한국음향학회지 제28권 제1호 pp. 70~76 (2009)

잡음 환경에서의 음성 명료도 향상 기술

Improvement of Speech Intelligibility in Noisy Environments

윤 제 열*, 김 중 회**, 오 은 미**, 박 호 종*
(Jae-Yul Yoon*, JungHoe Kim**, Eunmi Oh**, Hochong Park*)

*광운대학교 전자공학과, **삼성전자 멀티미디어랩

(접수일자: 2008년 11월 12일; 수정일자: 2008년 12월 29일; 채택일자: 2008년 12월 30일)

주변 잡음이 심한 환경의 음성 통신에서 음성 명료도는 주변 잡음의 마스킹 효과로 인하여 크게 저하된다. 본 논문에서는 잡음 환경에서 음성 명료도를 향상시켜 통화 품질을 높이는 새로운 방법을 제안한다. 청각 이론에 의하면 음성의 시간축 포락선은 명료도 결정에 중요한 역할을 한다. 이에 따라 본 논문에서는 대역별 시간축 포락선의 변화를 강화하여 명료도를 향상시키는 방법을 사용하며, 음질을 추가로 향상시키기 위한 피치 강화 동작을 포함한다. 또한, 실제 통화 상황에서의 정확한 주관적 성능 평가를 위하여 양 귀를 이용하는 새로운 주관적 성능 평가 방법을 제안한다. 제안하는 평가 방식을 통하여 제안하는 명료도 향상 기술의 성능을 평가하였으며, 명료도와 음질이 모두 향상되는 것을 확인하였고, 동작 파라미터 조정을 통하여 명료도와 음질 사이의 상호 관계가 조정되는 것을 확인하였다.

핵심용어: 음성 명료도, 잡음 환경, 시간축 포락선, 피치, 음질
투고분야: 음성처리 분야 (2)

In speech communications in noisy environments, speech intelligibility is seriously degraded due to the masking effect of ambient noise. In this paper, a new method to improve speech intelligibility in noisy environments is proposed. Based on the perception theory that the temporal envelope plays a major role in determining intelligibility, the proposed method uses a novel operation that enhances the fluctuation of band-wise temporal envelope and also contains pitch enhancement for improving speech naturalness. In addition, a new subjective evaluation scheme employing binaural listening is proposed in order to measure more reliable performance. The subjective performance measured with the proposed scheme shows that the proposed method improves both intelligibility and naturalness in various environments, whereas a function parameter can control the performance trade-off between intelligibility and naturalness.

Keywords: Speech intelligibility, Noisy environments, Temporal envelope, Pitch, Speech quality
ASK subject classification: Speech Signal Processing (2)

I. 서 론

주변 잡음이 심한 환경에서 음성 통신이 이루어지면 주변 잡음의 마스킹 효과에 의하여 음성의 명료도가 크게 저하되고 그에 따라 통화 품질이 저하된다. 예로, 그림 1과 같이 near-end 잡음이 심한 환경에서의 이동통신에서 사용자는 far-end 음성 신호를 청취하면서 동시에 near-end 잡음을 듣게 되어 far-end 음성에 대한 명료도가 저하된다. 음성 명료도가 저하되면 정확한 음성정보 전달이 불가능하므로 음성 통신에서 높은 음성 명료도는 반드시 필요한 성능 조건이다. 일반적으로 주변 잡음이 심한 환경에서 사용자는 스피커 출력 볼륨을 높여 명료도를 향상시키려 한다. 그러나 스피커 볼륨이 커지면 명료도와 관련이 없는 출력 음성의 모든 성분이 일정

책임저자: 박 호 종 (hcpark@kw.ac.kr)
서울 노원구 월계동 447-1 광운대학교 전자공학과
(전화: 02-940-5104; 팩스: 02-913-9057)

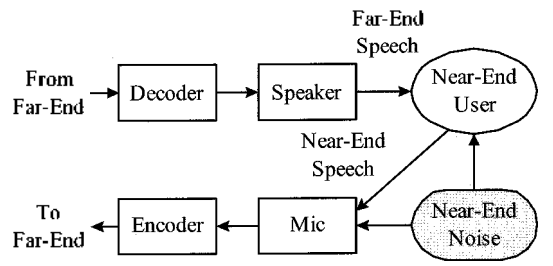

그림 1. 주변 잡음이 심한 환경의 이동통신에서의 사용자 청취 구조
Fig. 1. Near-end listening structure of cellular communications in noisy environments.

하게 증가되는 문제점이 나타난다. 따라서 음성 신호의 분석을 통하여 명료도와 밀접한 관련이 있는 성분만을 강화하여 보다 효율적으로 명료도를 향상시키는 새로운 기술이 필요하다.

