X}_{mel}(n,i) - \hat{X}_{mel\_sil}(i) \tag{17}$$

여기서, S는 비음성으로 추정된 초기 프레임들의 수이다. 이와 같은 과정으로 구한 음성 신호의 정규화된 필터뱅크 에너지를 매 프레임마다 추출하여 음성 여부에 대한 판정에 사용한다.

3.2. 임계치 설정의 자동화

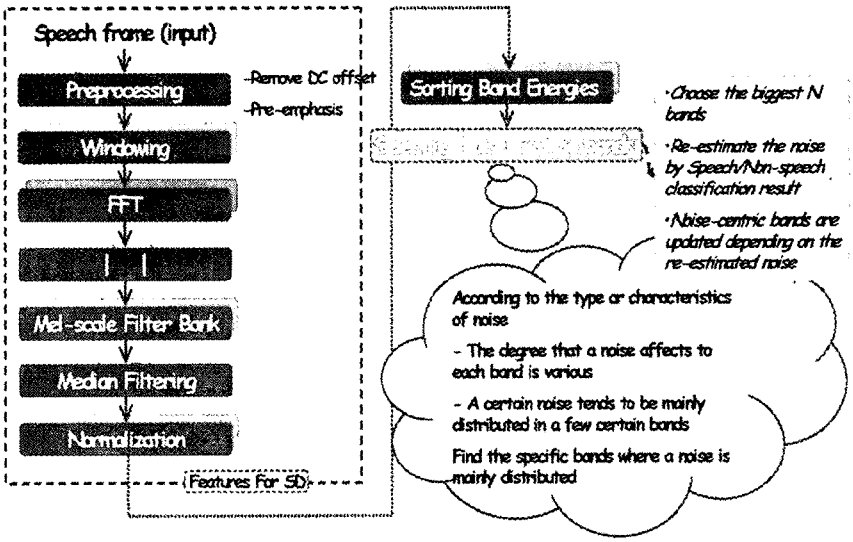
대부분의 음성 검출 알고리즘에서 음성 검출을 위한 특징값을 실험적으로 정해진 임계치와 비교하여 음성 경계 정보를 검출해 낸다. 여기서의 임계치란 음성

검출기의 좋은 성능을 위해서 실험 상황 또는 배경 잡음의 특성에 따라 적절히 변화해야 하는 값이다. 대부분의 알고리즘의 경우 음성의 초기 묵음 구간을 이용한 평균치에 비례하도록 임계치를 설정하고 이를 음성과 비음성을 구분하는데 사용한다. 이때 실험 상황이나 배경 잡음에 따라 수동적으로 임계치를 적절히 변화시킨다. 제안된 음성 검출 알고리즘은 이 임계치를 배경 잡음에 따라 적절히 변화하는 방법을 이용하였다. 즉 많은 노력에 의해 실험적 결과로 얻어지는 임계치를 배경 잡음에 따라 수동적으로 변화시킬 필요없이 자동적으로 계산하여 간단히 사용하고자 한다. 잡음 상황에서의 음성의 에너지는 대개 이항분포(binomial distribution), 즉 2개(음성과 배경 잡음의 분포)의 정규분포를 가지는데, 음성과 비음성을 구분하기 위한 통계적인 최적의 임계치는 이 두 정규분포가 만나는 곳이다[12]. 이 방법을 이용하려면 발성 전체를 히스토그램화 하여 최적의 임계치를 구할 수 있다. 그러나 이 방법은 실시간 음성 구간 검출이 불가능하기 때문에 음성의 초기 묵음 구간만을 히스토그램화하여 임계치를 정하였다. 우선 초기 묵음 구간을 이용하여 배경 잡음의 분포를 구하고, 누적분포를 계산하여 그 값이 전체의 80% 이상 되는 지점의 값을 이용하여 임계치로 정하였다.

