필터 뱅크 에너지 차감을 이용한 묵음 특징 정규화 방법의 성능 향상

정회원 신 광 호*, 최 숙 남*, 종신회원 정 현 열*

Performance Improvements for Silence Feature Normalization Method by Using Filter Bank Energy Subtraction

Guanghu Shen*, Sook-Nam Choi* *Regular Members*, Hyun-Yeol Chung* *Lifelong Member*

요 약

본 논문에서는 기존의 CLSFN (Cepstral distance and Log-energy based Silence Feature Normalization) 방법의 인식성능을 향상시키기 위하여, 필터 뱅크 서브 밴드 영역에서 잡음을 차감하는 방법과 CLSFN을 결합하는 방법, 즉 FSFN (Filter bank sub-band energy subtraction based CLSFN)을 제안하였다. 이 방법은 음성으로부터 특징 파라미터를 추출할 때 필터 뱅크 서브 밴드 영역에서 잡음을 제거하여 켑스트럼 특징을 향상시키고, 이에 대한 켑스트럼 거리를 이용하여 음성/묵음 분류의 정확도를 개선함으로써 기존 CLSFN 방법에 비해 향상된 인식성능을 얻을 수 있다. Aurora 2.0 DB를 이용한 실험결과, 제안하는 FSFN 방법은 CLSFN 방법에 비해 평균 단어 정확도 (word accuracy)가 약 2% 향상되었으며, CMVN (Cepstral Mean and Variance Normalization)과의 결합에서도 기존 모든 방법에 비해 가장 우수한 인식성능을 나타내어 제안 방법의 유효성을 확인할 수 있었다.

Key Words : speech recognition, feature enhancement, silence feature normalization, cepstral distance

ABSTRACT

In this paper we proposed FSFN (Filter bank sub-band energy subtraction based CLSFN) method to improve the recognition performance of the existing CLSFN (Cepstral distance and Log-energy based Silence Feature Normalization). The proposed FSFN reduces the energy of noise components in filter bank sub-band domain when extracting the features from speech data. This leads to extract the enhanced cepstral features and thus improves the accuracy of speech/silence classification using the enhanced cepstral features. Therefore, it can be expected to get improved performance comparing with the existing CLSFN. Experimental results conducted on Aurora 2.0 DB showed that our proposed FSFN method improves the averaged word accuracy of 2% comparing with the conventional CLSFN method, and FSFN combined with CMVN (Cepstral Mean and Variance Normalization) also showed the best recognition performance comparing with others.

Ⅰ. 서 론

음성인식은 인간과 기계간의 의사 전달을 위한 선도 기술 중의 하나라고 할 수 있다. 현재의 음성인식 기술을 여러 응용분야에 응용할 경우 잡음이 없는 환경에서는 매우 성공적으로 적용될 수 있다. 그러나 잡음이 존재하는 환경에서 보여주는 낮은 인식률 때문에 실제로 여러 분야에 적용되어 사용되고 있지 못하고 있는 실정이다. 따라서 많은 연구자들이 잡음환경하에서의 음성인식기의 성능을 높이는 데 꾸준한 노

* 영남대학교 정보통신공학과 (guanghosin@ynu.ac.kr, windy@ynu.ac.kr, hychung@ynu.ac.kr)
논문번호 : KICS2010-05-192, 접수일자 : 2010년 5월 1일, 최종논문접수일자: 2010년 6월 25일

력을 해오고 있다.

음성인식 시스템의 성능 저하는 주로 학습 환경과 인식 환경의 불일치 (mismatch)에서 초래된다. 이러한 불일치를 줄이기 위하여 다양한 접근 방법들이 제안되고 있는 데 크게 다음과 같이 세 가지 방식으로 나눌 수 있다. 첫 번째는 음성향상 (speech enhancement) 방식, 두 번째는 특징향상 (feature enhancement) 방식, 세 번째는 모델보상(model compensation) 방식이다[1].

