전화망 환경에서 한국어 숫자음 인식을 위한 잡음처리

김 규 홍, 김 회 린
한국정보통신대학교 공학부

Noise Reduction for Korean Connected Digit Recognition through Telephone Channel

Kyuhong Kim, Hoirin Kim
School of Engineering, Information and Communications University
E-mail : {kkh, hrkim}@icu.ac.kr

Abstract

일반적으로 음성 인식에서의 성능은 잡음의 영향으로 인하여 저하된다. 전화망을 통한 한국어 연속 숫자음 인식은 음성인식 분야에 있어서 어려운 영역에 속하는데, 이는 조음 현상으로 인한 인식률 저하되는 점과 전화망 채널의 영향으로 인하여 스펙트럼 포락이 왜곡되며 음성신호의 대역폭이 제한되기 때문이다. 본 논문에서는 잡음의 영향을 줄이기 위하여, 2WF(2-stage Wiener Filter) 와 SWP (SNR-dependent Waveform Processing) 그리고 CMN(Cepstrum Mean Normalization)을 사용하였다. 2WF는 음성 신호의 포만트 구조를 적게 왜곡시키면서 전체적인 가산잡음 뿐만 아니라 동적 가산잡음도 줄여준다. SWP는 음성파형에서 SNR값이 상대적으로 큰 부분을 강조하여 전체적인 SNR을 향상시킬 수 있다. 또한, CMN은 특징벡터로부터 채널잡음의 영향을 정규화하여 음성 인식 성능을 향상시킨다. 이러한 방법들을 전화망 한국어 연속 숫자음 DB를 이용하여 실험한 결과, 음성신호의 왜곡을 최소화하면서 잡음의 영향을 줄여 전화망에서의 숫자음 인식 성능을 향상시킬 수 있었다.

I. 서 론

음성신호가 전화망을 통과하게 되면 음성신호의 대역폭은 제한되고 그 전화망 특성 때문에 포만트 구조에도 왜곡이 생긴다. 왜곡된 음성 신호는 음성인식에서 에러를 발생시킨다. 전화망을 통한 음성인식에서 잡음이 섞이는 양상을 그림 1에 표시하였다. 이때 잡음에 오염된 음성 신호는 다음 식 (1)과 같이 나타낼 수 있다.

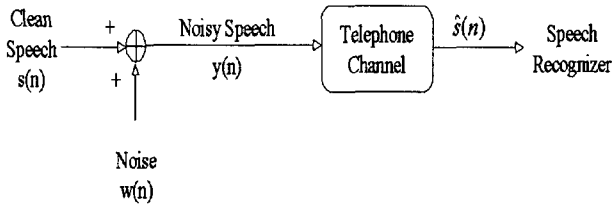

그림 1. 전화망을 통한 음성인식

$$\begin{aligned}\hat{s}(n) &= y(n) * h(n) \\ &= s(n) * h(n) + w(n) * h(n) \\ &= s(n) * h(n) + N(n)\end{aligned} \tag{1}$$

여기서 *는 convolution 연산자이고, h(n)은 전화망 채널의 전달함수이며, N(n)은 부가 잡음이다. 부가잡음 N(n)과 채널 전달 특성(h(n))의 영향을 줄이기 위하여 많은 기법들이 제안되어 왔지만, 여전히 명확한 해결 방법은 없다. 잡음 음성 인식에 있어서 노이즈 문제를 줄이기 위한 방식들이 여러 가지 있다. 그 중 한 가지는 음성신호의 상대적인 신호대 잡음비를 향상시키는 방법이다. 다른 한 가지는 특징 파라미터를 잡음에 둔감하도록 만드는 방법이다. 또 다른 방법은 음

