신호의 주기성에 따라 변형되는 스펙트럼 차감을 이용한 CMSBS

이우영*, 이상호*, 홍재근*
*경북대학교 전자전기컴퓨터학부
e-mail:wylee@ee.knu.ac.kr

CMSBS Extraction Using Periodicity-based Mel Sub-band Spectral Subtraction

Woo-Young Lee*, Sang-Ho Lee*, Jae-Keun Hong*
*School of Electrical Engineering and Computer Science,
Kyungpook National University

요 약

현재 음성인식에서 가장 많이 사용하고 있는 특징벡터는 MFCC(Mel-Frequency Cepstral Coefficients)이다. 그러나 MFCC도 잡음이 존재하는 환경에서는 인식 성능이 저하된다. 이러한 MFCC의 단점을 해결하기 위해 mel sub-band 스펙트럼 차감법과 신호대잡음비에 따른 에너지 압축을 이용하는 CMSBS(Compression and Mel Sub-Band Spectral subtraction) 방법을 사용한다. 본 논문에서는 CMSBS 방법 적용 시 음성이 발성되는 구간과 묵음 구간에서 mel sub-band 스펙트럼 차감법이 동일한 조건으로 이루어져 발생하는 중요한 음성정보의 손실을 보완하기 위하여 신호의 주기성을 이용하여 spectral flooring 파라미터를 변형하는 방법을 제안한다. 제안한 방법으로 실험을 한 결과 잡음이 거의 없는 음성신호에 대해서는 기존의 방법과 비슷한 인식률을 가지고, 잡음성분이 많을수록 변형된 mel sub-band 스펙트럼 차감법을 적용한 방법이 인식률에서 보다 높은 성능 향상을 가져왔다.

1. 서론

음성인식에 있어서 잡음에 강인한 인식 시스템의 필요성이 많이 대두되고 있고, 연구가 활발히 진행되고 있다. 현재 음성인식에서 가장 많이 사용하고 있는 특징벡터는 MFCC이다. 그러나 MFCC도 잡음이 존재하는 환경에서는 인식 성능이 저하된다. 이러한 MFCC의 단점을 해결하기 위해 Babak Nasersharif 는 mel sub-band 스펙트럼 차감법과 신호대잡음비에 따른 에너지 압축을 사용하는 CMSBS[1] 방법을 제안하였다. 그러나 CMSBS 방법을 적용할 때, 음성이 발성되는 구간과 묵음 구간에서 mel sub-band 스펙트럼 차감법이 동일한 조건으로 이루어지므로 음성이 발성되는 구간의 필터뱅크 에너지가 추정된 잡음신호의 필터뱅크 에너지보다 작은 경우, 해당 필터뱅크 에너지에 아주 작은 값의 spectral flooring 파라미터를 곱하게 되어 중요한 음성정보가 손실되는 단점이 있다.

본 논문에서는 각각의 구간에서 음성의 주기성(periodicity)의 정도를 파악하여 주기성이 약한 경우에는 음성정보의 존재 가능성이 낮다고 판단하여 spectral flooring 파라미터를 작게 설정하고, 주기성이 약하지 않은 경우에는 중요한 음성정보의 존재 가능성이 높다고 가정하여 주기성의 정도에 따라 spectral flooring 파라미터를 크게 설정하여 중요한 음성정보의 손실을 방지하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 CMSBS 특징벡터 추출 방법에 대해 소개한다. 3장에서는 신호의 주기성에 따라 변형되는 mel sub-band 스펙트럼 차감법을 사용한 CMSBS 특징벡터 추출 방법을 제안하고, 4장에서 실험 및 결과를 보인 후 5장에서 결론을 맺는다.

2. CMSBS

MFCC는 주파수 영역에서의 인간의 청각적 특성을 고려해 구성된 필터뱅크(filter bank)의 대수(log) 에너지를 DCT(discrete cosine transform)하여 구한다. 하지만, 잡음 환경에서 좋지 않은 인식 성능을 가지는 MFCC의 단점을 해결하기 위해 mel sub-band 스펙트럼 차감법과 신호대잡음비에 따른 에너지 압축을 이용하는 CMSBS 방법이 사용된다.