음성의 특징 파라미터로는 주로 주파수 기반의 켑스트럼 특징(예: MFCCs: Mel-Frequency Cepstral Coefficients)과 에너지 기반의 로그 에너지 (log-energy) 특징으로 나눌 수 있다. 본 논문은 특징 향상에 관한 연구로써 로그 에너지 특징 정규화 방법에 대해 연구초점을 맞춘다.

기존의 로그 에너지 정규화 방법들을 살펴보면, ERN (log-Energy dynamic Range Normalization)[1], SFN-I (Silence Feature Normalization-I)[3], SFN-II (Silence Feature Normalization-II)[4] 등의 방법이 우수한 인식성능을 나타내고 있다. 하지만, 이 방법들은 로그 에너지 특징 정보만을 이용하여 음성/묵음 분류를 수행하므로, 높은 SNR에서는 우수한 성능을 보이고 있으나, 낮은 SNR에서는 로그 에너지의 분별력이 떨어져 성능이 현저히 저하되는 문제점이 있다.

이를 해결하기 위하여 저자들은 이전의 연구[5]에서 켑스트럼 거리와 로그 에너지를 결합하여 음성/묵음을 분류하는 방법을 제안하였다. 즉, 저자들이 제안한 CLSFN (Cepstral distance and Log-energy based Silence Feature Normalization) 방법은 잡음의 켑스트럼 특징 분포의 분산값이 음성보다 작다는 특성에 기반을 두고 있으며, 특히 이 특성은 SNR이 낮을수록 더욱 선명하게 나타난다. 따라서 켑스트럼 거리와 로그 에너지를 결합하여 음성/묵음의 분류에 사용할 경우, 특히 낮은 SNR에서 음성/묵음 분류의 정확도가 많이 개선되어 전반적으로 우수한 인식성능을 얻을 수 있었다.

그렇지만 CLSFN 방법이 낮은 SNR에서는 켑스트럼 거리를 이용하여 음성/묵음 분류의 정확도를 어느 정도 개선할 수는 있으나, 불안정적 잡음, 채널 왜곡 (예: Aurora 2.0 DB Set C 테스트 데이터) 등의 환경에서는 묵음(또는 잡음) 구간의 켑스트럼 특징 분포의 분산값이 증가하여 켑스트럼 거리의 분별력이 떨어지는 현상이 발생하여 성능향상에 한계가 있었다.

한편, J. Chen[6]과 D. Yu[7]에 의하면 필터 뱅크 서브 밴드 (filter bank sub-band) 영역에서 잡음의 에너지를 추정하여 제거할 경우 잡음에 더욱 강인한 음성 특징 파라미터를 추출할 수 있다. 즉, 음성으로부터 특징 파라미터를 추출할 때, 우선 필터 뱅크 서브 밴드 영역에서 잡음의 에너지를 추정하여 미리 제거한 후 이로부터 켑스트럼 특징을 추출한다. 따라서 추출된 켑스트럼 특징을 이용하여 켑스트럼 거리를 계산하고 이를 음성/묵음 분류에 사용할 경우, 음성/묵음 분류의 정확도를 향상시킬 수 있을 뿐만 아니라, 이 방법을 기존의 CLSFN 방법에 결합할 경우 더욱 우수한 인식성능을 기대할 수 있다.

본 논문에서는 CLSFN의 문제점을 해결하기 위해 필터 뱅크 서브 밴드 영역 에너지 차감 기반의 CLSFN 방법, 즉 FSFN (Filter bank sub-band energy subtraction based CLSFN) 방법을 제안하고 그에 대한 유효성을 검증하기로 한다.

본 논문의 구성은 다음과 같다. 2장에서 저자들이 이전 연구에서 제안했든 묵음 특징 정규화 방법 (CLSFN)에 대해서 소개하고, 3장에서는 본 논문에서 제안하는 필터 뱅크 서브 밴드 에너지 차감 기반의 CLSFN (FSFN) 방법에 대해 소개한다. 4장에서 인식 실험을 통해 제안 방법의 유효성을 확인한 후, 마지막으로 5장에서 결론을 맺는다.