잡음 환경에서 음성 명료도를 향상시키는 기술에 대한 연구 결과가 많이 보고되고 있다 [1-3]. 그러나 [1]과 [2]와 같이 주변 잡음의 크기에 따라 음성 레벨을 조정하여 명료도를 향상시키는 방법은 휴대전화기의 음향특성에 따라 성능이 변하는 문제점을 가진다. 즉, 각 방법은 마이크로 입력되는 잡음의 크기를 분석하고 이를 기준으로 최적의 출력 음성 레벨을 결정하고 스피커를 통하여 사용자 귀로 전달한다. 따라서 사용자가 느끼는 음성 크기는 잡음 크기뿐만 아니라 마이크와 스피커의 특성에 따라 변한다. 반면, 잡음은 직접 귀로 전달되므로 사용자가 느끼는 잡음 크기는 휴대전화기 특성에 따라 변하지 않는다. 따라서 사용자가 느끼는 음성과 잡음의 레벨 차이는 명료도 향상 블록에서 최적으로 결정하였던 레벨 차이와 다르게 되고, 목표로 하는 명료도 향상을 제공하기 어렵다. 또한, 기존 연구에서 사용하였던 주관적 성능 평가 방법이 실제 통화 환경을 정확히 표현하지 못하고 있고 각 논문에서 제시하는 성능이 실제 환경에서의 성능에 비하여 과장되는 경향을 가진다. 예로, 실제 상황의 명료도를 정확하게 측정하기 위하여 양 귀로 음성과 잡음을 청취하면서 성능을 측정하여야 하지만, 아직 이 방법으로 성능을 평가한 사례는 없다.

본 논문에서는 이와 같은 기존 기술의 문제점을 해결하는 새로운 음성 명료도 향상 기술을 제안한다. 제안하는 기술은 주변 잡음의 크기와 무관하게 명료도와 밀접한 관련이 있는 음성 특성을 강화하는 방법에 기반을 둔다. 청각 이론에 의하면 음성의 시간축 포락선 (temporal envelope)은 음성 명료도를 결정하는데 중요한 역할을 하며, 이 이론에 따라 제안하는 기술은 음성의 대역별 시간축 포락선의 변화를 강화하여 명료도를 향상시키는 것을 핵심 동작으로 사용한다. 그리고 명료도 향상 과정에서의 신호 변형에 의한 음질 저하를 극복하기 위하여 음질 향상을 위한 피치 강화 모듈을 포함한다. 또한, 본 논문에서는 명료도에 대한 정확한 주관적 성능 평가를 위하여 양 귀를 이용하여 음성과 잡음을 청취하는 새로운 평가 방법을 제안한다. 제안한 성능 평가 방법을 사용하여 명료도 향상 기술에 의하여 명료도와 음질이 모두 향상되는 것을 확인하였고, 또한 명료도 향상 과정에서 동작 파라미터 값에 따라 명료도와 음질이 향상되는 정도가 조정되고, 두 평가 항목 사이의 보완 관계가 조정되는 것을 확인하였다. 이 결과는 통신의 사용 환경 또는 사용자 선호도에 따라 명료도와 음질 사이에서 보다 중요한 항목에서 높은 성능을 제공할 수 있도록 하여 준다.

II. 이론적 배경

음성 지각에 대한 많은 연구에 의하면 음성 신호의 시간축 포락선은 음성 명료도를 결정하는데 매우 중요한 역할을 한다. 명료도에 대한 시간축 포락선의 역할을 청각 이론을 기반으로 간단히 설명하면 다음과 같다. 그림 2는 인간 청각의 시간 영역 처리 모델을 보여준다 [4]. Non-Linear Device와 Temporal Integrator는 hair cell이 필터 출력 신호에 반응하는 과정에 해당하며, Non-Linear Device는 반파 정류기 또는 square-law 동작을 하고 Temporal Integrator는 저대역 통과 또는 smoothing 동작을 하며, 그에 따라 Temporal Integrator 출력 신호는 수학적으로 시간축 포락선에 해당한다. Decision Device는 신경 시스템에서 구체적으로 소리를 인식하는 과정을 모델링하며, 입력값, 입력 variance, 또는 신호의 최대/최소의 비가 일정값 이상이면 반응을 한다고 알려져 있다. 세 가지 이론은 서로 다른 동작을 의미하지만 개념적으로 시간축 포락선의 변화가 충분히 있고 동작 영역이 충분한 크기를 가져야 청각 반응을 하는 것을 의미하고, 이 조건을 만족하지 못하면 확실한 반응이 발생하지 않아 명료도가 저하된다고 할 수 있다.

따라서 음성의 시간축 포락선은 인지 과정에서 중요한 역할을 하며, Decision Device이 주변 잡음에 의하여 시간축 포락선의 변화를 인지하지 못하면 명료도가 저하된다. 결국, 시간축 포락선의 변화를 증폭시키면 잡음 효과를 감소시켜 음성의 명료도가 향상될 수 있다. 명료도와 시간축 포락선의 관계에 대한 실질적인 연구로서, 시간축 포락선이 완만해지면 명료도가 저하되는 것을 검증한 사례가 있고 [5], 신호 사이의 시간축 포락선을 비교하여 명료도 감소량을 정량적으로 표현하는 방법이 개발된 연구가 있다 [6]. 따라서 시간축 포락선은 음성 명료도를 향상시킬 수 있는 매우 중요한 음성의 특성 파라미터가 된다.