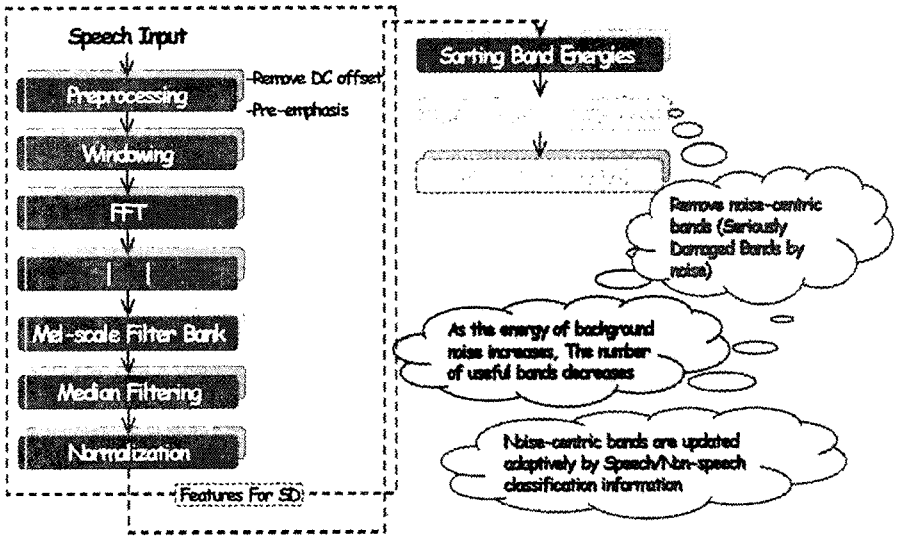
3.3. 유용한 대역 선정 기법

도메인이 한정적인 기존의 유용한 대역 기법과는[10] 달리 새로운 음성 검출 기법에서는 훈련 데이터를 사용하여 사전에 미리 유용한 대역을 선정하지 않고 입력 발화마다 유용한 대역을 선정한다. 인식 대상의 영역에 따라 다양한 잡음이 존재하지만, 잡음의 특성에 따라 특정 대역에 집중하는 현상 때문에 입력 발화의 초기 시작 프레임을 이용하여 잡음의 특성을 분석하여 집중 대역을 알아내어 유용한 대역에서 제외시킨다. 음성 검출 환경이 바뀔 때마다 훈련을 통해 유용한 대역을 재선택하는 과정이 불필요하다. 따라서 완전한 실시간 음성 검출이 가능하다. 그 절차는 다음과 같다. 먼저 초기 시작 프레임을 이용하여 잡음의 특성을 분석하여 잡음이 집중적으로 분포하는 대역을 알아낸다. 매 프레임마다 정규화된 필터뱅크 에너지를 구하고 이를 큰 것부터 내림차순으로 정렬한다. 실험 데이터의 신호대 잡음비에 따라 유용한 대역의 수를 결정하고, 발화 초기 프레임으로부터 알아낸 잡음 집중 대역을 유용한 대역에서 제외시키고 나머지 대역을 최종적인 유용한 대역으로 선정한다. 이때 신호대 추정된 잡음간의 신호대 잡음비가 0dB 이하인 대역은 유용한 대역에서 제외시킨다. 매 프레임마다 유용한 대역의 정규화된 필터뱅크 에너지를 구하여 각 대역의 임계치와 비교하여 그 프레임이 음성인지의 여부를 판정한다. 또한 현재 프레임의 음성/비음성 판정에 따라 잡음의 에너지를 갱신한다. 비음성 프레임으로 판명된 경우, 잡음의 대역 별 에너지를 다시 추정하고 이를 이용하여 음성/비음성 구간 분류에 사용되는 임계치를 갱신한다. 갱신된 잡음

의 대역별 에너지를 이용하여 잡음이 집중하는 대역을 재선정 한다. 이를 다음 프레임의 유용한 대역 선정에서 이용한다. 제안된 음성 검출 기법에서의 잡음 집중 분포 대역과 유용한 대역 선정 방법의 자세한 흐름도는 <그림 2>, <그림 3>과 같다.



<그림 2> 음성 검출기의 잡음 집중 분포 대역 선정 기법



<그림 3> 제안된 음성 검출기의 유용한 대역 선정 기법

4. 실험 및 성능 평가

4.1. 음성 데이터베이스

음성 검출 알고리즘의 성능 평가를 하기 위해 사용한 음성 DB는 ETRI DB로서 유선 전화 음성 DB(단어, 숫자)와 휴대폰 전화 음성 DB (단어, 숫자)로 구성되어 있다. 이 음성 DB는 각각 유선과 무선 전화망을 거쳐 8kHz의 표본화율과 16bit 양자화 단계로 수집되었으며, 유선 전화 음성 DB의 경우 약 54,000여 발화(단어, 숫자가 각각 27,000여 개)와 휴대 전화 음성 DB의 경우 약 49,000여 발화(단어, 숫자가 각각 26,000, 23,000여 개)로 구성되어 있다. 잡음이 없는 전화망 환경(Clean)에서의 실험을 위해서 이 DB로부터 신호대 잡음비가 25dB 이상인 발화들을 각각의 경우(유선 전화망 단어, 유선 전화망 숫자, 무선 전화망 단어, 무선 전화망 숫자)마다 1,300여 개 선별하였고 이를 양질의 전화 음성 DB로 사용하였다. 다양한 환경의 잡음이 내재된 전화 음성 DB를 구축하기 위해서 양질의 전화 음성 DB에 신호대 잡음비가 10dB, 15dB가 되도록 3종류의 AURORA 잡음(자동차 내부, 지하철 내부, 거리)을 부가하여 잡음이 내재된 전화 음성 DB(각각 1,300여 개의 발화로 구성된 12세트의 전화 음성 DB)를 구축하였고 AURORA front-end 잡음 제거 알고리즘의 적용 여부에 따라 총 24종류의 음성 검출 실험을 하였다.

