한국음향학회지 제29권 제4호 pp. 278~285 (2010)

강인한 음성인식을 위한 켑스트럼 거리와 로그 에너지 기반 묵음 특징 정규화

Cepstral Distance and Log-Energy Based Silence Feature Normalization for Robust Speech Recognition

신 광 호*, 정 현 열*
(Guanghu Shen*, Hyun-Yeol Chung*)

*영남대학교 정보통신공학과
(접수일자: 2010년 3월 22일; 채택일자: 2010년 4월 27일)

훈련 환경과 인식 환경의 차이가 음성인식 성능저하의 주요요인이다. 이러한 환경의 불일치를 줄이기 위한 방법으로 다양한 묵음특징 정규화 방법이 제안되고 있다. 기존의 묵음특징 정규화 방법은 낮은 SNR (Signal-to-Noise Ratio)에서 묵음구간의 에너지 레벨이 증가하여 음성/묵음 분류의 정확도가 떨어짐으로 인해 인식성능이 저하되는 문제점이 있었다. 본 논문에서는 로그 에너지와 음성/묵음 (또는 잡음)의 켑스트럼 특징의 분포 특성의 차이를 나타내는 켑스트럼 유클리디언 (Euclidean) 거리를 결합하여 음성/묵음을 분류하는 묵음특징 정규화 방법 (Cepstral distance and Log-energy based Silence Feature Normalization)을 제안하였다. 제안한 방법은 높은 SNR에서는 로그 에너지 특징이 잡음의 영향을 적게 받는 특성을 반영하여 기존의 묵음 특징 정규화 (Silence Feature Normalization)방법의 우수성을 그대로 유지하는 반면, 낮은 SNR에서는 로그 에너지 대신 음성/묵음 분류의 분별력이 우수한 켑스트럼 거리 정보를 이용함으로써 인식성능을 향상시킬 수 있다. 인식실험결과 기존의 SFN-I/II, CSFN 방법에 비해 전반적으로 향상된 인식성능을 얻을 수 있어 그 유효성을 확인할 수 있었다.
핵심용어: 음성 인식, 특징 향상, 묵음 특징 정규화, 켑스트럼 거리
투고분야: 음성처리 분야 (2.5)

The difference between training and test environments is one of the major performance degradation factors in noisy speech recognition and many silence feature normalization methods were proposed to solve this inconsistency. Conventional silence feature normalization method represents higher classification performance in higher SNR, but it has a problem of performance degradation in low SNR due to the low accuracy of speech/silence classification. On the other hand, cepstral distance represents well the characteristic distribution of speech/silence (or noise) in low SNR. In this paper, we propose a Cepstral distance and Log-energy based Silence Feature Normalization (CLSFN) method which uses both log-energy and cepstral euclidean distance to classify speech/silence for better performance. Because the proposed method reflects both the merit of log energy being less affected with noise in high SNR and the merit of cepstral distance having high discrimination accuracy for speech/silence classification in low SNR, the classification accuracy will be considered to be improved. The experimental results showed that our proposed CLSFN presented the improved recognition performances comparing with the conventional SFN-I/II and CSFN methods in all kinds of noisy environments.
Keywords: Speech recognition, Feature enhancement, Silence feature normalization, Cepstral distance
ASK subject classification: Speech Signal Processing (2.5)

책임저자: 정 현 열 (hychung@ynu.ac.kr)
712-749 경북 경산시 대동 214-1 영남대학교 전자정보공학부
(전화: 053-810-2496; 팩스: 053-810-4742)

I. 서 론

음성인식 기술은 컴퓨터가 인간의 음성을 식별하는데 필요한 여러 가지 특징 정보를 분석, 추출한 다음 이를 인식을 위한 특징 파라미터의 열로서 표현한 후 패턴 인식 방법을 이용하여 언어 정보를 추출하는 기술이다. 현재의 음성인식 기술은 잡음이 없는 환경에서는 좋은 인식 성능을 보여주고 있으나, 잡음이 존재하는 환경에서 보여주는 낮은 인식성능 때문에 실제 환경하의 응용에는 어려운 점이 많다. 이로 인하여 많은 연구자들이 잡음 환경 에서 음성인식 시스템의 성능을 높이는 데 꾸준한 노력을 기울이고 있다.