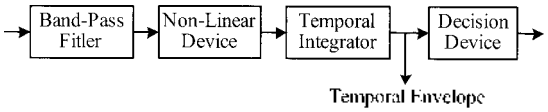

그림 2. 인간 청각의 시간 영역 처리 모델 [4]
Fig. 2. Temporal processing model of human ear [4].

제안하는 기술에서 대역의 시간축 포락선을 구하여 과정은 다음과 같다. 대역 b에 해당하는 대역통과 협대역 신호 $s_b[n]$는 $s_b[n]=a_b[n]\cos(2\pi f_b n+\theta_b)$와 같이 케리어 주파수 f_b와 진폭 $a_b[n]$을 가지는 진폭변조된 신호로 모델링된다. 따라서 대역 b의 시간축 포락선은 $a_b[n]$가 되고, $a_b[n]$는 $s_b[n]$의 Hilbert 변환 $s_b^H[n]=Hilbert(s_b[n])$를 구한 후 $a_b[n]=\sqrt{s_b[n]^2+s_b^H[n]^2}$ 를 통하여 구한다.

III. 제안하는 명료도 향상 기술

3.1. 개요

그림 3은 제안하는 명료도 향상 방법의 전체 구조를 보여준다. 입력 신호의 피치를 강화하는 과정을 거쳐, 대역별 시간축 포락선을 구하고 이의 변화를 강화하는 과정을 거친다. 시간축 포락선 강화 동작에서 강화량을 결정하여 명료도 향상 정도를 조정하는 강화 조절 파라미터를 가지며, 이는 명료도와 음질 사이의 상호 성능을 조절하는 역할을 한다.

명료도 향상을 위하여 음성 신호에 변형을 가하면 일반적으로 음질은 저하되는 경향을 가진다. 본 논문에서는 "명료도"와 "음질"을 음성의 품질을 평가하는 두 개의 서로 독립적인 항목으로 사용한다. 명료도는 음성과 잡음을 동시에 청취할 때 음성이 가지는 정보의 정확한 인지에 대한 평가하고, 음질은 잡음 없이 음성만 청취할 때 음성 고유 특성의 왜곡에 대한 평가이다. 실제 통신에서 전체 품질은 명료도와 음질의 가중치 적용된 합으로 표현

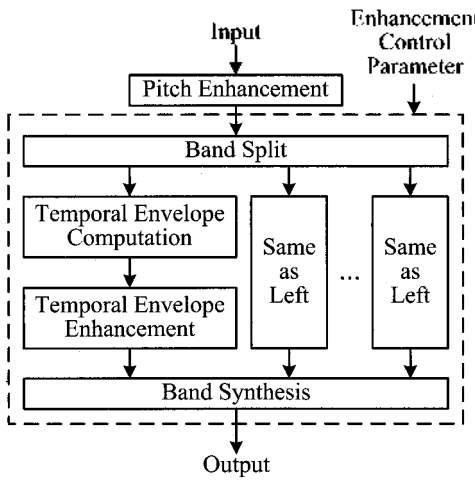

그림 3. 제안하는 명료도 향상 방법의 전체 구조
Fig. 3. Overall structure of the proposed intelligibility improvement method.

되므로 실질적인 통화 품질 향상을 얻기 위하여 명료도와 음질이 모두 높은 수준으로 유지되어야 하며 [2], 본 논문에서는 시간축 포락선을 강화하기 전에 피치를 추가로 강화하여 음질 저하가 최소가 되도록 한다.

3.2. 피치 강화 동작

CELP 기반의 음성 부호화기는 음질 향상을 위하여 음성 복원 이후에 피치 필터를 사용하여 피치 성분을 강화한다 [7][8]. 특히, VMR-WB는 피치 필터 앞에 대역통과 필터를 추가로 사용하여 저대역에만 피치 강화 동작을 적용하여 음질 향상 효과를 향상시킨다 [8]. 그러나 저대역 내부에서 세부 주파수 별로 서로 다른 강도의 피치 강화를 구현하는 것은 여전히 불가능하며, 하나의 피치 주기만을 처리하는 필터를 사용하므로 하모닉 피크가 피치 주파수의 정확한 정수배에 위치하지 않으면 잘못된 피치 강화 동작이 발생한다.

본 논문에서 제안하는 피치 강화 동작은 이와 같은 시간 영역에서의 문제점을 극복하기 위하여 주파수 영역 방법을 사용한다. 먼저, 전체 주파수 대역에 대하여 피치 주파수에 해당하는 대역폭을 가지면서 하모닉 피크가 대역 중심에 오도록 대역 분할을 한다. 다음, 각 피치 대역별로 하모닉 피크의 모양을 유지하면서 하모닉 피크와 하모닉 밸리 (valley) 영역의 상대적 차이를 확장하는 과정을 거친다. $\max_b$와 $\min_b$를 대역 b의 주파수 계수 크기 $|X_b[k]|$의 최대 및 최소라 할 때, 대역 b의 상대적 주파수 계수 크기는 식 (1)과 같이 정의된다.

$$|X_b'[k]|=\frac{\log\frac{|X_b[k]|}{\min_b}}{\log\frac{\max_b}{\min_b}},\quad 0.0\le|X_b'[k]|\le 1.0 \tag{1}$$

다음, 특정 함수 $W_b(\cdot)$를 상대적 주파수 계수에 적용하여 확장된 새로운 주파수 계수 $X_b^{new}[k]=W_b(|X_b'[k]|)\times X_b[k]$를 구한다. 함수 $W_b(\cdot)$는 단순증가 (monotonically increasing)하는 함수이고 $0<w_{\min}\le W_b(\cdot)\le 1.0$와 $W_b(1.0)=1.0$을 만족한다.