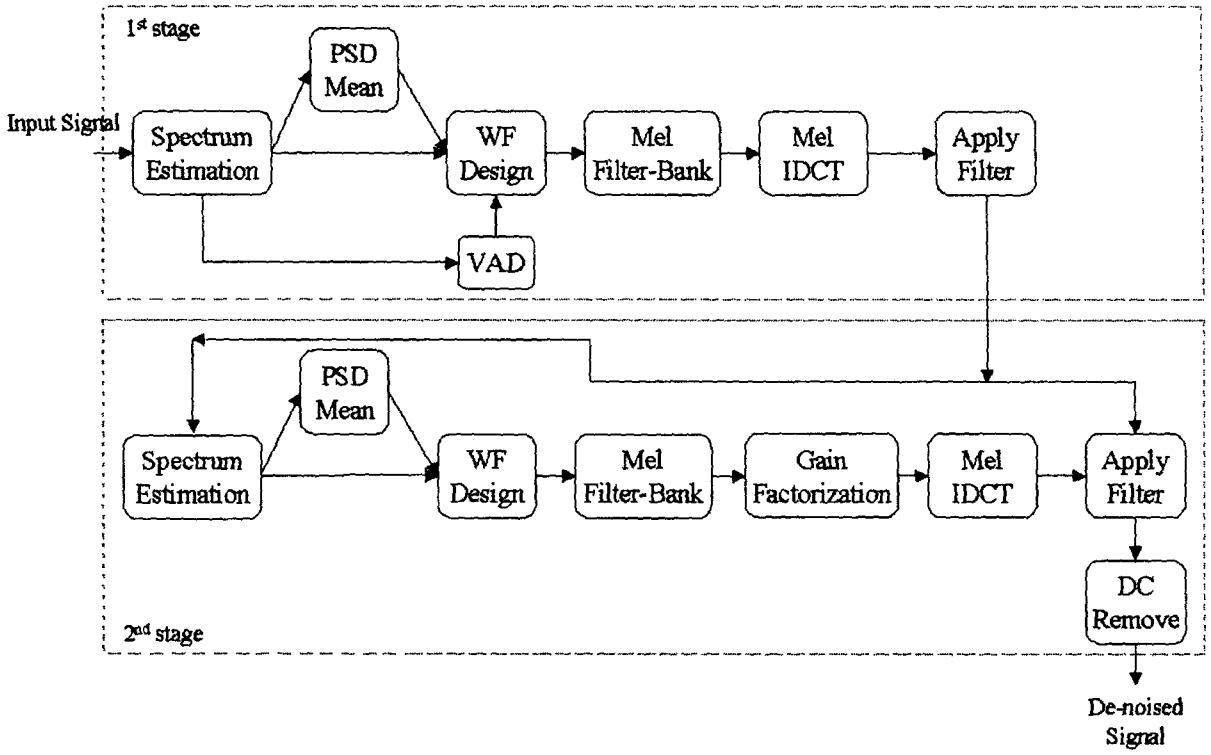

그림 2. 2-stage Wiener 필터의 블록도

향모델을 배경잡음에 적응하도록 만드는 방식이다. 본 논문에서는 잡음문제를 해결하기 위하여 2WF와 SWP 그리고 CMN 방법을 이용하여 전화망환경에서의 숫자음인식의 성능을 높였다.

본 논문의 구성은 다음과 같다. 잡음에 왜곡된 신호의 SNR을 향상시키기 위해 본 논문에서 사용한 2WF와 SWP에 대하여 2장과 3장에서 기술하고, 4장에서는 실험환경에 대하여 내용을 기술하고, 5장에서 결론을 맺는다.

II. 2WF (2-Stage Wiener Filter)

기존의 Wiener 필터의 단점들을 보완하면서, 잡음을 효과적으로 줄이기 위하여 본 논문에서는 음성인식의 전 처리 단계에서 2WF를 채택하였다 [1][2]. 이 필터는 크게 2개의 단계로 나누어진다. 첫 번째 단계에서는 전체적인 잡음을 줄여주며, 두 번째 단계에서는 동적으로 잡음을 제거해준다. 각 단계에서 필터의 전달함수는 다음 식(2)로 표현된다.

$$H(m) = \lambda(m)\left[\frac{(\hat{R}_y(m))^g - \rho\cdot(\hat{R}_n(m))^g}{(\hat{R}_y(m))^g}\right]^{\phi(m)} \quad (2)$$

여기서 $\hat{R}_y(m)$와 $\hat{R}_n(m)$는 각각 잡음음성과 잡음의 자기상관 추정치이고, m은 mel-index 이고 g는 root compression 상수, ρ는 0과 1사이의 상수이며, 모든 상수들은 실험에 의하여 결정된다. $\phi(m)$는 Wiener 필터 distortion을 제어하며, $\lambda(m)$은 유색잡음 억제를 위한 파라미터이다. 그림 2는 2WF의 블록도 이다. 노이즈제거 과정은 프레임단위로 이루어지는데, 각 프레임의 스펙트럼은 Spectrum Estimation 블록에서 계산된다. PSD Mean (Power spectrum Density Mean) 블록에서는 신호의 스펙트럼이 스무딩 된다. WF Design 블록에서는 현재 프레임의 스펙트럼 추정치와 노이즈 스펙트럼 추정치를 이용하여 주파수영역에서의 위너필터 계수를 구한다. 노이즈 스펙트럼은 VAD로 검출한 노이즈 프레임으로부터 추정되며, 위너필터 계수는 후에 멜 필터뱅크를 이용하여 스므딩되고, 이를 Mel-warped IDCT를 이용하여 위너필터의 전달함수가 계산된다. 첫 번째 단계에서 전체적인 노이즈가 제거된 후에 두 번째 단계 위너필터로 신호가 입력된다. 두 번째 단계에서는 Gain Factorization 에서 동적 잡음처리를 위한 이득(gain)이 계산되어 동적으로 잡음이 제거된다.