CMSBS를 이용한 특징벡터 추출 방법은 크게 두 단계로 나뉘어진다. 첫째, 잡음의 영향으로 증가된 필터뱅크 에너지를 줄이기 위해 mel sub-band 스펙트럼 차감을 하는 단계이다. 둘째, 잡음과 왜곡의 영향을 적게 받은 필터뱅크 에너지를 강조하기 위해 각각의 대역통과 된 에너지의 신호대잡음비에 따라 스펙트럼을 압축하는 단계이다.

2.1 mel sub-band 스펙트럼 차감법

CMSBS를 이용한 특징벡터 추출 방법은 일반적인 full-band 스펙트럼 차감법[2]보다 성능이 우수한 mel sub-band 스펙트럼 차감법[3]을 이용한다. mel sub-band 스펙트럼 차감법은 식 (1)과 같이 나타낸다.

$$E_i^{SS}=\begin{cases} E_i^X-\alpha_i E_i^N & \text{if. } E_i^X > \dfrac{\alpha_i}{1-\beta_i}E_i^N \\ \beta_i E_i^X & \text{otherwise} \end{cases} \quad (1)$$

여기서, E_i^{SS}는 mel sub-band 스펙트럼 차감법 후의 보상된 i번째 필터뱅크 에너지를 나타내고, α_i와 β_i는 각각 over-estimation 파라미터와 spectral flooring 파라미터를 나타낸다.

2.2 신호대잡음비에 따른 mel sub-band 에너지 압축

CMSBS를 이용한 특징벡터 추출 방법은 각각의 mel sub-band마다 신호대잡음비를 고려하여 스펙트럼 압축의 정도를 조절함으로써 중요한 음성정보의 손실을 줄인다. 압축 파라미터인 w_i는 식 (2)와 같이 구한다.

$$w_i=\gamma \cdot \left[1-\exp\left(-\frac{SNR_i}{\xi_i}\right)\right] \quad (2)$$

여기서, γ는 상수이고, SNR_i는 i번째 mel sub-band의 신호대잡음비를 나타내며, ξ_i는 압축 파라미터 w_i의 가파른 정도를 조절해 주는 파라미터로 식 (3)과 같이 구한다.

$$\xi_i=1-\frac{1}{1+\exp\left(-\dfrac{SNR_i-\mu_{SNR}}{\sigma_{SNR}}\right)} \quad (3)$$

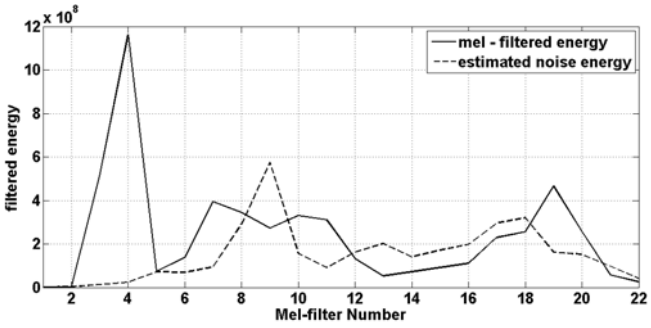
여기서, μ_{SNR}과 σ_{SNR}은 각 프레임의 모든 mel sub-band로부터 계산된 신호대잡음비의 평균과 표준편차를 나타낸다.

앞서 구한 E_i^{SS}를 식 (4)와 같이 각 sub-band의 신호대잡음비에 따른 압축 파라미터 w_i를 이용하여 제곱근을 취하고, 이를 이산 코사인 변환을 수행하여 CMSBS 방법에 의한 특징벡터 c_k를 구한다.

$$c_k=\sum_{i=1}^{M}\left(E_i^{SS}\right)^{w_i}\cos\left[\frac{\pi k(i-0.5)}{M}\right],\ 1\le k\le p \quad (4)$$