함수 $W_b(\cdot)$는 원하는 피치 강화량에 따라 정해지며, 각 프레임 및 피치 대역별로 서로 다르게 결정될 수 있으며, 이를 통하여 매우 유연하게 대역별 차별화된 피치 강화를 구현하게 된다. 예로, $w_{\min}$를 0.0 근처로 매우 작게 하면 하모닉 밸리 영역이 매우 작게 되어 강한 피치 강화가 이루어지고, 반대로 $w_{\min}$를 1.0 근처로 크게 하면 하

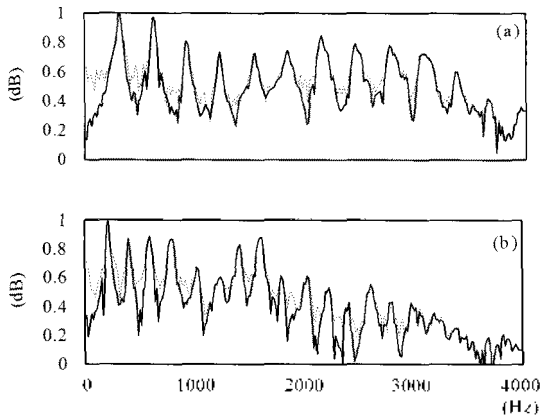

그림 4. 제안한 방법에 의하여 피치가 강화된 신호의 스펙트럼 (점선 : 원 신호, 실선 : 피치가 강화된 신호) (a) 여성 음성, (b) 남성 음성

Fig. 4. Spectrum of pitch-enhanced speech by the proposed method (dotted : original, solid : pitch-enhanced). (a) Female speech. (b) Male speech

모닉 밸리 변화가 거의 없으므로 매우 약한 피치 강화가 구현된다. 특히, 제안하는 방법은 하모닉 피크 위치가 아니라 주파수 계수의 상대적 크기를 기반으로 피치를 강화하므로 하모닉 피크가 정확한 위치에 오지 않더라고 정확한 피치 강화가 구현되는 장점을 가진다.

그림 4는 제안한 피치 강화 동작을 구현한 예를 보여준다. 여성과 남성 모두에 대하여 2 kHz 이하에서 하모닉 피크의 모양은 유지하면서 하모닉 밸리 영역이 3dB 감소하여 강한 피치 강화가 구현되고, 2 kHz 이상에서는 점진적으로 하모닉 밸리의 감소 정도가 줄어든다. 또한, 고대역에서 하모닉 피크가 정확한 위치에 오지 않더라고 원하는 피치 강화가 구현되는 것을 확인할 수 있다.

3.3. 시간축 포락선 강화 동작

시간축 포락선 강화 동작에서 가장 중요한 과정은 포락선의 시간축 변화량을 증폭시키는 것이다. 대역 b의 시간축 포락선을 $a_b[n]$라 할 때, 포락선의 시간 변화량은 $\Delta_b[n] = a_b[n]/a_b[n-1]$로 정의된다. $\Delta_b[n]$에 특정 함수 $g(\cdot)$를 적용하여 변화량 증폭을 구현하며, 그에 따라 증폭된 변화량 $g(\Delta_b[n])$을 얻는다. 마지막으로 변화량이 증폭된 포락선은 $a_b^{new}[n] = g(\Delta_b[n]) \times a_b^{new}[n-1]$이 되고, 해당 대역 신호의 원 신호가 $s_b[n]$라 할 때 증폭된 대역 신호는 $s_b^{new}[n] = (a_b^{new}[n]/a_b[n]) \times s_b[n]$이 된다.

본 논문에서는 $g(x) = |x|^p, p \geq 1.0$을 사용한다. $\Delta_b[n] > 1.0$가 되어 포락선이 증가하는 영역에서는 $g(\Delta_b[n]) > \Delta_b[n]$이 되므로 증가 속도가 더 커지고, 반면에 포락선이 감소하는 영역에서는 $g(\Delta_b[n]) < \Delta_b[n]$이 되어 감소 속도가 더 커진다. $g(\cdot)$에 포함되는 p 값은 포락선 강화 과정에서 강화레벨을 결정하는 파라미터의 역할을 하며, p 값이 커지면 더 많이 증폭된 변화량을 제공하여 포락선 강화량이 더 커진다.

