강인한 음성인식을 위한 MMSE-STSA 기반 후처리 가중필터뱅크분석을 통한 특징추출

정 성 윤, 배 건 성
경북대학교 전자공학과

Feature Extraction through the post processing of WFBA based on MMSE-STSA for Robust Speech Recognition

Sungyun Jung, Keunsung Bae
Department of Electronics
yunij@mir.knu.ac.kr, ksbae@ee.knu.ac.kr

요약

본 논문에서는, 잡음음성에 강인한 음성인식을 위한 특징추출 방법을 제시한다. 제시한 방법은 2 단계 잡음제거 과정으로 구성되어 있다. 첫번째 단계는 MMSE-STSA 음성개선기법을 통해 잡음음성신호를 개선시키는 과정이고, 두 번째 단계는, MMSE-STSA 의 개선된 음성에 후처리 가중필터뱅크분석을 통해 잔여잡음의 영향을 감소시키는 과정이다. 제안한 방법의 성능평가를 위해, AURORA2 의 잡음음성 DB 중 테스트 집합 A 에 대해 인식실험을 수행하고, 결과를 기존 방법들과 비교, 검토한다.

1. 서론

음성인식시스템은 훈련과 테스트가 서로 다른 환경에서 수행될 때, 심각한 성능저하를 가져온다. 심지어 훈련과 테스트가 동일한 환경에서 수행된다 하더라도, 신호대 잡음비가 10dB 이하로 배경잡음이 첨가된 상황에서는 인식성능 향상을 기대하기가 어렵다[1]. 배경잡음에 의해 훼손된 음성신호들의 성능저하를 최소화 하기위해, 음성개선을 통해 성능향상을 이룩하는 방법들이 제안되었다[2,3].

음성개선 기법들 중에서, MMSE-STSA(Minimum Mean Square Error Short-Time Spectral Amplitude)추정치 기반의 음성개선기법은 다른 기법들에 비해 나은 성능을 나타낸다. 따라서, 본 논문에서는 강인한 음성인식을 위해, MMSE-STS 음성개선기법에 기반한 특징추출 방법을 제안한다. 제안한 방법은 2 단계의 잡음제거 과정으로 구성된다. 첫번째 단계는 MMSE-STSA 추정치 기반 음성개선기법을 통해 잡음음성신호를 개선시키는 과정이고, 두 번째 단계는, 개선된 음성에 후처리 가중필터뱅크분석(WFBA :Weighted Filter Bank Analysis)을 통해 잔여잡음의 영향을 감소시키는 과정이다. 이는 MFCC 특징추출의 필터뱅크처리 시, 각 필터뱅크 대역에서 추정잡음의 에너지에 비례하는 에너지만큼을 빼줌으로 수행된다.

본 논문의 구성은 다음과 같다. MMSE-STSA 음성개선기법 및 잔여잡음에 대해 설명하고, 3 장에는

제안한 특징추출방법을 설명한다. 그리고 4 장에서는 인식실험결과를 제시하고, 5 장에서 결론을 맺는다.

2. MMSE-STSA 추정치 기반의 음성개선 및 잔여잡음

잡음음성신호 $x(n)$ 의 k 번째 스펙트럼 X_k 는 (1)과 같이 원음성 스펙트럼 S_k 와 잡음 스펙트럼 V_k 의 합으로 표현된다.

$$X_k = S_k + V_k, \quad R_k e^{j\vartheta_k} = A_k e^{j\alpha_k} + V_k \tag{1}$$

음성과 잡음의 스펙트럼이 통계적으로 서로 독립인 가우시안 랜덤변수라고 가정하면, t 번째 프레임의 MMSE-STSA 추정치 $\hat{A}_k(t)$ 는 식 (2)로 주어진다[3]. 이때 사후 신호대잡음비(a posteriori SNR) $\gamma_k(t)$ 는 식 (3)과 같이 잡음음성과 추정된 잡음의 분산을 이용하여 직접 구하고, 사전 신호대잡음비(a priori SNR) $\xi_k(t)$ 는 식 (4)에 정의된 "Decision-Directed" 방법으로 구한다[3]. 식 (4)에서 $\lambda_{sk}(t)$ 와 $\lambda_{vk}(t)$ 는 각각 원음성과 잡음의 k 번째 전력 스펙트럼을 말하며, α 는 망각지수(forgetting factor)이고 $P[\cdot]$ 는 양의 값을 가지기 위한 연산자이다. 그리고 식 (2)에서 $\Gamma(\cdot)$ 는 Gamma 함수를 의미하고, $M(\theta)$ 는 식 (5)와 같이 정의되는 함수로서 $I_0(\theta)$ 와 $I_1(\theta)$ 는 각각 0 차와 1 차 modified Bessel 함수를 의미한다[3].