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

본 논문은 특징향상 방식에 관한 연구로서, 음성인식에서 주로 사용되고 있는 특징 파라미터로는 MFCC (Mel-Frequency Cepstral Coefficients), PLP (Perceptual Linear Predictive) 등의 주파수 기반 특징과 에너지 기반 특징이 있으며, 이 가운데 에너지 특징 향상에 대한 연구에 초점을 맞추고 있다.

일반적으로 에너지 특징은 유성음과 무성음을 구분시켜주는 등 음소간의 변별력을 향상시키는 데 중요한 특징으로 이용되고 있으나, 잡음환경 하에서 이 특징을 그대로 사용할 경우 사람의 목소리 크기나 환경 변화에 따라 에너지 분포가 크게 달라진다. 이를 완화하기 위해 에너지 특징을 로그 에너지 또는 delta 로그 에너지로 변환해서 사용하는 경우가 많다. 하지만 로그 에너지 특징도 잡음이 많이 존재하는 환경에서는 여전히 잡음의 오염을 효과적으로 억제하지 못하는 문제점을 가지고 있다. 이와 같은 이유로 로그 에너지 특징을 정규화 하거나 보상해주는 다음과 같은 방법들이 많이 연구되고 있다.

Tai et al., [2]은 로그 에너지 특징을 IIR HPF (Infinite Impulse Response High-Pass Filter)를 통과시켜 음성/묵음을 분류하여 묵음구간만 작은 값으로 낮춰주는 묵음 특징 정규화 (SFN-I: Silence Feature Normalization I) 방법을 제안하였다. 이 방법은 높은 SNR (Signal-to-Noise Ratio)에서는 상당히 우수한 성능을 나타내는 반면, 낮은 SNR에서 성능향상이 미미하다는 문제점이 있다. 그 이유는 로그 에너지 특징을 이용한 음성과 묵음(또는 잡음)간의 분별력이 잡음이 많은 환경에서는 현저하게 저하되기 때문이다. Wang et al., [3]은 SFN-I 방법을 개선하기 위하여 SFN-I 방법을 이용하여 처리한 출력에 로그 에너지 값의 크기에 비례하는 가중치 (weighting)를 곱해주는 방법 SFN-II를 제안하였다. SFN-II 방법의 인식성능은 SFN-I 방법에 비해 많이 향상되었으나, 낮은 SNR에서의 성능향상이 여전히 부족함을 나타내었다.

저자들은 선행연구 [4]에서 켑스트럼의 유클리디언 거리 [5]를 이용하여 음성/묵음을 분류한 후 묵음특징을 정규화하는 CSFN (Cepstral distance based Silence Feature Normalization) 방법을 제안하였다. 이 방법은 대부분의 잡음환경에서, 특히 낮은 SNR환경에서 우수한 인식성능을 나타내고 있다. 그러나 이 방법의 문제점은 불안정적인 잡음환경 (예, Aurora 2.0의 테스트 데이터 Set B)에서는 인식성능이 저하되는 문제점이 있었다. 이것은 불안정한 배경잡음 (때로는 사람의 말소리 포함)으로 인하여 음성과 묵음의 켑스트럼 거리의 차이가 줄어들어 음성/묵음의 분별력이 저하되기 때문으로 분석되었다.

위의 분석결과로부터 높은 SNR에서 우수한 성능을 나타내는 SFN-I 방법과 낮은 SNR에서 우수한 음성/묵음의 분별력을 나타내는 켑스트럼 거리를 결합하면 잡음환경 하에서 더 향상된 인식성능을 나타낼 것으로 생각된다. 즉, 잡음의 켑스트럼 특징은 음성에 비해 상대적으로 작은 분산 값을 가지므로 켑스트럼 유클리디언 거리로 음성/묵음을 분류할 경우, 특히 낮은 SNR에서의 정확도를 많이 향상시킬 수 있기 때문이다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 묵음 특징 정규화방법에 대해서 설명하고 3장에서는 기존의 묵음특징 정규화 방법에 대해 분석한 후, 4장에서 제안하는 켑스트럼 거리와 로그 에너지 기반의 묵음특징 정규화 방법에 대해서 기술한다. 5장에서는 제안한 방법에 대해 실험을 통해 그 유효성을 확인한 후 마지막으로 6장에서 결론을 맺도록 한다.