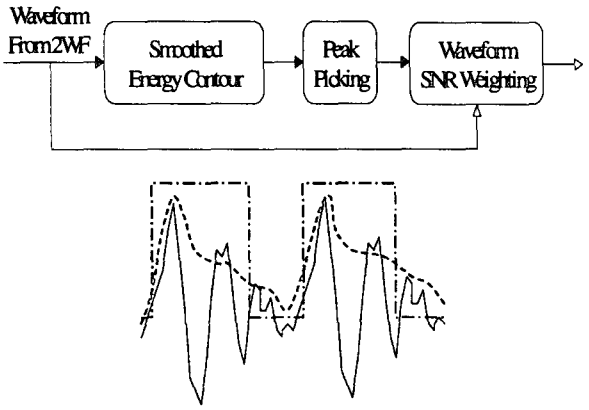

그림 3. SNR-dependent Waveform Processing

III. SNR-dependent Waveform Processing

유성음구간에서 상대적으로 SNR이 높은 구간의 신호 에너지를 높여주고, 상대적으로 낮은 SNR의 구간을 감쇠시켜주어 전체적인 SNR을 높이는 효과를 주는 것이 SWP이다 [2]. 전체적인 처리과정은 그림 3에 나타나있다. Teager에너지를 이용하여 점선에 해당하는 에너지 포락선을 구한 후, 간단한 peak-picking 알고리즘을 이용하여 에너지 포락의 피크를 찾는다. 이로부터 각 피크들 사이의 거리의 비율로 윈도우 함수 $\pi(t)$ 를 결정한다. 이 윈도우 함수를 기준으로 하여 높은 SNR을 갖는 부분과 낮은 SNR을 갖는 부분을 결정한다. 결과적으로 개선된 SNR을 갖는 음성신호는 다음 식 (3)에 의하여 계산된다.

$$\hat{s}(t) = \alpha \cdot \pi(t)s(t) + \beta(1-\pi(t))s(t)$$
$$\alpha \geq 1.0, \quad 0 < \beta \leq 1.0 \qquad (3)$$

여기서 s(t)는 2WF의 출력신호이고, α, β 는 각각 enhancing 상수와 attenuating 상수이다.

V. 실험 및 결과

실험에 사용된 음성 데이터는 전화망 4연속 숫자음 DB이다. 모든 음성데이터는 8kHz로 표본화되고, 16 비트로 양자화되어 있다. 실험에 사용된 총 문장 수는 60,064개 이며, 남녀 각각 37,554와 22,510 문장의 비율로 이루어져 있다. 음향모델을 만들기 위하여 전체 DB의 약 70%를 사용하였으며, 인식테스트에서는 나머지 30%를 사용하였다. 음성 데이터베이스는 BPF(Band Pass Filter)를 통과시켜 대역폭을 부가적으로 제한하였으며, 이때 BPF의 유효대역은 340Hz에서 3,400 Hz 이다. 다양한 신호대 잡음비를 갖는 데이터를 만들어 주기 위하여 백색잡음을 BPF에 통과시켜 노이즈로 사용하였으며, 이를 이용하여 5 dB, 10 dB, 15 dB의 SNR을 갖는 신호를 구성하였다. 프레임의 크기는 20 msec이고 10 msec씩 이동시키면서 잡음처리와 특징벡터 추출 작업을 하였다. 음성인식에 사용된 특징벡터는 MFCC (Mel-Frequency Cepstral Coefficients) 13차와 delta MFCC 그리고 acceleration MFCC를 구하여 총 39차를 이용하였다. 전화망 채널 노이즈의 영향에 의한 인식률 저하를 줄이기 위하여 모든 실험에서는 CMN(Cepstrum Mean Normalization)을 사용하였다. 실험에 사용된 인식기는 CHMM(Continuous HMM)을 기본으로 한 인식기이다 [4]. 음성인식 단위

표 1. 2WF와 SWP의 인식 성능 (%)

The performance of baseline					
	Clean	15 dB	10 dB	5 dB	Average
Clean Training	97.74	92.81	86.73	76.79	88.52
Multi-condition Training	97.55	95.13	92.19	87.12	93.00
The performance of 2WF					
	Clean	15 dB	10 dB	5 dB	Average
Clean Training	97.74	95.47	92.16	84.93	92.68
Multi-condition Training	97.60	96.49	94.75	90.71	94.89
The performance of 2WF and SWP					
	Clean	15 dB	10 dB	5 dB	Average
Clean Training	97.77	95.67	92.23	85.36	92.76
Multi-condition Training	97.53	96.55	94.82	91.15	95.00