$$\begin{aligned}\hat{A}_k(t) &= E\{A_k(t) \mid X_k(t)\} \\ &= \Gamma(1.5)\cdot\frac{\sqrt{\frac{\xi_k(t)\cdot\gamma_k(t)}{1+\xi_k(t)}}}{\gamma_k(t)}\cdot M\left(\frac{\xi_k(t)\cdot\gamma_k(t)}{1+\xi_k(t)}\right)\cdot R_k(t) \\ &= G_{MMSE}\left(\xi_k(t),\gamma_k(t)\right)\cdot R_k(t)\end{aligned} \tag{2}$$

$$\gamma_k(t) \equiv \frac{R_k^2(t)}{\lambda_{vk}(t)} \tag{3}$$

$$\xi_k(t) \equiv \frac{\lambda_{sk}(t)}{\lambda_{vk}(t)} = \alpha\frac{\hat{A}_k^2(t-1)}{\lambda_{vk}(t-1)} + (1-\alpha)P\left[\gamma_k(t)-1\right] \tag{4}$$

$$M(\theta) = e^{-\frac{\theta}{2}}\cdot\left[(1+\theta)\cdot I_0\left(\frac{\theta}{2}\right)+\theta\cdot I_1\left(\frac{\theta}{2}\right)\right] \tag{5}$$

MMSE-STSA 추정치에 음성존재확률(SPP: Speech Present Probability)을 도입하면 수정된 MMSE-STSA 추정치를 얻을 수 있다[3]. 본 논문에서는 음성존재확률을 적용한 MMSE-STSA 추정기법을 사용하였다.

그림 1 은 잡음음성과 MMSE-STSA 추정치에 의해 개선된 음성을 나타낸 것이다. 그림에서 개선된 음성에 여전히 고려해야 할 정도의 잔여잡음이 존재함을 확인할 수 있다.

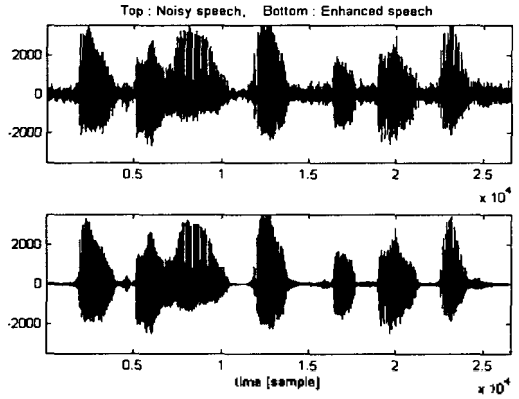

그림 1. Babble 잡음이 부가된 경우의 잡음음성(위)과 개선된 음성(아래)의 예 (SNR=10dB)

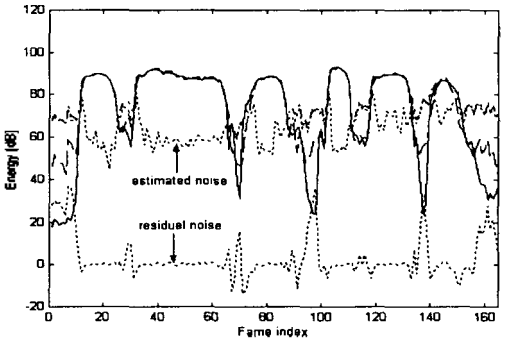

그림 2. MMSE-STSA 추정치 기반 잡음제거 후의 깨끗한 음성(굵은실선), 개선된 음성(dashed), 추정잡음(점선), 그리고 잔여잡음(아래 점선)의 시계열

그림 2 는 깨끗한 음성, MMSE_STSA 에 의해 개선된 음성, 추정잡음, 잔여잡음에 대한 로그필터뱅크에너지의 시계열 특성을 나타낸 것이다. 그림에서 잔여잡음이 음성의 시작과 끝부분의 묵음영역과 음성사이의 짧은 묵음구간에서 큰 값을 갖는 반면, 음성에너지가 큰 구간, 즉 높은 SNR 영역에서는 작은 값을 가짐을 알 수 있다. 또한, 추정잡음이 잔여잡음의 묵음구간에서 유사한 패턴을 나타내고 있음을 알 수 있다.