II. 에너지 특징 향상

2.1. SFN-I (Silence Feature Normalization I)

SFN-I 방법은 작은 로그 에너지 값을 갖는 묵음 구간을 찾아서 일정 값 이하로 줄여주는 방법이다. 수행절차는 다음과 같다.

먼저, 로그 에너지 값을 식 (1)과 같이 IIR HPF를 통과시킨다. 여기서, $\log E[n]$은 n번째 프레임의 로그 에너지를, $\log\bar{E}[n]$은 필터링의 출력 값을 나타낸다.

$$\log\bar{E}[n] = \frac{1}{2}(\log E[n+1] - \log\bar{E}[n-1]) \quad (1)$$

다음은 식 (2)를 이용하여 음성/묵음 분류의 기준값 (T_0)를 계산한다. 여기서, N은 각 음성 데이터의 프레임 수를 의미한다.

$$T_0 = \frac{1}{N}\sum_{n=1}^{N}\log\bar{E}[n] \quad (2)$$

마지막으로 정규화된 로그 에너지 $\log\hat{E}[n]$는 식 (3)을 이용하여 계산한다.

$$\log\hat{E}[n] = \begin{cases} \log E[n] & \text{if } \log\bar{E}[n] > T_0 \\ \log(\varepsilon) + \delta & \text{if } \log\bar{E}[n] \le T_0 \end{cases} \quad (3)$$

즉, 식 (1)에서 계산된 $\log\bar{E}[n]$값이 기준값 (T_0)보다 크면 음성으로 분류하여 로그 에너지 값을 원래의 값으로 유지하고, 반면에 작으면 묵음으로 분류하여 매우 작은 값ε로 정규화한다. 여기서, ε는 상수 10^{-3}이고, δ는 평균값 0, 분산값 10^{-8}인 매우 작은 수를 의미한다.

2.2. SFN-II (Silence Feature Normalization II)

SFN-II 방법은 SFN-I 을 개선한 방법이다. 이는 잡음에 오염된 음성구간은 잡음만 존재하는 구간에 비해 더 넓은 주파수 대역폭 (bandwidth)을 가지고 있으며 [3], 큰 로그 에너지 값을 갖는 구간은 작은 로그 에너지 값을 갖는 구간에 비해 잡음의 영향을 상대적으로 적게 받는다는 관측 [2, 3]에 근거하여, 기존의 SFN-I 방법의 출력값에 식 (4)와 같이 가중 함수 ($w[n]$)를 곱해주는 방법이다. 여기서, 가중 함수는 식 (5)와 같이 정의하며, 이는 그림 1과 같이 수정된 시그모이드 곡선 (sigmoid curve)의 형태를 취하고 있다. 즉, 음성구간에는 큰 값을 갖는 곡선 (그림에서 우측 부분)으로, 묵음구간에는 작은 값을 갖는 곡선 (그림에서 좌측 부분)으로 가중치를 곱해주므로, 기존의 SFN-I 방법에 비해 잡음에 더욱 강인한 로그 에너지 특징을 얻을 수 있게 된다.

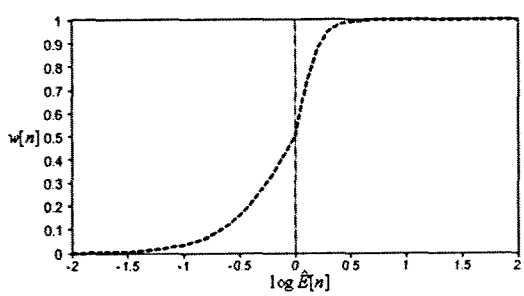

그림 1. 가중함수$w[n]$($T_0=0, \sigma_2=1, \sigma_1=3, \beta=0.1$으로 가정)
Fig. 1. The weighting function $w[n]$ (assuming $T_0=0$, $\sigma_2=1, \sigma_1=3, \beta=0.1$).

$$\log\hat{E}[n] = w[n] \cdot \log\hat{E}[n] \quad (4)$$

$$w[n] = \begin{cases} 1/(1+\exp(-(\log\bar{E}[n] - T_0)/\beta\sigma_1)) & \text{if } \log\bar{E}[n] > T_0 \\ 1/(1+\exp(-(\log\bar{E}[n] - T_0)/\beta\sigma_2)) & \text{if } \log\bar{E}[n] < T_0 \end{cases} \quad (5)$$