Ⅱ. Cepstral Distance and Log-Energy Based Silence Feature Normalization

CLSFN[5] 방법은 켑스트럼 거리와 로그 에너지를 결합하여 음성/묵음을 분류 한 후, 묵음 특징만을 찾아서 매우 작은 값으로 정규화 하여, 음성인식의 학습 환경과 인식 환경의 불일치를 줄여주므로 잡음환경하의 음성인식 성능을 향상시키는 방법이다. 이하 이 방법의 중심이 되는 켑스트럼 거리 및 그의 기준값, 로그 에너지 및 그의 기준값을 구한 후 이를 결합하여 묵음 특징을 정규화하는 과정을 간략하게 소개한다.

• 켑스트럼 거리 및 그의 기준값

음성의 시작 묵음구간의 평균 켑스트럼 특징 벡터를 식 (1)을 이용하여 구한다. 여기서 N_F는 음성의 시작 묵음구간의 프레임 수를 나타내며, $c_i[n]$는 n번째 프레임의 i번째 차수의 켑스트럼 특징 벡터이다. ($N_F = 30$)

$$\bar{c}_i = \frac{1}{N_F}\sum_{n=1}^{N_F} c_i[n] \qquad (1)$$

n번째 프레임에 대한 음성의 시작 묵음구간의 평균 켑스트럼 벡터와의 유클리디언 (euclidean) 거리 $d[n]$는 식 (2)을 이용하여 구할 수 있다. 여기서 p는 켑스트럼 특징 벡터의 차수를 의미한다.

$$d[n] = \sum_{i=1}^{p} (c_i[n] - \overline{c_i})^2 \tag{2}$$

다음은 식 (3)과 같이 미디언 (median) 필터링을 하여 켑스트럼 거리를 스무딩 (smoothing) 한다. 여기서 필터 길이 (length)는 $2k+1$이며, k는 5로 설정한다.

$$\tilde{d}[n] = median(d[j] | j = n-k, \ldots, n+k) \tag{3}$$

음성/묵음의 분류를 위한 켑스트럼 거리에 대한 기준값(T_0)은 식 (4)을 이용하여 계산 할 수 있다.

$$T_0 = \frac{1}{N_F} \sum_{n=1}^{N_F} \tilde{d}[n] \quad (N_F = 30) \tag{4}$$

• 로그 에너지 및 그의 기준값

로그 에너지에 대한 기준값을 구하기 위하여 먼저 식 (5)를 이용하여 필터링을 한다. 여기서 $\log E[n]$은 n번째 프레임의 로그 에너지를, $\log \overline{E}[n]$은 필터링의 출력 값을 나타낸다.

$$\log \overline{E}[n] = \frac{1}{2}(\log E[n+1] - \log \overline{E}[n-1]) \tag{5}$$

따라서 로그 에너지에 대한 기준값 (T_1)은 식 (6)을 이용하여 계산할 수 있으며, 여기서 N은 음성 데이터의 프레임 수를 나타낸다.

$$T_1 = \frac{1}{N} \sum_{n=1}^{N} \log \overline{E}[n] \tag{6}$$

최종적으로 켑스트럼 거리와 로그 에너지를 결합하여 묵음의 로그 에너지 특징을 식 (7)과 같이 정규화 한다.

$$\begin{aligned}
&IF(\tilde{d}[t] > 1.2T_0) \ //1\text{차분류} \\
&\quad IF(\log \hat{E}[n] > T_1) \ //2\text{차분류} \\
&\quad\quad \log \hat{E}[n] = \log E[n] \\
&\quad ELSE \\
&\quad\quad IF(\tilde{d}[t] > 3T_0) \ //3\text{차분류} \\
&\quad\quad\quad \log \hat{E}[n] = \log E[n] \\
&\quad\quad ELSE \\
&\quad\quad\quad \log \hat{E}[n] = \log(\epsilon) + \delta \\
&ELSE \\
&\quad \log \hat{E}[n] = \log(\epsilon) + \delta
\end{aligned} \tag{7}$$

즉, 켑스트럼 거리를 기준값 ($1.2T_0$)과 비교하여, 작은 값을 갖는 구간은 묵음으로 분류하고, 큰 값을 갖는 구간은 음성으로 분류하는 1차 분류를 수행한다.