3. 변형된 mel sub-band 스펙트럼 차감법을 사용한 CMSBS

CMSBS 방법은 잡음의 영향으로 증가된 필터뱅크 에너지를 줄이고, 잡음과 왜곡의 영향을 적게 받은 필터뱅크 에너지를 강조하는 장점이 있지만, 잡음신호가 음성신호 전 구간에서 정적(stationary)이라고 가정하는 스펙트럼 차감법은 실생활 잡음(real noise)처럼 시간에 따라 임의로 변하는 환경에서는 적절하게 대응하지 못한다. 또한, 음성이 발성되는 구간과 묵음 구간에서 mel sub-band 스펙트럼 차감법이 동일한 조건으로 이루어져 그림 1과 같이 음성이 발성되는 구간의 필터뱅크 에너지가 추정된 잡음신호의 필터뱅크 에너지보다 작은 경우, 식 (1)에 의해 아주 작은 값의 spectral flooring 파라미터를 곱하여 중요한 음성정보가 손실되는 경우가 발생한다.



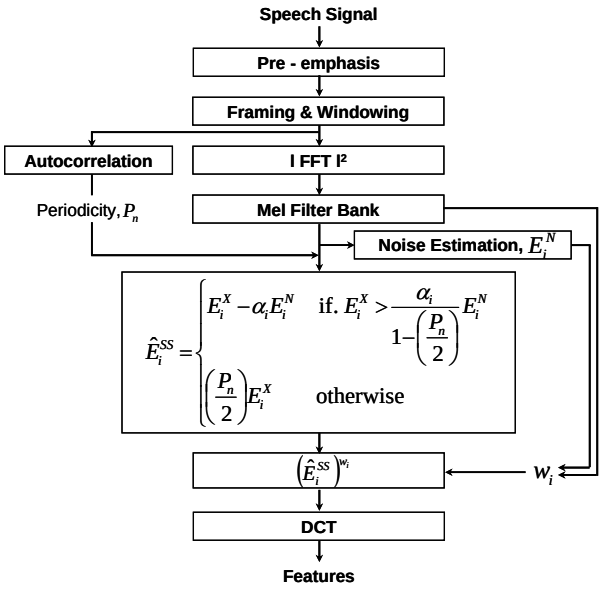
[그림 1] 추정된 잡음신호와 음성이 발성되는 구간의 필터뱅크 에너지

[표 1] 추정된 잡음 신호의 sub-band 에너지가 각 프레임의 sub-band 에너지보다 큰 경우의 수

periodicity	프레임의 수	sub-band의 수	$E_i^N > E_i^X$의 수
0 ~ 0.2	60	1320	674
0.2 ~ 0.4	47	1034	526
0.4 ~ 0.6	11	242	77
0.6 ~ 0.8	107	2354	431
0.8 ~ 1	96	2112	475
Total	321	7062	2183

표 1은 AURORA2 데이터베이스의 한 utterance에서 추정된 잡음 신호의 sub-band 에너지 E_i^N이 E_i^X보다 큰 경우의 수를 주기성에 따라 나타낸 것이다. 표 1에서와 같이, 주기성이 강한 프레임에서도 잡음 신호의 sub-band 에너지가 각 프레임의 sub-band 에너지보다 큰 경우가 많이 발생함을 알 수 있다.

본 논문에서는 음성 정보의 손실을 최소화 하면서 배경잡음을 감쇄하기 위해 신호의 주기성에 따라 β_i를 변형하는 방법을 제안한다. 변형된 mel sub-band 스펙트럼 차감법을 사용한 특징벡터 추출 방법의 전체 블록다이어그램은 그림 2와 같다.



[그림 2] 변형된 mel sub-band 스펙트럼 차감법을 이용한 CMSBS 특징벡터 추출

일반적으로 실생활 잡음은 그 주기성이 약한 특성을 가지므로 각 프레임에서 음성의 주기성의 정도를 파악하여 해당 프레임 내에 중요한 음성정보의 존재 가능성을 판단한다. 즉, 주기성이 아주 약한 경우에는 음성정보의 존재 가능성이 낮다고 판단하여 β_i를 작게 설정하고, 주기성이 약하지 않은 경우에는 음성이 발성되고 있는 프레임으로서 중요한 음성정보의 존재 가능성이 높다고 가정하여 주기성의 정도에 따라 β_i를 크게 설정하여 중요한 음성정보의 손실을 방지한다.