여기서 T_0는 SFN-I 방법에서와 동일한 기준값이고, σ_1, σ_2는 각 $\{\log\bar{E}[n] | \log\bar{E}[n] > T_0\}$, $\{\log\bar{E}[n] | \log\bar{E}[n] \le T_0\}$ 구간에 존재하는 $\log\bar{E}[n]$값에 대한 표준편차를 의미하며, β는 0.1로 설정한다.

2.3. Cepstral distance based Silence Feature Normalization (CSFN)

저자들에 의한 연구 [4]에서는 켑스트럼 특징의 유클리디언 거리 [5]를 이용하여 음성/묵음을 분류한 후 로그 에너지를 정규화하는 CSFN (Cepstral distance based Silence Feature Normalization) 방법을 제안하였다. 이 방법은 잡음의 켑스트럼 특징은 음성의 켑스트럼 특징에 비해 작은 분산 값을 갖는 분포 특성을 지니고 있다는 데에 근거를 두고 있다. 특히, SNR이 낮을수록 이 특성은 더욱 선명하게 나타나고 있다. 따라서 켑스트럼 특징의 거리를 이용하여 음성/묵음을 분류할 경우, 낮은 SNR에서의 분류의 정확도를 향상시킬 수 있어 전체적으로 묵음 특징 정규화 방법의 성능을 향상시킬 수 있다.

먼저, 식 (6)을 이용하여 음성의 시작 묵음구간의 평균 켑스트럼 특징 벡터를 구한다.

$$\overline{C_i} = \frac{1}{N_F}\sum_{n=1}^{N_F} c_i[n] \quad (6)$$

여기서 N_F는 음성의 시작구간의 묵음 프레임 수를 나타내며, $c_i[n]$는 n번째 프레임의 i번째 차수의 켑스트럼 특징 벡터이다. ($N_F=30$)

따라서 n번째 프레임에 대한 음성 시작구간의 평균 켑스트럼 벡터와의 유클리디언 거리 $d[n]$는 식 (7)과 같이 구한다. 여기서 p는 켑스트럼 특징 벡터의 차수를 의미한다.

$$d[n] = \sum_{i=1}^{p}(c_i[n] - \overline{C_i})^2 \tag{7}$$

보다 안정적인 켑스트럼 거리의 변화를 구하기 위하여 식 (8)과 같이 미디언 (median) 필터링을 수행한다. 여기서 필터의 길이 (length)는 $2k+1$이며, k는 5로 설정한다.

$$\bar{d}[n] = median(d[j] | j = n-k, \ldots, n+k) \tag{8}$$

SFN-I 방법과 유사하게, 식 (8)에서 계산된 켑스트럼 거리를 식 (9)의 IIR HPF를 통과시킨 후, 그에 대한 평균값 (식 (10))을 계산하여 기준값으로 설정하고, 식 (11)을 이용하여 각 프레임에 대하여 음성/묵음 분류 및 정규화를 수행한다.

$$\bar{d}[n] = \frac{1}{2}(\bar{d}[n+1] - \bar{d}[n-1]) \tag{9}$$

$$T_1 = \frac{1}{N}\sum_{n=1}^{N}\bar{d}[n] \tag{10}$$

$$\log\hat{E}[n] = \begin{cases} \log E[n] & \text{if } \bar{d}[n] > T_1 \\ \log(\varepsilon) + \delta & \text{if } \bar{d}[n] \le T_1 \end{cases} \tag{11}$$

III. 기존 특징 향상 방법에 대한 분석

일반적으로 실제 잡음환경에서 잡음의 오염은 대부분 일정하지가 않다. 특히 불안정 (non-stationary)적인 잡음인 경우, 어떤 구간에는 잡음이 많이 나타나는 반면, 어떤 구간에는 잡음이 상대적으로 적게 나타나는 경우가 있다. 예를 들어 Aurora 2.0 database의 공항 (airport) 잡음 5 dB 환경하의 "3615" 발성 데이터와 이 발성의 clean 환경에서의 에너지 곡선을 그림 2에 나타냈다. 잡음이 포함된 경우 그림에서 보는 바와 같이 41~51번째 음성구간은 75~85번째 묵음구간에 비해 에너지 레벨이 작은 것을 확인할 수 있다. 따라서 이러한 구간에는 각각 음성을 묵음으로, 묵음을 음성 등으로 오분류 할 가능성이 높을 것으로 예상된다.