다음은 1차 분류의 결과에서 음성구간에 대하여, 로그 에너지를 이용하여 2차 분류를 수행하며, 그 중에서 로그 에너지 값이 기준값 (T_1) 보다 작아 묵음으로 분류된 묵음구간에 대하여 큰 켑스트럼 기준값 ($3T_0$)을 이용하여 3차 음성/묵음 분류를 수행한다. 즉, 2차 분류에서 묵음구간으로 오 분류한 경우를 보상하기 위한 작업이다. 여기서, ε는 상수 10^{-3}이고, δ는 평균값 0, 분산값 10^{-8}인 매우 작은 수를 의미한다.

Ⅲ. Filter Bank Sub-Band Energy Subtraction Based CLSFN

일반적으로 묵음(또는 잡음)의 켑스트럼 특징 분포의 분산값은 음성의 경우 보다 작은 값을 가진다. 기존의 CLSFN 방법은 이러한 특성에 근거하여, 켑스트럼 거리와 로그 에너지를 결합하여 음성/묵음의 분류에 사용하여 분류의 정확도를 향상시키는 방법이다.

특히, CLSFN 방법에서는 켑스트럼 거리를 이용하여 낮은 SNR에서 음성/묵음 분류의 정확도를 어느 정도 개선시킬 수 있었다. 그러나, 불안정한 잡음, 채널 왜곡 (예: Aurora 2.0 DB에서 Set C 테스트 데이터) 등의 환경에서는 묵음(또는 잡음)구간의 켑스트럼 특징 분포의 분산값이 크게 증가하므로 켑스트럼 거리의 분별력이 떨어지는 문제점이 발생하였다.

이를 해결하기 위해, 본 논문에서는 필터 뱅크 서브 밴드 에너지 차감 기반의 CLSFN 방법, 즉 FSFN (Filter bank sub-band energy subtraction based CLSFN) 방법을 제안하였다. 그림 1은 제안하는 FSFN 방법의 블록 다이어그램을 나타내고 있다. 즉, 먼저 음성신호로부터 주파수 변환 (FFT)을 수행한 후, 필터 뱅크 서브 밴드 영역에서 잡음 성분의 에너지를 추정하여 이를 원 신호로부터 차감한 후, 로그 (logarithm) 변환, DCT (Discrete Cosine Transform) 변환을 차례로 수행하여 향상된 켑스트럼 특징을 추출한다. 다음은 이 특징에 대한 켑스트럼 거리

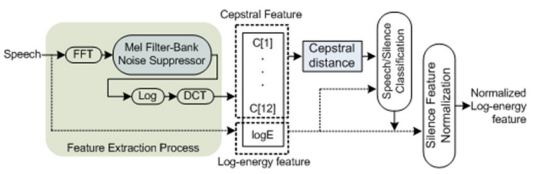

그림 1. FSFN 방법의 블록 다이어그램

(cepstral distance)를 음성/묵음 분류에 이용한다. 따라서 음성/묵음 분류의 정확도를 개선할 수 있을 뿐만 아니라, 묵음 특징 정규화의 인식성능도 함께 향상시킬 수 있다. 이하 제안한 방법에 대해 순차적으로 설명한다.

• Filter Bank Sub-Band Energy Subtraction[6,7]

본 논문에서는 필터 뱅크 서브 밴드 영역에서 잡음 성분의 에너지를 차감하기 위하여 식 (8)을 이용한다.

$$E_{\hat{S}}(i,t)=\begin{cases}E_Y(i,t)-\alpha^* E_{\hat{N}}(i,t) & \text{if } E_Y(i,t)>\frac{\alpha}{1-\beta}E_{\hat{N}}(i,t)\\ \beta^* E_Y(i,t) & otherwise\end{cases} \quad (8)$$

여기서 $E_Y(i,t)$, $E_N(i,t)$ 및 $E_{\hat{S}}(i,t)$는 각각 t번째 프레임, i번째 필터 뱅크 서브 밴드의 입력 에너지 값, 잡음 성분의 에너지 추정값 및 에너지 차감의 출력값을 나타낸다. α, β는 over-estimation factor, spectral flooring 파라미터를 의미한다. ($\alpha=3.0$, $\beta=0.1$)

• Sub-Band Minimum Power Estimation[8,9]

서브 밴드에서 잡음 성분의 에너지 값을 추정하는 과정은 다음과 같다.