자기상관계수 열의 주기성의 정도를 나타내는 인자 P_n은 다음 식 (5)와 같이 각 프레임의 자기상관계수 $r_n(0)$와 $r_n(l)$을 이용하여 구한다.

$$P_n = \frac{r_n(l)}{r_n(0)}, \qquad 0 \le P_n \le 1 \tag{5}$$

여기서, $r_n(l)$은 n번째 프레임에서 자기상관계수 열의 두 번째 극대값을 나타낸다. 이렇게 구해진 신호의 주기성 P_n을 이용하여 β_i를 식 (6)과 같이 변형한다.

$$\beta_i = \frac{P_n}{2} \tag{6}$$

주기성에 따라 변형되는 β_i를 이용한 mel sub-band 스펙트럼 차감법을 적용하여 개선된 필터뱅크 에너지 $\hat{E}_i^{SS}$ 를 식 (7)을 이용하여 구한다.

$$\hat{E}_i^{SS} = \begin{cases} E_i^X - \alpha_i E_i^N & \text{if. } E_i^X > \dfrac{\alpha_i}{1-\left(\frac{P_n}{2}\right)} E_i^N \\ \left(\dfrac{P_n}{2}\right) E_i^X & \text{otherwise} \end{cases} \tag{7}$$

이렇게 구한 $\hat{E}_i^{SS}$를 식 (4)의 E_i^{SS} 대신 사용함으로써 제안된 방법에 의한 특징벡터를 구한다.

4. 실험 및 결과

본 논문에서는 음성인식 성능의 평가를 위해 European Telecommunications Standards Institute (ETSI) STQ-AURORA DST Working Group[4] 에서 제시한 AURORA2 데이터베이스를 사용하였다. AURORA2 데이터베이스는 8㎑ 표본화 주파수에 표본 당 16 bits로 양자화 된 음성으로 구성되어 있고, 세 가지 테스트음성 집합과 두 가지 훈련음성 집합인 clean과 multicondition을 포함하고 있다. 본 논문에서는 훈련음성 집합으로 clean 조건의 데이터베이스를 사용하고 테스트 음성으로는 AURORA2 Set A로 6가지 신호대잡음비 레벨(20dB, 15dB, 10dB, 5dB, 0dB, -5dB)의 Subway, Babble, Car, Exhibition 잡음을 첨가하여 만들어진 테스트음성 집합을 사용하였다.

특징벡터 추출 과정은 공통적으로 한 프레임을 32msec로 정하고 한 프레임 길이의 해밍 창함수를

곱하며, 프레임 이동률은 10msec로 한다. 분석 프레임에 256포인트 FFT를 취하여 파워 스펙트럼을 구한다. 구해진 파워 스펙트럼은 22개의 멜 스케일 대역 필터로 구성된 필터뱅크를 적용하여 각 대역별 에너지를 구한다. 이 에너지를 특징벡터 추출 방법에 따라 대역별 대수 또는 스펙트럼 차감 후 제곱근 에너지를 계산하여 이산 코사인 변환을 통해 대수(log) 에너지를 포함한 13차의 특징벡터를 얻는다. 그리고 구해진 특징벡터의 delta 성분과 acceleration 성분을 구하여 최종적으로 분석 프레임 당 총 39차의 특징벡터를 구한다.

기본 인식 시스템은 AURORA2 에서 제공하는 AURORA2-HTK를 사용하였다. 단어모델은 one, two, three. four, five, six, seven, eight, nine, zero, oh의 11개로 정의되어 있고 각 단어모델은 16개의 상태(state)와 상태 당 3개의 mixtures로 구성된 HMM으로 이루어져 있다. 11개의 단어모델 외에 2개의 묵음 모델이 있는데, 각각 3상태와 1상태 모델로 구성되어 있다.