그림 3은 기존의 SFN-I/II 방법을 이용한 묵음특징 정규화 결과의 일례를 나타냈다. 전반적으로 로그 에너지 기반의 SFN-I 방법은 안정적으로 음성/묵음을 분류 및 묵음특징을 정규화한 것을 확인할 수 있다. 그리고 SFN-II 방법은 SFN-I 방법의 결과에 에너지 크기에 비례하는 가중치를 곱하므로, 묵음구간의 로그 에너지는 더욱 작아지고, 음성구간의 시작과 끝나는 부분은 중간 부분에 비해 로그 에너지 값이 원래 값보다 약간 작아지는 것을 확인 할 수 있다. 따라서 이렇게 변환된 로그 에너지 특징은 잡음에 더욱 강인해지므로 인식성능 또한 향상되는 것을 실험으로 확인 할 수 있다.

그림 3으로부터 SFN-I 방법은 높은 SNR에서 우수한 인식성능을 나타내는 것을 확인 할 수 있으나, 낮은 SNR 환경에서는 잡음의 에너지 레벨이 증가하므로 음성/묵음의 분별력이 떨어지게 되고, 이로 인해 음성/묵음의 오분류가 증가하여 인식성능의 저하가 나타난다. 즉, 인식성능의 저하는 시작 묵음구간 (0~20번째 프레임)과 80번째 프레임 부분의 묵음 구간을 음성으로 오분류 하거나, 또는 50 번째 프레임 부분에서 약간의 음성구간이 묵음으로 오분류됨으로 나타난다.

CSFN 방법은 기존의 SFN-I/II 방법들에 비해 전반적으로 향상된 인식결과를 나타냈으나, 특정 환경 (Set B)

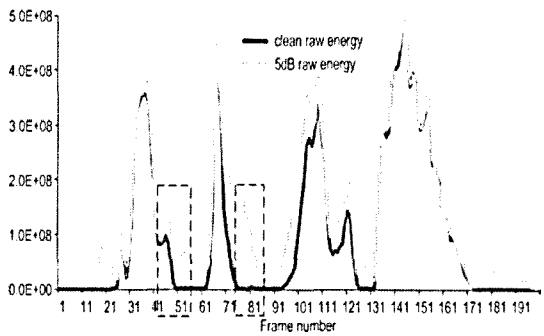

그림 2. 에너지 곡선 비교
Fig. 2. Comparison of energy curves.

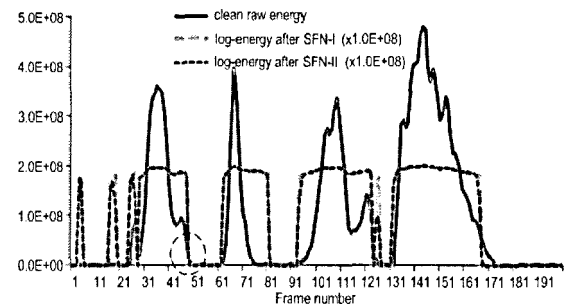

그림 3. SFN-I/II 방법을 이용한 묵음특징 정규화 비교
Fig. 3. Comparison of silence feature normalization using SFN-I/II methods.

에서는 인식성능이 오히려 저하되는 문제점을 실험을 통해 확인 할 수 있다. 이는 그림 4를 통해서 설명할 수 있다. 즉, 공항 잡음은 공항내부에서 발생하는 다양한 잡음으로써 사람의 말소리 등의 다양한 소리들이 포함되어 있어, 음성 데이터의 묵음 (또는 잡음) 구간과 음성 구간의 켑스트럼 거리의 차이가 일반 잡음환경에 비해 많이 줄어든다. 따라서, 잡음구간을 음성구간으로 (51번째 프레임 부분), 음성구간을 잡음구간으로 (101번째 프레임 부분) 오 분류하는 경우가 많이 발생하게 된다.