음성 검출 알고리즘의 정확도가 음성인식기의 성능에 미치는 영향을 알아보는 실험도 아울러 행하였다. 그러나 음성인식 실험에서는 위에서 기술한 전화 음성 DB가 대상 어휘의 규모나 내용적인 측면에서 부적합하였기 때문에, ETRI에서 제작한 사연숫자음 전화 음성 DB를 사용하였다. 이 DB의 발화들은 유선 전화망 채널을 거쳐 8kHz의 표본화율과 16bit 양자화 단계로 수집되었으며 각 발화는 4개의 연속적인 숫자음들로 구성되어 있고 각 숫자음은 공을 포함해서 영에서 구까지 11개 중의 하나에 해당한다. 이 숫자음 DB는 신호대 잡음비가 20dB 이상인 비교적 양질의 전화 음성 DB로서 비슷한 비율의 남녀 화자 292명에 의해 발성된 약 60,000개의 발화들로 구성되었으며 각 발화들은 음성부의 전후에 각각 200msec 정도의 비음성 구간을 포함하고 있다. 이 데이타 중에서 약 50,000여 개의 발화들을 음성 인식기의 훈련에 사용하였고, 나머지 10,000여 개의 발화로 구성된 양질의 전화 음성 DB를 평가에 사용하였다. 또한 이 평가용 DB에 위에서 언급한 3종류의 AURORA 잡음을 인위적으로 가산하여 각각 5dB, 10dB, 15dB의 신호대 잡음비가 되도록 생성된, 잡음이 내재된 전화 음성 DB 9세트를 잡음 환경에서의 평가에 사용하였다.

4.2. 실험 방법

4.2.1. 음성 검출

전화망 환경 음성인식을 위한 강인한 음성 구간 검출 방법의 성능을 평가하기 위하여 유무선 전화망 환경에서 단어와 숫자음 데이터를 사용하였다. 음성 검출 성능 평가로서는 사전에 수작업으로 검출된 음성의 경계와 제안된 음성 검출 알고리즘에 의한 음성 경계 간의 차이를 구하고 이들의 통계적인 분포를 기존의 음성 검출 방법으로 구한 결과와 비교하는 방식을 택하였다. 즉, 음성 검출 알고리즘을 사용하여 구한 음성 경계가 수작업으로 검출된 음성 경계로부터 N 프레임 (10msec 단위) 이내일 정확도에 대한 분포를 구하였다. 음성 검출의 전처리 단계로서 AURORA 잡음 처리 알고리즘의 유효성을 확인하기 위하여 잡음 처리 과정의 유무에 따른 음성 검출 실험도 하였다. 이와 같은 성능 평가 방안을 사용하여 에너지와 영교차율에 기반한 기존의 음성 검출 방법과 기존의 훈련으로부터 선정된 유용한 대역 에너지를 이용한 음성 검출 방법과 유용한 대역을 매 프레임마다 선정하는 방식의 제안된 음성 검출 방법 간의 음성 검출 성능을 비교, 분석하였다. 개선된 멜 대역 에너지 기반의 음성 검출기의 경우, 일부 음성 DB(약 200여 개의 발화)를 이용하여 유용한 대역을 사전에 훈련하였다. 제안된 음성 검출 방법에서 사용되는 유용한 대역의 수는 13개 중 총 8개로 2kHz 이상을 고주파 대역 그 이하를 저주파 대역으로 나누어 각각 6개와 2개로 정하였다.