$a_b^{new}[n]/a_b[n]$는 각 대역 신호 $s_b[n]$에 적용되는 이득의 역할을 수행한다. 일반적으로 사용하는 대역별 레벨 증폭 또는 고대역 증폭 과정을 사용하면 각 대역에 적용되는 이득값은 시간에 따라 일정하게 된다. 반면에, 제안하는 방법에서는 대역별 이득이 $a_b^{new}[n]/a_b[n]$가 되며 포락선의 시간축 모양 변화에 따라 매우 정교하게 결정되며, 명료도 향상의 측면에서 시간에 따라 가장 효과적인 동작을 할 수 있는 모양으로 정해진다. 이와 같이 시간에 따라 변하는 이득을 사용함으로써 신호 파형의 왜곡을 최소화 하고 레벨 증가를 억제하며, 이는 다시 음질의 저하를 최소로 하는데 많은 도움을 준다.

제안하는 시간축 포락선 강화 동작은 고대역에만 적용하며, 이는 저대역에 포락선 변형을 가하면 음질에 매우 민감한 성분이 변형되어 음질 저하가 크게 발생하기 때문이다. 기존 연구 결과에 의하면 명료도 향상을 위한 최적 고대역 통과 필터의 차단 주파수는 2 kHz이며 [9], 이 결과를 바탕으로 제안하는 포락선 강화 동작은 2 kHz 이상의 고대역에만 적용한다. 즉, 2 kHz 미만에는 $p = 1.0$을 사용하고 2 kHz 이상에 $p > 1.0$을 사용한다.

그림 5는 시간축 포락선 강화 동작의 예를 보여준다. (a)는 100 msec 길이의 '처'에 해당하는 한글 남성 음성 파형이고, (b)과 (c)는 각각 2 kHz와 2.3 kHz에 중심을 가지는 대역 신호 $s_b[n]$와 그의 시간축 포락선 $a_b[n]$을 보여준다. (d)와 (e)는 각각 $p = 1.1$와 $p = 1.2$를 사용하여 변화량을 증폭시킨 포락선 $a_b^{new}[n]$과 그에 대한 대역신호 $s_b^{new}[n]$를 보여준다. (f)와 (g)는 앞의 두 대역의 포락선 크기 비 $a_b^{new}[n]/a_b[n]$를 보여주며, 대역신호 $s_b[n]$에 적용되는 이득에 해당한다. 앞에서 언급하였듯이 시간에 따라 변하는 성질을 가지며, 이에 따라 명료도 향상에 필요한 성분만을 증폭시키는 역할을 수행한다. 마지막으로 (h)는 명료도가 향상된 최종 신호이며, 자음 영역과 고주파 성분의 레벨이 증가한 것을 볼 수 있다.

IV. 성능 평가

4.1. 평가 장치 및 방법

명료도 향상에 대한 정확한 주관적 성능을 평가하기

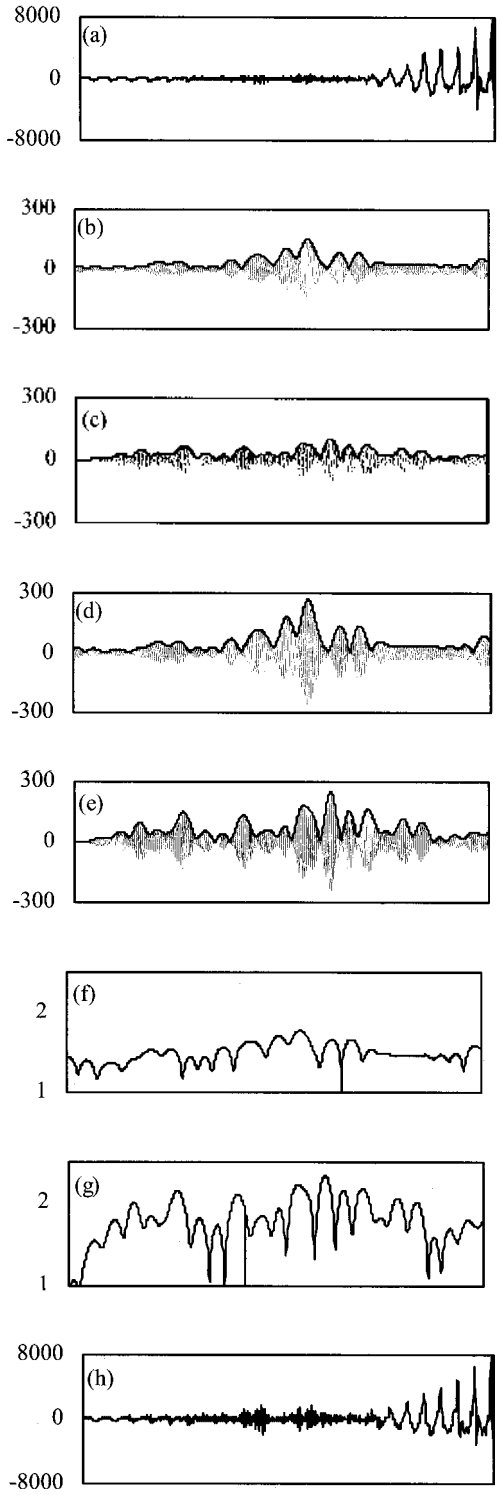

그림 5. 시간축 포락선 강화 동작의 예. (a) 100 msec 길이의 남성 음성 파형, (b)- (c) 2 kHz 및 2.3 kHz 대역 신호와 시간축 포락선, (d)- (e) 포락선이 강화된 신호, (f)- (g) 포락선 비, (h) 최종 명료도가 향상된 신호