그리고 CSFN 방법에 대한 인식실험결과 (표 1)를 살펴보면, 인식 성능향상이 주로 낮은 SNR에서 이루어지는 반면, 높은 SNR에서는 SFN-I/II 방법에 비해 오히려 저하되는 것을 확인할 수 있다. 이는 음성과 묵음의 켑스트럼 거리 값의 차이가 로그 에너지 값의 차이보다 크므로, 식 (10)과 같이 기준값을 계산하여 음성/묵음을 분류할 경우, 음성의 시작과 끝 부분의 음소 (예: 무성음)들이 큰 기준값 때문에 묵음으로 분류되어 버리는 현상이 많이 발생하기 때문이다. 따라서 켑스트럼 거리를 SFN-I/II 방법과 유사하게 식 (9)를 이용하여 필터링한 후의 평균값을 음성/묵음 분류의 기준값으로 사용하는 것은 다양한 잡음환경하에서의 강건한 음성인식을 위해서는 부족한 것으로 판단된다.

IV. 제안하는 방법

4.1. Cepstral distance and Log-energy based Silence Feature Normalization (CLSFN)

앞에서 논의한 바와 같이 기존의 SFN-I 방법은 높은 SNR에서 우수한 성능을 나타내고 있고, 켑스트럼 거리를 이용한 음성/묵음 분류 및 묵음특징 정규화 방법 CSFN은 낮은 SNR에서 우수한 성능을 나타냄을 알 수 있다. 따라서 본 논문에서는 켑스트럼 거리와 로그 에너지를 결합하여 음성/묵음 분류를 위해 사용하는 정규화 방법을 제안한다.

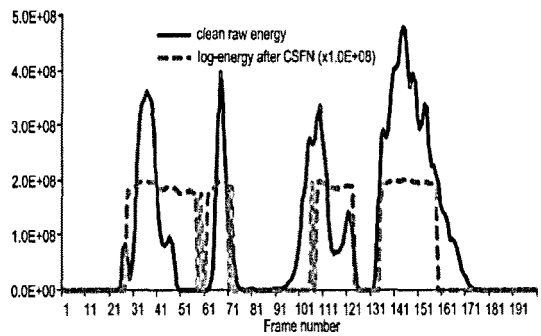

그림 4. CSFN 방법을 이용한 묵음특징 정규화
Fig. 4. Silence feature normalization using CSFN method.

그림 5는 본 논문에서 제안한 CLSFN 방법의 블록 다이어그램이다. 즉, 음성의 켑스트럼 특징으로부터 유클리디언 거리를 계산하여 기존의 로그 에너지특징과 함께 이용하여 음성/묵음을 분류한 후, 음성구간은 원래 로그 에너지를 그대로 유지하고 묵음구간은 매우 작은 값으로 정규화를 수행한다.

이때 음성의 시작 묵음구간의 평균 켑스트럼 거리는 식 (12)를 이용하여 계산한다.

$$T_2 = \frac{1}{N_F}\sum_{n=1}^{N_F}\tilde{d}[n] \quad (N_F = 30) \tag{12}$$

CLSFN 방법의 수행 절차는 다음과 같다.

$IF(\tilde{d}[t] > \alpha * T_2)$ //잡음을 음성으로 오 분류 방지
  $IF(\log\hat{E}[n] > T_0)$ //기존의 $SFN-I$ 방법
    $\log\hat{E}[n] = \log E[n]$
  $ELSE$
    $IF(\tilde{d}[t] > 3 * T_2)$ //음성을 잡음으로 오 분류 방지
      $\log\hat{E}[n] = \log E[n]$
    $ELSE$
      $\log\hat{E}[n] = \log(\epsilon) + \delta$
$ELSE$
  $\log\hat{E}[n] = \log(\epsilon) + \delta$

먼저 켑스트럼 거리를 기준값 ($\alpha * T_2$)과 비교하여, 작은 값을 갖는 구간은 묵음으로 분류하는 반면, 큰 값을 갖는 구간에서만 기존의 SFN-I 방법을 수행한다. 다음은 기존의 SFN-I 방법의 문제점인 음성을 묵음으로 오 분류하는 경우를 대비하여, 비록 묵음구간으로 분류를 하였지만 켑스트럼 거리의 값이 기준값 ($3 * T_2$)보다 크면 음성으로 판단하여 다시 음성구간의 복원시키는 절차를 수행하게 된다.