4.2.2. 음성인식

음성 검출기의 성능은 다음 단계에 위치하고 있는 음성인식기의 성능에도 중요한 영향을 미친다. 따라서 발성된 음성 신호를 음성 검출기에 통과시켜 검출된 음성 부분을 음성인식기의 입력으로 사용하였을 때, 인식 성능의 상대적인 차이를 조사함으로써 음성 검출기가 음성인식기의 성능에 미치는 영향을 파악할 수 있고 음성 검출기의 성능도 간접적으로 평가할 수 있다. 이를 위하여 4.1.절에서 설명한 3가지의 AURORA 잡음 환경과 잡음을 부가하지 않은 양질의 경우를 포함한 4가지의 신호대 잡음비 환경을 반영하는 평가용 전화 음성 데이터 10세트에 대해 기존 음성 검출 방법과 제안된 음성 검출 방법을 각각 적용하여 검출된 전화 음성 데이터에 대한 음성인식 실험을 하였다. 음성 검출기의 전처리 단계에서 적용된 AURORA front-end 잡음 처리 과정의 효과를 확인하기 위하여 잡음 처리 과정의 유무에 따른 음성인식 실험도 아울러 행하였다. 음성인식 실험을 위한 음성인식기로는 HTK Tools을 사용하였다[9].

다양한 잡음이 내재된 전화 음성 환경에 대한 AURORA front-end 잡음 처리 알고리즘의 최적화를 위하여, 적용된 Wiener filter의 tap 수 조정에 따른 잡음 처리 효과를 조사하기 위한 음성인식 실험도 하였다.

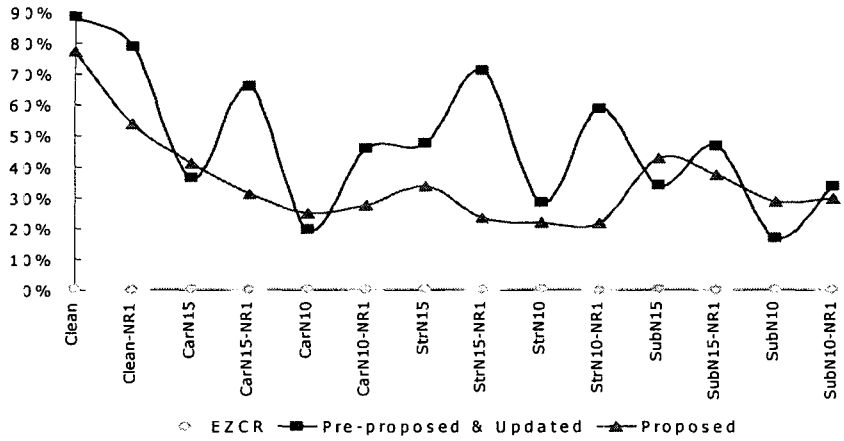
성능 평가를 위해서 다양한 잡음 환경을 반영하는 10세트의 평가용 DB로부터 잡음 제거를 하지 않은 경우와 AURORA front-end 잡음 처리 과정을 거친 경우에 해당하는 20세트의 평가용 DB를 추가로 생성하였다. 이 전화 음성 DB에 수작업에 의한 음성 검출 방법, 에너지-영교차율 기반의 음성 검출기, 그리고 제안된 멜 대역 에너지에 기반한 음성 검출기를 적용하여 음성 검출된 각각의 데이터를 음성인식기의 입력으로 사용하였을 때의 인식 성능을 조사하였다.

음성인식에 사용한 특징 파라미터는 20msec의 프레임 길이로 매 10msec마다 추출된 음성 프레임으로부터 12차 MFCC 계수와 프레임 에너지 및 이들의 델타값과 델타-델타 파라미터를 포함하는 39차 계수를 추출하여 사용하였고 전화 채널 특성을 제거하기 위해 켑스트럼 평균치 정규화 (cepstral mean normalization) 과정을 추가하였다.