Fig. 5. Example of the proposed temporal envelope enhancement. (a) 100msec-long male speech, (b)-(c) Band-passed signal centered at 2 kHz and 2.3 kHz and temporal envelope, (d)- (e) Envelope-enhanced signal, (f)- (g) Envelope ratio, (h) Final intelligibility-enhanced speech.

위하여 새로운 평가 방법을 제안한다. 잡음이 심한 환경에서 수화기를 사용하는 음성 통신의 청취 구조는 그림 6과 같다. $H_s(f)$, $H_1(f)$, $H_2(f)$는 각각 음성 출력 또는 잡음원으로부터 각 귀까지의 전달함수를 나타내고, 이때 음성을 청취하는 귀는 수화기에 의하여 외부 소리가 약간 차단되므로 오른쪽 귀에 입력되는 잡음 레벨은 왼쪽 귀에 입력되는 잡음 레벨에 비하여 감소한다. 이와 같은 상황을 정확히 포함하여 성능을 측정하는 것이 필요하지만, 실험의 재현을 위하여 장치의 규격화가 필요하고 통화 환경에 따라 변하는 전달함수를 성능 평가에서 그대로 반영하는 것은 현실적으로 불가능하므로, 그림 6의 구조를 현실에 맞도록 간단히 하여 성능 평가를 진행한다. 즉, point 잡음원이 청취자의 가운데 있다고 가정하여 $H_1(f) = GH_2(f), G < 1.0$로 간략화 하고, 전달함수의 차이를 무시하고 잡음과 음성 사이의 시간 지연은 명료도와 관련이 없으므로 잡음과 음성을 청취하는 전달함수를 동일하게 $H_2(f) - H_s(f)$로 설정한다.

이상과 같이 간략화된 청취 구조는 식 (2)의 스테레오 파일을 청취하는 것으로 구현될 수 있으며, 이를 통하여 최종 성능을 평가한다.

$$\begin{aligned} \text{Right Signal} &= s[n] + Gw[n], G < 1.0 \\ \text{Left Signal} &= w[n] \end{aligned} \tag{2}$$

물론 이와 같은 청취 방법은 많은 간략화 과정을 포함하여 실제 상황을 완벽하게 표현하지는 못하지만, 정확한 주관적 평가 방법의 핵심 사항인 양 귀로 청취하는 구조를 제대로 구현하므로 기존의 평가 방법에 비하여 보다 높은 신뢰도를 가지는 평가 결과를 제공하여 준다. G 값은 수화기의 구조에 따라 변하는데 본 논문에서는 $G = -6$ dB를 사용하며, 이 값은 다양한 종류의 휴대전화기를 사용하여 음성을 청취하는 귀에서 잡음이 차단되는 레벨을 실험적으로 구한 결과이다.

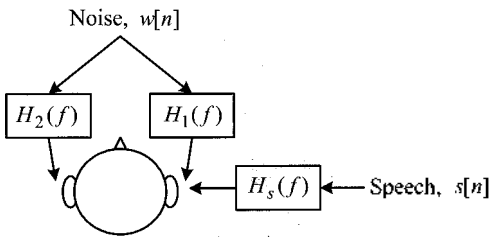

그림 6. 잡음 환경에서 수화기를 통하여 양 귀로 음성과 잡음을 청취하는 구조

Fig. 6. Binaural listening structure with a handset in noisy environments.

동일한 조건의 음성과 잡음에 대하여, 두 신호를 하나의 파일로 합하여 한 귀로 청취하여 측정한 명료도가 제안한 방법과 같이 양 귀로 청취하여 측정한 명료도에 비하여 높게 측정되는 것을 확인하였다. 이는 양 귀로 청취할 경우 마스킹 효과가 감소하여 잡음의 인지 효과가 증가하여 나타난 현상이다 [10]. 따라서 한 귀로 성능을 측정한 기존 논문의 결과는 실제 상황에서 사용자가 느끼는 명료도보다 과장된 성능이며, 정확한 성능 평가를 위하여 본 논문에서 제안하듯이 양 귀로 청취하는 구조가 반드시 필요하다.

잡음 환경에서의 명료도 향상 기술의 가장 대표적인 응용 분야가 디지털 이동통신 영역이므로 평가에 사용할 음성 신호는 한국 표준 음성 부호화기인 EVRC를 사용하여 복원한 신호로 제작한다 [7]. 음성 DB는 남성과 여성 각각 4명의 화자에 대하여 각 2 문장씩 총 16 개의 4 ~ 6초 한글 문장으로 구성되며, 잡음은 NOISEX DB의 factory noise와 babble noise를 사용하고, 평가자는 총 10명으로 구성한다.