먼저, 식 (9)를 이용하여 각 프레임에 대하여 각 서브 밴드의 출력 에너지 값을 1차 재귀 평균 스무딩(smoothing)을 한다. α_s는 스무딩 파라미터이며 0.9로 설정한다.

$$P(i,t)=\alpha_s P(i,t-1)+(1-\alpha_s)E_Y(i,t) \quad (9)$$

각 프레임의 각 서브 밴드에 대한 최소 에너지 값을 추정하기 위하여, 식 (10), (11)과 같이 먼저 변수 초기화를 수행한다. 현재의 검색창 (길이 $L=20frames$) 안에서 식 (12), (13)을 이용하여 현재 프레임의 각 서브 밴드의 최소 에너지 값 ($P_{\min}(i,t)$)과 최소 임시 에너지 값 ($P_{tmp}(i,t)$)을 구한다.

$$P_{\min}(i,0)=P(i,0) \quad (10)$$

$$P_{tmp}(i,0)=P(i,0) \quad (11)$$

$$P_{\min}(i,t)=\min\{P_{\min}(i,t-1),P(i,t)\} \quad (12)$$

$$P_{tmp}(i,t)=\min\{P_{tmp}(i,t-1),P(i,t)\} \quad (13)$$

현재 검색창의 모든 프레임을 다 읽은 후, 다음 검색창의 데이터를 읽기 위하여, 식 (14), (15)를 이용하여 다시 변수 초기화를 수행한다. 즉, 현재 프레임의 입력 에너지 값을 현재 프레임의 최소 임시 에너지 값 ($P_{tmp}(i,t)$)으로 설정해준다.

$$P_{\min}(i,t)=\min\{P_{tmp}(i,t-1),P(i,t)\} \quad (14)$$

$$P_{tmp}(i,t)=P(i,t) \quad (15)$$

위 과정을 반복하면 각 프레임의 각 서브 밴드의 최소 에너지 값을 추정할 수 있으며, 그 다음은 현재 프레임의 각 서브 밴드에 음성의 유무를 판별하기 위하여 식 (16)을 계산한다.

$$\vartheta=\frac{P(i,t)}{P_{\min}(i,t)} \quad (16)$$

마지막으로 식 (16)의 결과를 바탕으로 식 (17)을 이용하여 현재 프레임에 대한 잡음 성분의 에너지 값을 추정한다. 즉, ϑ값이 10보다 크면, 현재 서브 밴드에 음성이 포함되어 있음을 의미하므로 전 프레임에서 동일 서브 밴드에서 추정된 잡음 에너지 값을 그대로 사용하는 반면, 작으면 잡음 성분만 존재하는 것으로 판별하여 전 프레임의 동일 서브 밴드의 잡음 에너지 추정 값과 현재 프레임의 동일 서브 밴드의 입력 에너지 값과 결합하여 잡음 에너지 값을 다시 추정한다.

$$E_{\hat{N}}(i,t)=\begin{cases}E_{\hat{N}}(i,t-1) & \text{if } \vartheta>10\\ \alpha_s E_{\hat{N}}(i,t-1)+(1-\alpha_s)E_Y(i,t) & otherwise\end{cases} \quad (17)$$

위 과정을 통하여 잡음 성분을 음성신호로부터 제거하므로 개선된 켑스트럼 특징 파라미터를 추출할 수 있게 된다. 따라서 다음 단계에서는 개선된 켑스트럼 특징을 이용하여 2.1절에서 설명한 CLSFN 방법의 실행 절차를 동일하게 적용 하면 된다.