그 결과, 그림 6에서와 같이 기존의 SFN-I 방법으로

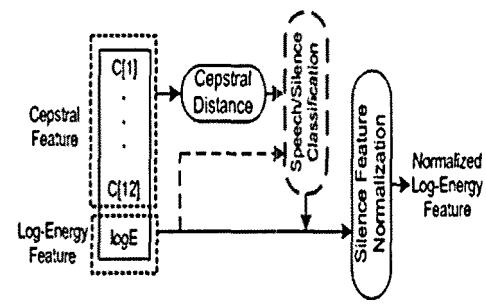

그림 5. CLSFN 방법의 블록 다이어그램
Fig. 5. Block diagram of CLSFN method.

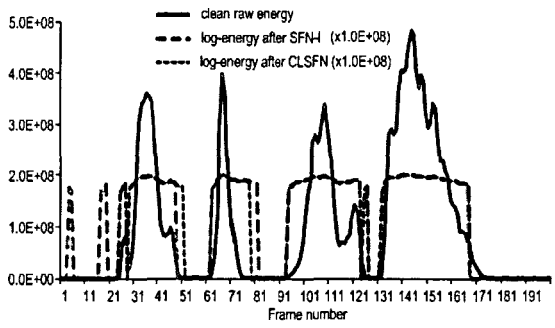

그림 6. CLSFN 방법을 이용한 묵음특징 정규화
Fig. 6. Silence feature normalization using CLSFN method.

시작묵음구간과 81번째 프레임 부분을 음성으로 오 분류하는 것을 방지할 수 있었으며 또한 51번째 프레임 부분에서 음성구간이 약간 잘리는 부분을 음성구간으로 다시 복원되는 것을 확인 할 수 있었다.

V. 실험 및 결과

5.1. 실험 환경

인식실험 및 평가를 위하여 Aurora 2.0 데이터베이스 [6]를 사용하였다. Aurora 2.0에는 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, exhibition)과 각각의 5종류의 잡음 레벨 (clean, 20 dB, 15 dB, 10 dB, 5 dB)로 구성되어 있다. 테스트 데이터는 세 종류의 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 -5 dB에서 clean 까지 7가지의 잡음 레벨로 구성되어있다.

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

5.2. 실험 결과

표 1은 기존의 여러 묵음특징 정규화방법들에 대한 인식실험 결과이다. SFN-I/II 방법은 Baseline에 비해 Set A, B에서 인식 성능이 현저히 향상된 것을 확인 할 수 있다. 하지만 Set C와 같은 채널 왜곡이 존재하는 환경에서는 두 방법 모두 효과가 없는 것으로 나타났다. 이는 로그 에너지 특징만을 사용하여 음성/묵음을 분류하였으므로 잡음의 주파수 특성을 고려해주지 못하기 때문으로 분석된다. CSFN 방법은 Set A, C에서 SFN-I/II 방법에 비해 향상된 인식결과를 나타내었다. 특히 성능향상이 주로 낮은 SNR에서 이루어졌다는 사실이 중요한 의미를 갖는다. 즉, 잡음이 많은 환경에서는 켑스트럼 거리의 분별력이 로그 에너지에 비해 우수함을 입증하고 있다는 것이다. 하지만 Set B에서는 SFN-I/II 방법의 성능에 미치지 못하고 있다. 이는 Set B에서의 airport 잡음과 같이 사람의 말소리까지 포함하는 다양한 잡음이 존재하는 불안정적인 환경에서는 켑스트럼 거리만을 이용한 음성/묵음 분류 방법도 여전히 부족함을 의미한다. 하지만 Set C에서의 성능향상은 음성의 주파수 분포 특성을 나타내는 켑스트럼 특징에 대한 유클라디언 거리를 이용하여 음성/묵음을 분류하였으므로 분류의 정확도를 향상시킬 수 있었기 때문으로 분석된다.

표 1. 기존의 묵음특징 정규화 방법에 대한 인식성능 비교
Table 1. Recognition performance comparisons of the conventional silence feature normalization methods.