명료도 측정은 식 (2)에 따라 음성과 잡음을 동시에 청취하여 진행하고, EVRC 출력 신호에 제안 기술을 적용한 후의 상대적 명료도를 표 1의 comparison scale로 평가한다. 명료도 평가는 일반적으로 단어에 대한 Diagnostic Rhyme Test (DRT) 방법을 사용한다 [11]. 그러나 음성통신에서 사용자가 실제로 느끼는 "통화 명료도"는 단어 Rhyme의 정확한 이해뿐만 아니라 문장 맥락을 기반으로 진행되는 전체 문장의 이해에 따라 결정된다. 따라서 음성통신에서 단어보다는 문장을 사용하는 평가가 보다 정확한 명료도를 측정한다고 판단되어 DRT 대신에 문장 청취를 통하여 주관적 명료도를 측정하였다 [12][13]. 음질 측정은 음성을 잡음 없이 단독으로 한 귀로 청취하여 진행하고, EVRC 출력 신호에 제안 기술을 적용한 후의 음질을 원 신호 (EVRC 입력 신호) 음질과 비교하여 표 1의 degradation scale로 평가한다.

표 1. 명료도와 음질의 주관적 평가 기준
Table 1. Performance scale for intelligibility and quality.

Comparison Scale		Degradation Scale	
3	Much Better	5	Degradation Not Perceived
2	Better	4	Degradation Perceived but Not Annoying
1	Slightly Better	3	Slightly Annoying
0	About the Same	2	Annoying
-1	Slightly Worse	1	Very Annoying
-2	Worse		
-3	Much Worse		

4.2. 평가 결과 및 분석

표 2에 명료도 측정 결과가 정리되어 있다. L0, L1, L2, L3은 각각 서로 다른 p 값을 가지는 시간축 포락선의 강화레벨을 나타낸다. L0은 $p = 1.0$를 사용하며 시간축 포락선 강화 없이 피치 강화만 적용한 경우에 해당한다. L1, L2, L3은 p의 최대값을 각각 1.05, 1.10, 1.20으로 한 경우이며, 각 경우에서 대역별 p 값은 최대값 이하에서 서로 다른 값을 가진다. 표 2에 의하면, 제안한 명료도 향상 기술에 의하여 잡음 종류, SNR, 성별 등에 관계없이 매우 강인하게 명료도가 향상되는 것을 확인할 수 있고, 강화레벨이 증가하면 명료도 향상도 증가하는 것을 알 수 있다.

음질 평가에 대한 결과는 표 3에 정리되어 있으며, EVRC 일은 원음에 대하여 EVRC 부호화기에 의한 음질 저하가 DMOS = 3.85 인 것을 나타낸다. L0에서 피치 강화를 통하여 음질이 크게 향상된 것을 볼 수 있고, L1까지 음질이 증가하다가 L1 이후에는 포락선 변형이 너무 심하여 음질이 저하되는 것을 알 수 있다. 그러나 L3에서도 DMOS = 3.45를 가져 실제 응용에 사용할 수 있는 음질을 유지하고 있다.

이상의 평가 결과에 의하면 제안한 기술을 사용하여

표 2. 제안한 기술에 대한 음성 명료도 성능 측정 결과 (B : babble noise, F : factory noise)
Table 2. Subjective performance of intelligibility by the proposed method.

		L0		L1		L2		L3	
Noise SNR (dB)		B	F	B	F	B	F	B	F
Male	5	0.18	0.06	0.56	0.40	1.01	0.57	1.82	1.72
	0	0.08	0.22	0.40	0.42	0.87	0.80	1.73	1.50
	-5	0.07	0.07	0.41	0.31	0.68	0.86	1.51	1.65
	-10	0.14	0.19	0.40	0.39	0.60	0.85	1.54	1.47
Female	5	0.14	0.05	0.50	0.40	1.08	0.81	1.93	1.69
	0	0.13	0.09	0.36	0.41	0.74	0.98	1.90	1.80
	-5	0.09	0.18	0.60	0.54	0.93	0.89	1.93	1.79
	-10	0.02	0.08	0.22	0.35	0.81	0.94	1.73	1.80
Ave.		0.11	0.12	0.43	0.40	0.84	0.85	1.76	1.68

표 3. 제안한 기술에 대한 음질 측정 결과
Table 3. Subjective performance of speech quality by the proposed method.