Ⅳ. 실험 및 결과

4.1 실험 환경

인식실험 및 평가를 위하여 Aurora 2.0 DB[10]를 사용하였다. Aurora 2.0 DB에는 2가지의 훈련환경이 있는데, 8440개의 clean 발성으로 구성된 clean-condition과 동일한 발성을 20개의 잡음환경에 나누어 각 422개의 발성으로 구성된 multi-condition이 있다. 잡음환경은 4종류의 잡음 (subway, babble, car, exhi-

bition)과 각각의 5종류의 잡음 레벨 (clean, 20dB, 15dB, 10dB, 5dB)로 구성되어 있다. 테스트 데이터는 3가지의 subset로 구성되어 있으며, 훈련에서 이용한 4종류 잡음을 포함한 Set A와 훈련에서 이용되지 않은 새로운 4종류의 잡음 (restaurant, street, airport, station)을 포함한 Set B, 그리고 Set A와 Set B에 나타난 2종류의 잡음 (subway, street)에 훈련환경과 다른 채널특성을 포함한 Set C의 총 10종류 잡음으로 -5dB에서 clean 까지 7가지의 잡음 레벨로 구성되어 있다.

기본 인식기는 Aurora2-HTK를 사용하였다. 단어 모델은 one, two, three, four, five, six, seven, eight, nine, zero, oh의 11개로 정의되었고, 각 단어 모델은 3 혼합수 (mixture), 16 상태 (state)를 갖는 CHMM (Continuous Hidden Markov Models)으로 구성되었다. 인식 시스템에는 11개의 단어 모델 외에 2개의 묵음 모델 (silence model)이 포함되어 있는데, 각각 3 상태와 1 상태의 CHMM으로 구성되었다. 특징 파라미터는 12차 MFCCs와 1차 로그 에너지, 그리고 각각의 delta 및 delta-delta 계수를 포함한 총 39차로 구성하였다. 그리고 분석 프레임의 크기는 25ms이며, 10ms씩 이동하면서 특징 파라미터를 추출하였다.

본 논문에서는 clean-condition에서 실험을 수행하였으며, 성능 평가에서는 20dB에서 -5dB까지의 평균 단어 정확도 (word accuracy)를 비교하였다.

4.2 실험 결과

표 1은 Baseline, 기존의 로그 에너지 정규화 방법 (ERN, SFN-I/II, CLSFN) 및 제안 방법에 대한 평균 단어 정확도를 나타낸 것 이다. 기존의 로그 에너지 정규화 방법은 Baseline에 비해 Set A, B에서 인식 성능이 현저히 향상된 것을 확인 할 수 있다. 하지만 Set C와 같은 채널 왜곡이 존재하는 환경에서는 기존 방법의 효과가 거의 없는 것으로 나타났다. 이는 기존 방법이 로그 에너지 특징만을 사용하여 음성/묵음을 분류하였으므로 잡음의 주파수 특성을 고려해주지 못하기 때문으로 분석된다.

저자들의 이전 연구에서 제안한 CLSFN 방법은 Set A, B에서 우수한 인식성능을 얻을 수 있었을 뿐만 아니라, Set C에서도 기존 방법에 비해 어느 정도 향상된 인식성능을 얻을 수 있어 기존의 로그 에너지 정규화 방법의 고질적인 문제를 해결하는데 켑스트럼 거리 측도를 도입하는 것이 많은 도움이 되었음을 의미한다.

본 논문에서 제안한 FSFN 방법은 사전에 기대했던 것과 같이 CLSFN 방법에 비해 Set A, B에서 안정적인 성능향상을 나타냈을 뿐만 아니라, Set C에서 약 5%의 성능향상, 전체 평균 단어 정확도에서 약 2%의 성능향상을 얻어, 전반적으로 가장 우수함을 확인 할 수 있었다. 그 이유는 필터 뱅크 서브 밴드에서 잡음 성분의 에너지를 차감하므로 개선된 켑스트럼 특징을 얻을 수 있었으며, 따라서 음성/묵음 분류에 대한 켑스트럼 거리의 분별력도 함께 개선되어, 최종 인식성능 향상에 기여한 것으로 판단된다.