3. 제안한 특징추출 방법

앞 장에서 개선된 음성에 존재하는 잔여잡음이 개선된 음성신호의 에너지가 낮은 대역에 많은 영향을 끼치는 것을 보였다. 특징추출 시, 잔여잡음의 영향을 감소시키기 위해, 본 논문에서는 MMSE-STSA 추정치 기반 음성개선기법과 가중 필터뱅크 분석을 사용한 특징추출 방법을 제안한다. 그림 3 에 제안한 방법의 전체 블록도를 나타내었다. 먼저, 잡음음성 $y(m)$ 은 MMSE-STSA 추정치 기법을 적용 후, 개선된 음성 $\hat{x}(m)$ 과 잡음음성으로부터 개선된 음성을 뺀 추정잡음 $\hat{n}(m)$ 으로 분리된다. 추정잡음 $\hat{n}(m)$ 에 대한 필터뱅크 분석을 통해, 각 필터뱅크의 대역에 대해 가중치 λ_i 를 계산한다. 가중치가 계산되면, 가중필터뱅크 분석을 통해 잡음에 강인한 MFCC 특징파라미터가 추출된다.

그림 4 는 추정잡음으로부터 가중치를 계산하기위해, 필터뱅크분석의 블록도를 나타낸 것이다. 추정잡음 프레임 $\hat{n}(m)$, $1 \le m \le N$ 은 단구간 푸리에 변환을 통해 주파수 영역으로 변환된 후, 전력스펙트럼 $P_{\hat{N}}(k)$ 가 계산된다. 일단, 각 프레임에 대한 전력스펙트럼이 구해지면, i 번째 멜스케일 대역통과필터 $\Phi_i(k)$ 를 통해 필터뱅크 에너지 e_i 가 구해진다. 필터뱅크 에너지 e_i 에 로그를 취한 후, 가중치, λ_i 를 식 (6)과 같이 구한다.

$$\lambda_i = \frac{\log(e_i)}{\sum_{j=1}^{T} \log(e_j)} \qquad (6)$$

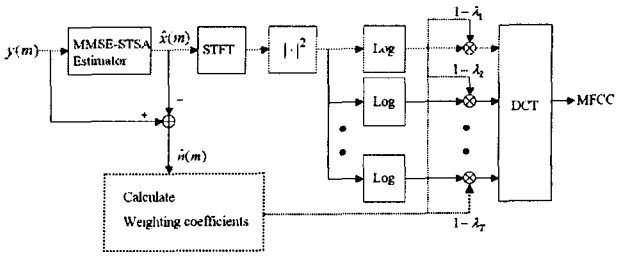

그림 3. 제안한 특징추출 방법의 블럭도

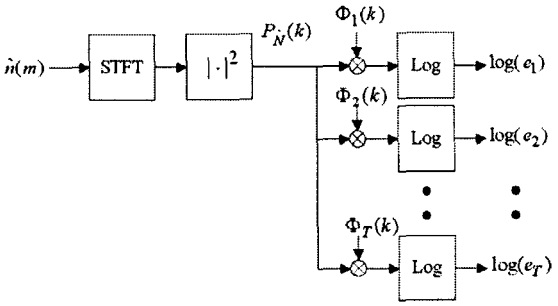

그림 4. 추정잡음으로부터 가중치 계산을 위한 필터뱅크 분석

식 (6)에서 T 는 멜스케일 대역통과필터의 수를 의미한다. 일단, 각 대역에 대한 가중치가 결정되면, 로그필터뱅크에너지 영역에서 잔여잡음의 영향을 줄이기 위해 가중필터뱅크 분석이 식 (7)와 같이 수행된다.

$$Weighted\ \ LogFBE_i = (1-\lambda_i) \cdot LogFBE_i \qquad (7)$$

여기에서 $LogFBE_i$ 는 개선된 음성의 i 번째 필터뱅크 에너지를 나타낸다. 최종적으로, $Weighted\ \ LogFBE_i$ 에 DCT(Discrete Cosine Transform)를 적용하여, 강인한 음성인식을 위한 MFCC 를 추출한다.