[표 2] AURORA2 set A 잡음음성에서 MFCC 방법의 인식실험 결과

	Subway	Babble	Car	Exhibition	Avg.
Clean	99.42	99.33	99.37	99.57	99.42
20dB	98.34	98.52	98.63	97.93	98.36
15dB	95.79	95.71	97.08	95.50	96.02
10dB	86.80	84.25	87.44	86.98	86.37
5dB	64.48	60.55	56.07	62.70	60.95
0dB	33.34	33.71	18.37	27.95	28.34
-5dB	12.80	17.41	8.98	10.52	12.43
Avg.	70.14	69.93	66.56	68.74	68.84

[표 3] AURORA2 set A 잡음음성에서 CMSBS 방법의 인식실험 결과

	Subway	Babble	Car	Exhibition	Avg.
Clean	99.48	99.27	99.43	99.38	99.39
20dB	98.07	97.76	98.39	96.73	97.74
15dB	95.86	93.02	97.11	93.37	94.84
10dB	88.33	80.35	90.69	82.23	85.40
5dB	71.66	56.41	70.5	57.95	64.13
0dB	44.18	31.89	39.31	32.18	36.89
-5dB	25.48	18.02	19.8	16.78	20.02
Avg.	74.72	68.10	73.60	68.37	71.20

[표 4] AURORA2 set A 잡음음성에서 제안한 방법의 인식실험 결과

	Subway	Babble	Car	Exhibition	Avg.
Clean	99.42	99.33	99.46	99.48	99.42
20dB	98.31	98.52	98.51	96.98	98.08
15dB	96.59	95.68	98.15	95.09	96.38
10dB	91.13	87.27	94.27	87.94	90.15
5dB	76.39	66.48	80.14	66.95	72.49
0dB	48.88	38.81	48.14	37.64	43.37
-5dB	27.54	21.77	21.95	19.04	22.58
Avg.	76.89	72.55	77.23	71.87	74.64

표 2, 3, 4는 MFCC, CMSBS 특징벡터 추출 방법과 제안한 변형된 mel sub-band 스펙트럼 차감법을 이용하여 특징벡터를 추출하는 방법의 인식실험 결과를 나타낸다. CMSBS 방법에서 γ=0.08, α_i=1, β_i=0.1로 고정시키고 실험을 진행하였다.

표 2, 3, 4의 결과로 기존의 특징벡터 추출방법과 제안한 방법을 비교해 볼 때 깨끗한 음성일 경우는 비슷한 성능을 나타내고 있지만, 신호대잡음비가 작아질수록 제안한 방법이 우수한 성능을 나타내는 것을 확인할 수 있다.

5. 결론

본 논문에서는 잡음환경에서의 강인한 음성인식을 위한 특징벡터 추출을 목적으로, 기존의 CMSBS 특징벡터 추출 방법에서 사용하였던 mel sub-band 스펙트럼 차감법을 각 프레임의 자기상관계수의 첫 번째 극대값과 두 번째 극대값의 크기에 따라 결정되는 주기성(periodicity) 인자 P_n을 이용하여 β_i를 변형함으로써 음성 정보의 손실을 줄일 수 있는 방법을 제안하였다. 제안한 방법은 전체적인 인식률에 있어서 성능이 향상되며, 특히 신호대잡음비가 작아질수록 높은 성능 향상을 보였다.

참고문헌

[1] B. Nasersharif and A. Akbari,“SNR-dependent compression of enhanced Mel sub-band energies for compensation of noise effects on MFCC features,” Pattern Recognition Letters, Vol.28, No.11, 1320-1326, 2007.

[2] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. on Acoustic Speech Signal Processing, vol. ASSP-27, no. 2, pp. 113-120, Apr. 1980.

[3] Chen, J., Paliwal, K.K., Nakamura, S., 2001. Sub-band based additive noise removal for robust speech recognition. In: Proc. Eurospeech, pp. 571 - 574.

[4] ETSI standard document, “Speech processing, transmission and quality aspects(STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithm,” ETSI ES 201 108 v1.1.1