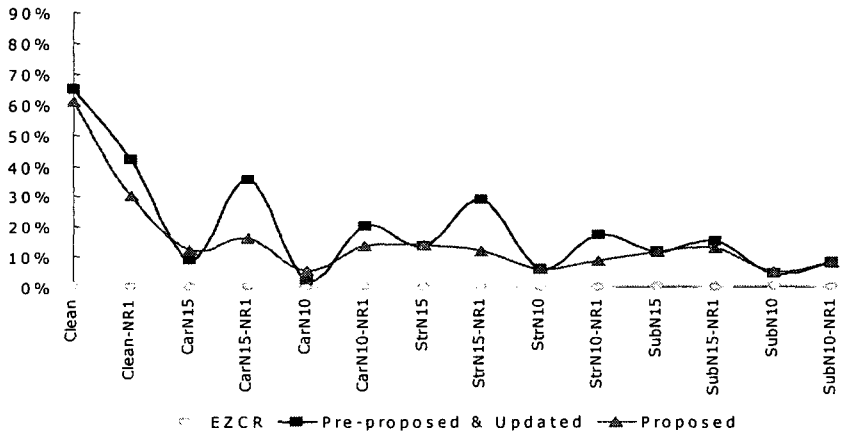
4.3. 성능 평가

4.3.1. 음성 검출

아래의 <그림 4>, <그림 5>는 다양한 잡음 환경(Clean: 양질의 전화 음성, Car: 자동차 내부 잡음, Sub: 지하철 내부 잡음, Str: 거리 잡음)에서의 무선 전화망 숫자음 DB에 대해 잡음 제거 유무에 따른 음성 검출 알고리즘에 의해 검출된 경계와 수작업에 의한 레이블링 위치와의 거리 분포를 나타내었다. 그림에서 보듯이, 음성 검출 전에 훈련을 통해 유용한 대역을 선정한 음성 검출기(Pre-proposed & Updated)[10]의 성능이 가장 좋음을 알 수 있고, 여기에 잡음 제거를 적용했을 경우 가장 뛰어난 성능 향상을 보이고 있다. 제안된 음성 검출기(Proposed)의 성능 또한 우수한 것을 알 수 있다. 그러나 끝점과는 달리 시작점의 경우 잡음 제거에 의한 성능 향상은 거의 없다. 마지막으로 에너지-영교차율 기반의 음성 검출기(EZCR)의 경우, 다양한 잡음 환경에서 음성 검출기가 거의 동작하지 않는 것을 알 수 있다. <표 1>은 다양한 잡음 환경에서 음성 검출기들의 성능을 나타낸 것이다. <그림 4>, <그림 5>와 비슷한 성향을 보여주고 있다.



<그림 4> 다양한 잡음 환경에서 AURORA 잡음 처리 여부에 따른 음성 검출기의 정확도 (검출된 시작점이 0~5 프레임 이내에 올 확률, 유선 전화망 숫자음)



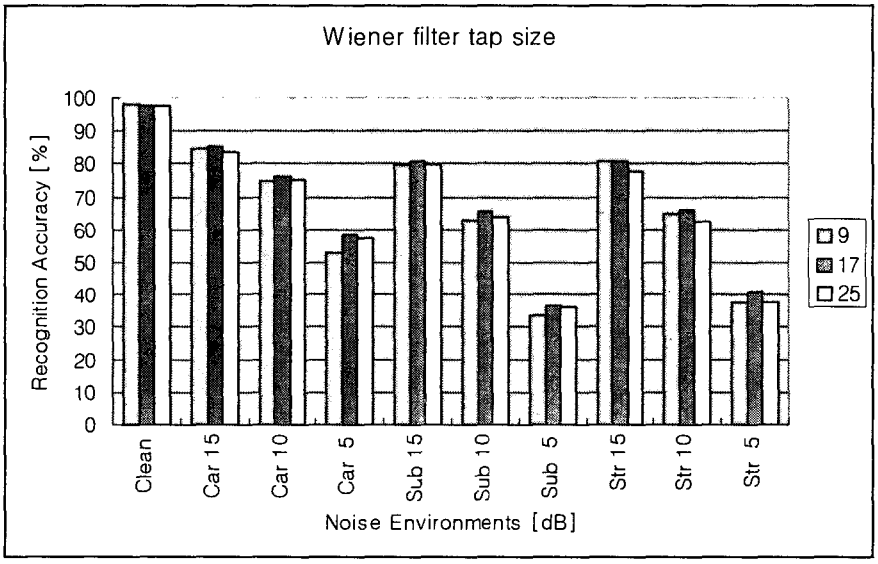
<그림 5> 다양한 잡음 환경에서 AURORA 잡음 처리 여부에 따른 음성 검출기의 정확도 (검출된 끝점이 0~5 프레임 이내에 올 확률, 유선 전화망 숫자음)

<표 1> 다양한 잡음 환경에서 음성 검출기의 성능 (0~5 프레임 이내에 검출할 확률[%])

	DB / Detectors	TW Clean	TW Str15	TD Clean	TD Sub15	MW Clean	MW Car15	MD Clean	MD Car15
Start	EZCR	8.7	4.4	0.2	0.2	0.0	0.0	0.3	1.1
	Prev. Pro.	84.3	50.5	88.5	33.9	74.5	26.2	59.7	11.5
	Pro.	70.7	36.4	77.2	42.7	39.8	22.0	36.4	15.4
End	EZCR	4.8	2.6	0.1	0.2	0.0	0.1	0.8	1.4
	Prev. Pro.	60.5	15.7	65.0	11.5	81.0	9.8	55.2	8.5
	Pro.	55.9	11.6	60.9	11.6	54.3	12.5	50.1	10.4