4. 실험결과

제안된 특징추출방법은 ETSI 에서 표준으로 제시한 Aurora2 DB 를 사용하여 실험되었다. 훈련은 Clean condition 으로 수행하였고, 테스트 집합 A 에 대해 인식테스트를 행하였다. 테스트 집합 A 는 4 가지 잡음(subway, babble, car, exhibition)에 대해, 20dB 부터 -5dB 까지 6 가지 잡음레벨에 대한

잡음음성으로 구성된다. 기본 음성인식기는 HTK 를 사용하였고, 16 개의 state 와 2 개의 묵음모델을 가진 CHMM 을 이용하였다. 특징파라미터는 23 차 필터뱅크를 이용한 12 차 MFCC, 1 차 로그에너지, 그리고 각각의 delta 및 acceleration 을 포함한 총 39 차로 구성되었다. 표 1 은 Baseline 의 Word Accuracy 를 인식률로 나타낸 것이고, 표 2 는 MMSE-STSA 음성개선기법에 대한 인식률, 그리고, 표 3 은 제안한 특징추출 기법을 적용한 인식결과를 나타낸 것이다. 실험결과, 제안한 방법이 Baseline 보다 약 8.4%, MMSE-STSA 보다 약 1.8%의 평균 성능향상을 나타내었다. 이는 Baseline 인식률을 기준으로 상대적인 WER(Word Error Rate)감소(R)가 각각 16.11% 20.44% 향상됨을 의미한다. 여기서 사용된 상대적인 WER 감소(R)는 식 (8)로 정의 된다

$$R = 100\% - \frac{WER_{method2}}{WER_{method1}} \times 100\% \qquad (8)$$

5. 결론

본 논문에서는 강인한 음성인식을 위해, MMSE-STSA 추정치에 기반한 후처리 가중필터뱅크분석을 통한 특징추출 방법을 제안하였다. Aurora2 DB 를 사용한 인식실험 결과, 제안한 방법이 더 나은 성능향상을 보였다.

SNR	Types of noise				Avg.
	Subway	Babble	Car	Exhibition	
Clean	98.93	99.00	98.96	99.20	99.02
20 dB	97.05	90.15	97.41	96.39	87.29
15 dB	93.49	73.76	90.04	92.39	95.24
10 dB	78.72	49.43	67.01	75.66	67.62
5 dB	52.20	26.81	34.09	44.83	39.39
0 dB	26.01	9.28	14.46	18.05	16.90
-5 dB	11.18	1.57	9.39	9.60	7.92
Avg.	65.37	50.00	58.77	62.25	59.05

표 1. Baseline 에 대한 인식결과 [%]

SNR	Types of noise				Avg.
	Subway	Babble	Car	Exhibition	
Clean	98.28	98.22	98.30	98.43	96.93
20 dB	92.72	87.45	94.78	90.10	91.28
15 dB	89.16	77.66	91.74	85.50	86.02
10 dB	81.95	62.55	85.71	77.35	73.35
5 dB	67.61	39.84	71.43	59.95	59.72
0 dB	42.74	18.23	41.96	38.29	35.28
-5 dB	16.06	3.69	12.79	15.83	12.06
Avg.	69.79	55.38	70.96	66.49	65.65

표 2. MMSE-STSA 개선기법에 대한 인식결과 [%]

SNR	Types of noise				Avg.
	Subway	Babble	Car	Exhibition	
Clean	98.16	98.67	98.36	98.49	97.16
20 dB	93.55	89.78	95.14	91.18	92.44
15 dB	90.64	81.74	92.81	87.94	88.29
10 dB	84.53	66.51	87.29	80.44	76.31
5 dB	70.83	43.59	73.78	68.22	62.75
0 dB	46.24	19.92	44.68	40.02	37.69
-5 dB	18.88	3.05	13.60	15.71	12.77
Avg.	71.83	57.54	72.24	68.07	67.42

표 3. 제안한 특징추출방법에 대한 인식결과 [%]

본 연구는 한국과학재단 목적기초연구(R01-2003-000-10242-0)지원으로 수행되었음.

참고문헌

[1] Acero, A., *Acoustical and environmental robustness in automatic speech recognition*, Norwell, M.A.Kluwer, 1993

[2] S.F.Boll, "Supression of acoustic noise in speech using spectral subtraction," IEEE Trans.Acoust. Speech Signal Process., vol.ASSP-27, pp.113-120, 1992

[3] Y. Ephraim and D.Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-32, no.6, pp.1109-1121, Dec.1984