표 2에서는 켑스트럼 정규화 방법 (CMVN: Cepstral Mean and Variance Normalization)[11] 및 그와 로그 에너지 정규화 방법을 결합할 경우에 대한 인식결과를 나타내었다. 먼저, CMVN 방법과 표 1의 로그 에너지 정규화 방법들과 비교해보면, 로그 에너지 정규화 방법들은 주로 Set A, B에서 우수한 성능을, CMVN은 Set C에서 비교적 우수한 성능을 나타내고 있었다. 이것은 Set C의 채널왜곡 오염을 완화하는데 있어서 기존의 로그 에너지 정규화 방법이 취약했다는 것을 의미한다.

로그 에너지 특징과 켑스트럼 특징은 서로 독립적인 관계를 가지고 있으며, 일반적으로 이 두 특징에 대한 정규화 방법을 동시에 적용할 경우 향상된 인식성능을 얻을 수 있다. 따라서 두 가지 방식을 결합하였을 경우의 인식실험 결과는 결합하기 전의 CMVN과 로그 에너지 정규화 방법들과 각각 비교하였을 경우, 전자에 비해 약 11~15%, 후자에 비해 5~7%의 성능이 향상되었음을 확인 할 수 있었다.

특히, 본 논문에서 제안한 FSFN 방법과 CMVN을 결합 할 경우, 즉 FSFN-CMVN은 기타 방법에 비해

표 1. 로그 에너지 정규화 방법에 대한 인식 성능 비교

Method	Set A	Set B	Set C	Avg.
Baseline	52.44	47.73	57.03	52.40
ERN	62.68	59.09	57.58	59.78
SFN-I	62.25	63.95	53.10	59.76
SFN-II	63.09	65.27	53.07	60.48
CLSFN	65.85	65.66	58.46	63.33
FSFN	**66.53**	**66.49**	**63.04**	**65.35**

표 2. 결합 방법에 대한 인식 성능 비교

Method	Set A	Set B	Set C	Avg.
CMVN	53.74	53.61	58.59	55.31
ERN-CMVN	66.79	66.76	67.52	67.02
SFN-I-CMVN	67.45	68.27	63.54	66.42
SFN-II-CMVN	67.78	68.19	62.84	66.27
CLSFN-CMVN	69.95	69.67	69.54	69.72
FSFN-CMVN	**70.60**	**71.11**	**68.03**	**69.92**

가장 우수한 인식성능을 나타냈다. 하지만 CLSFN-CMVN과 비교했을 경우, Set A와 B에서는 성능이 향상되었으나, Set C에서는 성능향상이 약간 떨어지는 현상도 발생하였다. 그 이유는 FSFN에서의 필터 뱅크 서브 밴드 영역의 잡음 에너지를 차감과 CMVN에서의 켑스트럼의 평균값을 차감하므로, 비록 실행하는 영역은 다르지만 음성의 최종 켑스트럼 특징 파라미터에서 왜곡이 누적되기 때문으로 분석된다. 따라서 필터 뱅크 서브 밴드 영역에서의 잡음 성분을 좀 더 정확하게 추정 및 차감을 실행할 필요가 있을 것으로 판단되며, 이에 대한 연구는 향후 연구에서 계속 진행할 계획이다.