	EVRC	L0	L1	L2	L3
Male	3.91	4.14	4.18	4.20	3.43
Female	3.79	4.10	4.29	3.97	3.27
Ave.	3.85	4.12	4.24	4.09	3.45

명료도와 음질을 동시에 조정 가능하고, 강화레벨에 따라 명료도와 음질 사이의 뚜렷한 trade-off가 존재하는 것을 알 수 있다. 따라서 사용 환경 또는 사용자의 선호도에 따라 제안 기술의 최적 동작점을 정할 수 있다. 즉, 사용자가 느끼는 통합 품질은 명료도와 음질의 가중치 합으로 결정되는데, 각 사용 환경 및 사용자에 따라 각 가중치가 다르므로 각 상황에 최적인 명료도와 음질을 가지도록 강화레벨을 설정하는 것이 가능하다. 예로, 잡음이 매우 심한 환경에서는 명료도를 높여 정보 전달 성능을 높이는 것이 필요하므로 L3이 최적점이 되고, 잡음이 매우 적은 환경에서는 명료도 감소가 거의 없으므로 음질을 향상시켜 통화 품질을 향상시키는 L1이 최적점이 된다. 또는 사용자의 선호도에 따라 동작점 설정이 가능하며, 예로 음질보다는 정확한 정보 전달을 원할 경우 높은 명료도를 제공하는 L3이 개인 고유의 최적점이 된다.

V. 결 론

본 논문에서는 잡음 환경에서 음성 명료도를 향상시키는 새로운 기술을 제안하였다. 음성의 시간축 포락선이 명료도를 결정하는데 중요한 역할을 한다는 청각 이론에 따라 대역별 시간축 포락선을 강화하여 명료도를 향상시키는 새로운 개념을 도입하였고, 추가적인 음질 향상을 위하여 주파수 영역에서의 피치 강화 동작을 포함하였다. 또한, 정확한 명료도 성능을 평가하기 위하여 양 귀를 통하여 음성과 잡음을 청취하는 구조를 도입한 새로운 주관적 성능 평가 방법을 제안하였다. 제안한 성능 평가 방법에 따라 제안 기술의 성능을 평가하여 명료도와 음질이 모두 향상되는 것을 확인하였다.

제안하는 기술은 강화레벨에 따라 명료도 향상과 음질 향상 사이에 뚜렷한 trade-off를 가지며, 그에 따라 사용 환경 또는 사용자 선호도에 따라 제안 기술의 최적 동작점을 정할 수 있다. 이 기능을 활용하면 다양한 환경 및 다양한 성향의 사용자의 요구 조건을 만족시킬 수 있으며, 그에 따라 보다 넓은 영역에 응용이 가능할 것이다.

감사의 글

본 연구는 2007년 삼성전자의 연구비 지원으로 이루어졌습니다.

참 고 문 헌

1. B. Sauert and P. Vary, "Near end listening enhancement : speech intelligibility improvement in noisy environments," *ICASSP* 2006, pp.493-496, 2006.
2. J. Shin and N. Kim, "Perceptual reinforcement of speech signal based on partial specific loudness," *IEEE Signal Processing Letters*, 14(11), 2007.
3. P. Shankar and S. Park, "Speech intelligibility enhancement using tunable equalization filter," *ICASSP* 2007, pp.613-616, 2007.
4. B. C. J. Moore, *An introduction to the psychology of hearing*, 4th Ed., Academic Press, 1996.
5. R. Drullman, J. Festen and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoustical Society of America*, 95(2), Feb., 1994.
6. T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoustical Society of America*, 77(3), Mar., 1985.
7. TIA/EIA IS-127, "Enhanced Variable Rate Codec (EVRC), Speech Service Option 3 for Wideband Spread Spectrum Digital Systems," 1997.
8. 3GPP2 C.S0014-0, "Source-Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB)," 2004.
9. R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Trans. ASSP*, 24(4), 1976.
10. J. C. R. Licklider, "The Influence of Interaural Phase Relations upon the Masking of Speech by White Noise," *The Journal of the Acoustical Society of America*, 20(2), 1948.
11. W. D. Voiers, "Evaluating processed speech using the Diagnostic Rhyme Test (DRT)," *Speech Technology*, vol.1, 1983.

저자 약력

•윤 제 열 (Jae-Yul Yoon)
2007년 2월 : 광운대학교 전자공학과 (공학사)
2007년 3월~현재 : 광운대학교 전자공학과 박사과정
※주관심분야 : 음성/오디오 신호처리, 통신 신호처리

•김 중 회 (JungHoe Kim)
1998년 2월 : 고려대학교 전기공학과 (공학사)
2000년 2월 : 고려대학교 전기공학과 (M.S.)
2000년 3월~현재 : 삼성종합기술원 전문연구원
※주관심분야 : 음성/오디오 압축, 다채널 오디오 신호처리, 양자화 이론

•오 은 미 (Eunmi Oh)
1990년 2월 : 연세대학교 심리학과 (문학사)
1997년 12월 : Univ. of Wisconsin-Madison, 심리학과 (Ph.D.)
2000년 7월~현재 : 삼성전자 종합기술원 연구원
※주관심분야 : 심리음향학, 음성/오디오 신호처리

•박 호 종 (Hochong Park)
1986년 2월 : 서울대학교 전자공학과 (공학사)
1987년 12월 : Univ. of Wisconsin-Madison, 전자공학과 (M.S.)
1993년 5월 : Univ. of Wisconsin-Madison, 전자공학과 (Ph.D.)
1993년 9월~1997년 8월 : 삼성전자 선임연구원
1997년 9월~현재 : 광운대학교 전자공학과 교수
※주관심분야 : 음성/오디오 신호처리, 통신 신호처리, 영상 신호처리