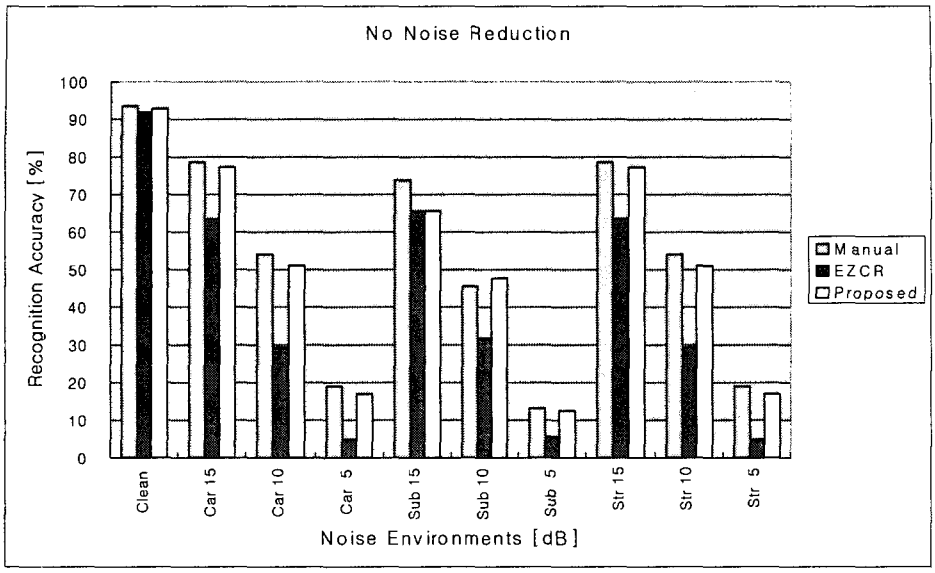
4.3.2. 음성인식

<그림 6>에서는 AURORA front-end 잡음 처리 알고리즘에서 적용한 Wiener filter의 tap 수에 따른 음성인식률의 차이를 나타내었다. 다양한 잡음 환경이 반영된 전화망 환경에 대한 음성인식 실험 결과로부터 tap 수가 17일 경우에 최고의 인식 성능을 나타냄을 알 수 있으며 이는 AURORA 표준안에서 정한 값과 같다. 따라서 이 실험 결과로부터 전화망 채널 왜곡에 대해 Wiener filter의 tap 수는 크게 영향을 받지 않음을 알 수 있다.



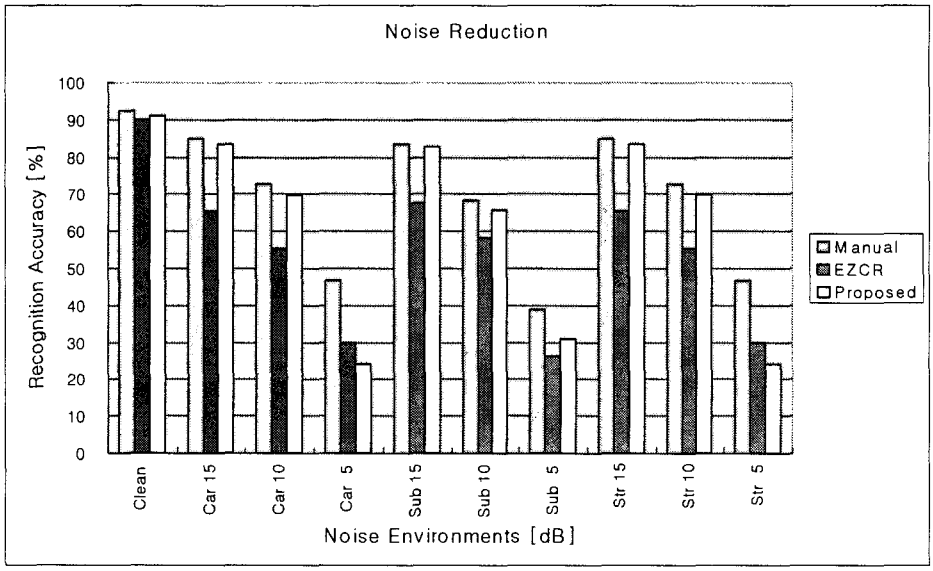
<그림 6> Wiener filter의 tap 수에 따른 전화 음성 인식률

<그림 7>은 AURORA front-end 잡음 제거 과정을 거치지 않았을 경우의 다양한 잡음 환경에 대한 전화 음성인식 결과를 나타낸다. 이 그림에서는 신호대 잡음비가 증가할수록 음성인식률도 거의 선형적으로 증가하였다. 음성 검출 방법에 따른 인식률을 비교하면, 제안된 음성 검출 방법(Proposed)을 적용하였을 때의 음성인식률이 수작업(Manual)의 경우와 거의 대등한 성능을 보이면서 기존의 에너지-영교차율 기반의 음성 검출 방법(EZCR)에 비해 훨씬 우수함을 알 수 있다.