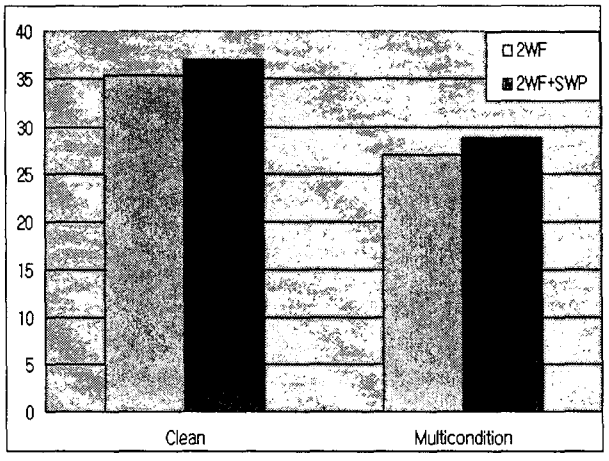

그림 4. Error reduction rate (%)

는 left-to-right tied-state 트라이폰을 사용하였으며, 각 state는 5개의 mixture로 구성되었다. 인식을 위한 음향모델은 clean 모델과 multi-condition 모델을 구성

하여 실험에 사용하였다. 여기서 clean 모델은 Gaussian 노이즈를 섞지 않은 데이터를 이용하여 학습시켰고, multi-condition 모델은 만들어놓은 모든 잡음 DB와 잡음을 섞지 않은 모든 데이터로 학습시켜 구성하였다. CMN만 이용하여 인식 성능을 계산한 경우와 CMN과 2WF를 이용하여 인식성능을 계산한 경우 그리고 CMN, 2WF, SWP 모두를 이용하여 계산한 경우를 각각 정리하여 표 1에 나타내었고, 이때의 ERR을 그림 4에 나타내었다. Baseline 시스템에서 잡음이 증가함에 따라 인식률이 급격히 감소되는데, 2WF와 SWP를 사용함에 따라서 인식률 저하가 개선되는 것을 보이고 있다. 또한, 잡음이 섞이지 않은 테스트 데이터로 실험한 경우에 2WF와 SWP를 사용함에 따라서 인식률이 저하되지 않는 특징을 보여주고 있는데, 이는 필터의 스펙트럼 왜곡으로 인한 성능저하가 없다는 것을 의미한다. Clean 훈련모델을 이용하여 인식률을 테스트한 경우는 기본시스템(CMN 만 사용)의 경우 약 88.5%의 인식률을 보여줬지만, 2WF와 SWP를 인식기의 전처리로 사용한 경우 약 92%이상의 성능을 보여줬다. 그림 4에 나타나 있듯이 인식테스트에서 음향모델로 clean 모델을 사용한 경우는 2WF와 2WF+SWP의 경우가 각각 ERR (Equal Reduction Rate)이 35%, 37%이고 , multi-condition 모델을 사용한 경우는 2WF와 2WF+SWP의 경우 각각 ERR이 27%, 29%였다.

VI. 결론

음성이 잡음에 의하여 왜곡될 때 음성인식기의 성능은 급격히 저하되며, 또한 한국어 연속 숫자음의 경우는 조음현상으로 인하여 음성인식 성능이 낮다. 본 논문에서는 전화망 환경에서 한국어 연속 숫자음 인식기의 성능은 향상시키기 위하여 2WF와 SWP을 이용하였으며, 다양한 SNR을 갖는 노이즈 DB를 구성하여 인식 성능 실험에 사용 하였다. 2WF와 SWP를 한국어 숫자음 인식에 적용한 결과 ERR이 clean, multi-condition 훈련 환경에서 각각 37%, 29%였다. 실험결과 2WF와 SWP이 유선 전화망을 통한 연속 숫자음 인식에서도 우수한 결과를 보여주었다.

참고문헌

[1] ETSI ES 202 050 V1.1.1, "Speech Processing, Transmission and Quality aspects; Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms", 2002.

[2] Y. M. Cheng, D. Macho, Y. Wei, D. Ealey, H. Kelleher, D. Pearce, W. Kushner and T. Ramabadran, " A robust front-end algorithm for distributed speech recognition", Proc. of Eurospeech2001, 425-429, 2001.

[3] R. Reddy, "Spoken Language Processing", CMU, 2000.

[4] Steve Young, "The HTK BOOK version 3.0", 2000.