Ⅴ. 결 론

CLSFN 방법은 켑스트럼 유클리디언 거리와 로그 에너지를 결합하여 음성/묵음을 분류하여 묵음의 로그 에너지 특징을 매우 작은 값으로 정규화하는 방법이다. 본 논문에서는 기존의 CLSFN 방법의 인식성능을 향상시키기 위하여, 필터 뱅크 서브 밴드 영역에서 잡음을 차감하는 방법과 CLSFN 결합하는 방법, 즉 FSFN (Filter bank sub-band energy subtraction based CLSFN)을 제안하였다. 이 방법은 켑스트럼 특징을 향상시킬 수 있을 뿐만 아니라, 이 특징에 대한 켑스트럼 거리를 이용한 음성/묵음(또는 잡음) 분류의 분별력을 개선할 수 있어 기존의 묵음 특징 정규화 방법의 인식성능을 향상시킬 수 있다. Aurora 2.0 DB를 이용한 실험결과, 제안된 FSFN 방법은 기존의 CLSFN 방법에 비해 평균 단어 정확도가 약 2% 향상되었으며, CMVN과의 결합에서도 기존 모든 방법에 비해 가장 우수한 인식성능을 나타내었다.

참 고 문 헌

〔1〕 K.S. Yao, E. Visser, O.W. Kwon and T.W. Lee, “A Speech Processing Front-End with Eigenspace Normalization for Robust Speech Recognition in Noisy Automobile Environments,” *Proc. Eurospeech*, pp.9-12, Sep. 2003.

〔2〕 W.Z. Zhu and D.O. Shaughnessy, “Log Energy Dynamic Range Normalization for Robust for Robust Speech Recognition,” *Proc. ICASSP*, Vol.1, pp.245-248, 2005.

〔3〕 C.-F. Tai and J.-W. Hung, “Silence Energy Normalization for Robust Speech Recognition in Additive Noise Environments,” *Proc. ICSLP*, pp.2558-2561, Sep. 2006.

〔4〕 C.-C. Wang, C.-A. Pan and J.-W. Hung, “Silence Feature Normalization for Robust Speech Recognition in Additive Noise Environments,” *Proc. ICSLP*, pp.1028-1031, Sep. 2008.

〔5〕 신광호, 정현열, “강인한 음성인식을 위한 켑스트럼 거리와 로그 에너지 기반 묵음 특징 정규화,” *한국음향학회지*, Vol.29, No.4, pp. 278-285, 2010.

〔6〕 J. Chen, K.K. Paliwal and S. Nakamura, “Sub-Band Based Additive Noise Removal for Robust Speech Recognition,” *Proc. Eurospeech*, pp. 571-574, 2001.

〔7〕 D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong and A. Acero, “A Minimum Mean Square Error Noise Reduction Algorithm on Mel Frequency Cepstra for Robust Speech Recognition,” *Proc. ICASSP*, Las Vegas, USA, 2008.

〔8〕 G.I. Cohen and B. Berdugo, “Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement,” *IEEE Signal Process. Lett.*, Vol.9, No.1, pp. 12-15, Jan. 2002.

〔9〕 R. Martin, “Spectral Subtraction Based on Minimum Statistics,” *Proc. 7th EUSIPCO94*, pp.1182-1185, 1994.

〔10〕 H.-G Hirsch and D. Pearce, “The Aurora Experimental Framework for The Performance Evaluation of Speech Recognition Systems Under Noisy Conditions,” *ISCA ITRW ASR*, France, Sep. 2000.

〔11〕 O. Viikki and K. Laurila, “Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition,” *Speech Communication*, Vol.25, pp.133-147, 1998.

신 광 호 (Guanghu Shen) 정회원

2002년 8월 (中) 연변대학교 사범대학 수학교육학과 이학사
2005년 8월 영남대학교 정보통신공학과 공학석사
2005년 9월~현재 영남대학교 정보통신공학과 박사과정
<관심분야> 잡음처리, 음성인식, 디지털 신호처리

최 숙 남 (Sook-Nam Choi) 정회원

1995년 8월 영남대학교 전자공학과 공학사
2007년 8월 영남대학교 전기전자통신교육 교육학석사
2008년 9월~현재 영남대학교 정보통신공학과 박사과정
<관심분야> 잡음제거, 음성인식

정 현 열 (Hyun-Yeol Chung) 종신회원

1989년 (日) 동북대학교 정보공학과 공학박사
1989년~현재 영남대학교 공과대학 정보통신공학과 교수
2006년1월~2006년12월 한국음향학회 회장
<관심분야> 음성인식, 화자인식, 음성합성 및 DSP 응용분야