<그림 7> 음성 검출 방법에 따른 전화 음성 인식 성능
(AURORA front-end 잡음 처리 과정을 거치지 않은 경우)

<그림 8>은 AURORA front-end 잡음 제거 과정을 거쳤을 경우의 다양한 잡음 환경에 대한 전화 음성 인식 결과를 나타낸다. 잡음 제거 과정을 적용한 경우에는 자동차 내부 잡음 환경이 나머지 두 종류의 잡음 환경에 비해 더 좋은 성능을 보였다. 이는 자동차 소음이 비교적 정적인 특성을 띠고 있으며 주파수 성분도 음성 신호의 주요 주파수 성분과 크게 겹치지 않기 때문에 Wiener filter에 의해 효과적으로 제거되었음을 추정할 수 있다. 반면에 거리 잡음에 대한 잡음 제거 과정이 적용된 실험에서, 신호대 잡음비가 아주 낮은 5dB의 경우에는 제안된 음성 검출 방법이나 에너지-영교차율 기반의 방법이 수작업에 의한 음성 검출에 비해 다소 큰 성능 차이를 나타내었다. 이는 거리 잡음이 간헐적인 임펄스성 잡음을 다른 두 잡음에 비해 많이 포함하고 있어서 정적인 잡음의 제거에 초점을 맞추고 있는 Wiener filter가 효과적으로 잡음을 제거하지 못한데서 기인하였다고 추측된다. 지하철 내부 잡음은 자동차 내부 잡음과 어느 정도 비슷한 주파수 특성을 띠고 있지만 지하철이나 기차의 선로에서 발생하는 특유의 규칙적인 주기를 가지는 잡음이 그 주기 내에서 시간에 따라 변하는 잡음 역할을 하였다.



<그림 8> 음성 검출 방법에 따른 전화 음성 인식 성능
(AURORA front-end 잡음 처리 과정을 거친 경우)

따라서 지하철 내부 잡음의 경우 자동차 내부 잡음에 비해 훨씬 낮은 인식 성능을 나타내었다. 결국, 지하철 내부나 거리와 같은 환경에서 발생하는 잡음을 효과적으로 제거하기 위해서는 시간에 따라 변하는 특성을 띤 잡음을 효과적으로 제거할 수 있는 알고리즘의 개발이 필수적이라 하겠다.

세 종류의 부가 잡음들에 대해서 AURORA front-end 잡음 처리를 하였을 때의 결과를 잡음 처리를 하지 않았던 이전의 결과와 비교하면, 세 음성 검출 방법 모두에 걸쳐서 5dB, 10dB, 15dB의 잡음 환경에서 현저한 인식률 개선을 확인할 수 있다. 특히 제안된 음성 검출 방법에 의한 인식 성능은 수작업 음성 검출 방법에 의한 결과와 거의 유사함을 알 수 있다. 따라서 음성 검출기에 전처리 단계로서 AURORA front-end 잡음 처리 알고리즘을 추가할 경우, 잡음이 제거된 특징 파라미터를 사용하는데서 기인하는 음성인식 성능 향상뿐만 아니라 음성 검출 성능의 개선에서 오는 추가적인 음성인식 성능 향상도 얻을 수 있음을 알 수 있다.

5. 결 론

전화망 환경에서 음성인식은 일반적인 사무실 환경에 비해 무시할 수 없을 정도의 인식률 저하를 가져온다. 이러한 인식률 저하 요인들 중 하나로 음성 부분을 정확하게 검출하지 못한다는 점을 들 수 있다. 본 연구에서는 이 문제점을 해결하

기 위하여 전화망 환경에서 음성인식을 위한 잡음에 강인한 음성 검출 알고리즘을 제안하였다. 이 알고리즘은 먼저 여러 종류의 잡음이 상당한 수준으로 존재하는 전화망 환경을 고려하여 AURORA 프로젝트의 front-end의 잡음 처리에서 표준안으로 채택된 2단계 mel-warped Wiener filter를 사용하여 음성에 부가된 잡음을 효과적으로 제거하고자 하였다. 음성 검출 부분에서는 음성을 정확하고 효과적으로 검출하기 위해 음성 검출에 유용한 대역을 제안된 방법에 의해 훈련 과정 없이 실시간으로 선정하고, 선정된 대역들의 필터뱅크 에너지를 음성인지의 여부를 판정하는 파라미터로 사용하여 음성과 비음성을 구분하였다.

음성 검출기의 성능 평가 실험에서 사전에 훈련을 통하여 유용한 대역을 구하고 임계치 자동화 기법과 신호대 추정된 잡음비에 의해 유용한 대역을 재선정하는 개선된 방법이 사전 훈련 과정 없이 매 프레임마다 유용한 대역을 선정하는 방법에 비해 좋은 성능을 보였다. 전처리 잡음 제거 과정을 적용하여 음성 검출의 성능 향상을 볼 수 있었고, 개선된 멜 밴드 기반의 음성 검출 알고리즘의 경우 기존의 방법에 비해 향상된 성능을 보여주었다.

음성 검출기를 적용한 음성 검출된 음성 데이터에 대한 음성인식 실험에서 제안된 음성 검출 방법은 기존의 에너지-영교차율 기반의 음성 검출 방법에 비해 훨씬 우수한 성능을 나타내었으며, 특히 신호대 잡음비가 10dB 이상인 전화망 환경에서는 수작업에 의한 음성 검출 방법에 근접하는 성능을 보였다. 따라서 제안된 음성 검출법이 전화 음성에서 일반적으로 나타나는 잡음 수준에 대해서 음성 검출의 신뢰도를 더 높일 수 있을 것으로 예상한다.

참 고 문 헌

[1] J. C. Junqa, B. Mark, B. Reaves, "A Robust Algorithm for Word Boundary Detection in the Presence of Noise", *IEEE Trans. Speech and Audio Processing,* Vol. 2, No. 3, pp.406-412, July 1994.

[2] ETSI final draft standard doc., "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms", *ETSI ES 202 050, V1.1.1 (2002-07),* July 2002.

[3] A. Agarwal, Y. M. Cheng, "Two-Stage Mel-Warped Wiener Filter for Robust Speech Recognition", *Proc. ASRU '99,* 1999.

[4] D. Macho, Y. M. Cheng, "SNR-Dependent Waveform Processing for Robust Speech Recognition", *Proc. ICASSP 2001,* pp.305-308, 2001.

[5] Y. M. Cheng, D. O'Shaughnessy, "Speech Enhancement based Conceptually on Auditory Evidence", *IEEE Trans. Signal Processing,* Vol. 39, No. 9, pp.1943-1954, 1991.

[6] D. E. Tsoukalas, J. N. Mourjopoulos, G. Kokkinakis, "Speech Enhancement based on Audible Noise", *IEEE Trans. Speech and Audio Processing,* Vol. 5. No. 6, pp.497-514, 1997.

[7] N. Virag, "Speech Enhancement based on Masking Properties of the Auditory System", *Proc. ICASSP '95,* pp.796-799, 1995.

[8] S. Davis and P. Mermelstein, "Comparison of Parametric Representation for Monosyllable Word Recognition in Continuously Spoken Sentences", *IEEE Trans. ASSP,* Vol. 28, No. 4, pp.357-366, 1980.

[9] S. Young, *The HTK BOOK (Ver. 2.2),* Entropic Ltd., Jan. 1999.

[10] M. Ji, H. Kim, "Keyword Spotting using Distance-Dependent Multiple Modeling," *Proc. of ICSP2001,* pp.162-166, Aug. 2001.

[11] G. D. Wu, C. T. Lin, "Word boundary detection with Mel-scale frequency bank in noisy environment", *IEEE Trans. Speech and Audio Processing,* Vol. 8, No. 5, pp.541-554, Sept. 2000.

[12] S. V. Gerven, F. Xie, "A Comparative Study of Speech Detection Methods", *Eurospeech 97,* pp.1095-1098, 1997.

접수일자: 11월 7일
게재결정: 12월 12일

▶ 서영주(Youngjoo Suh)
주소: 305-732 대전광역시 유성구 화암동 58-4번지 한국정보통신대학원대학교
소속: 한국정보통신대학원대학교(ICU) 음성인식기술 연구실
전화: 042) 866-6221
E-mail: yjsuh@icu.ac.kr

▶ 지미경(Mikyong Ji)
주소: 305-732 대전광역시 유성구 화암동 58-4번지 한국정보통신대학원대학교
소속: 한국정보통신대학원대학교(ICU) 음성인식기술 연구실
전화: 042) 866-6207
E-mail: lindaji@icu.ac.kr

▶ 김회린(Hoi-Rin Kim)
주소: 305-732 대전광역시 유성구 화암동 58-4번지 한국정보통신대학원대학교
소속: 한국정보통신대학원대학교(ICU) 음성인식기술 연구실
전화: 042) 866-6139
E-mail: hrkim@icu.ac.kr