Set A	Baseline	SFN-I	SFN-II	CSFN
20 dB	95.25	96.58	96.69	95.26
15 dB	87.33	93.10	93.22	92.17
10 dB	67.71	83.06	83.54	82.93
5 dB	39.48	61.04	62.33	62.78
0 dB	16.95	29.88	31.38	34.44
-5 dB	7.94	9.84	11.39	15.16
Avg.	52.44	62.25	63.09	63.79

Set B	Baseline	SFN-I	SFN-II	CSFN
20 dB	92.77	96.87	96.95	94.89
15 dB	81.34	94.28	94.43	91.65
10 dB	59.01	85.47	86.34	82.55
5 dB	31.93	64.95	66.64	61.57
0 dB	13.70	33.05	35.95	32.64
-5 dB	7.65	9.06	11.30	13.34
Avg.	47.73	63.95	65.27	62.77

Set C	Baseline	SFN-I	SFN-II	CSFN
20 dB	94.30	93.43	93.75	93.95
15 dB	87.84	84.76	85.28	89.00
10 dB	74.15	67.42	67.88	76.83
5 dB	50.24	41.79	41.50	55.20
0 dB	24.17	21.09	19.80	27.88
-5 dB	11.49	10.10	10.22	13.21
Avg.	57.03	53.10	53.07	59.34

본 논문에서 제안하는 CLSFN 방법에 대한 최적의 파

표 2. CLSFN 방법에서 α에 따른 인식성능 비교
Table 2. Recognition performance comparisons according to α in CLSFN method.

α	Set A	Set B	Set C	Avg.
1.0	65.20	65.55	57.88	62.87
1.1	65.46	65.66	57.97	63.03
1.2	65.85	65.66	58.46	63.33
1.3	65.48	65.63	58.31	63.14
1.4	65.54	66.12	57.75	63.13
1.5	65.18	65.85	57.01	62.68

표 3. CLSFN 방법에 대한 인식성능 비교 ($\alpha = 1.2$)
Table 3. Recognition performance comparisons for CLSFN method ($\alpha = 1.2$).

SNR	Set A	Set B	Set C
20 dB	96.52	96.65	94.32
15 dB	93.74	93.94	88.60
10 dB	85.82	86.50	75.08
5 dB	66.69	67.58	52.93
0 dB	36.71	36.20	27.58
-5 dB	15.65	13.11	12.27
Avg.	65.85	65.66	58.46

라미터 α의 값을 찾기 위하여, α값을 1.0에서 1.5까지 변화시켜서 인식실험을 수행하였다 (표 2). 그 결과, α=1.2에서 가장 우수한 성능을 얻을 수 있었으며, 이 경우의 실험결과를 표 3에 나타내었다. 표 3과 표 1을 비교할 경우, 제안한 CLSFN 방법은 낮은 SNR에서 CSFN 방법의 우수한 성능을 유지하였을 뿐만 아니라, 높은 SNR에서도 SFN-I 방법의 우수한 성능을 유지할 수 있어서 전반적으로 가장 우수함을 확인할 수 있었다.

각 방법에 대한 평균 인식성능을 비교한 결과를 표4에 나타내었다. 여기서 로그 에너지 정규화 방법들은 기존의 대표적인 켑스트럼 정규화 방법 (CMVN: Cepstral Mean and Variance Normalization)에 비해 전반적으로 우수한 성능을 나타냈었다. 다만, 채널 잡음환경인 Set C에서 CMVN 방법은 주파수 기반의 켑스트럼 특징을 정규화 해주므로 기존의 로그 에너지 정규화 방법 (SFN-I/II)에 비해 상대적으로 우수한 인식성능을 나타냈다.

본 논문에서 제안하는 CLSFN 방법은 기존의 대표적인 묵음특징 정규화 방법 SFN-I에 비해 Set A, B, C에서 평균 단어 정확도가 각각 약 7 %, 2 %, 5 % 향상되었음을 확인할 수 있었고, 모든 테스트 세트에 대한 평균 단어 정확도에서도 기존방법 (SFN-I, SFN-II, CSFN)에 비해 약 4 %, 3 %, 1 % 향상되었음을 확인 할 수 있어 제안한 방법의 유효성을 확인할 수 있었다.

표 4. 모든 방법에 대한 평균 인식성능 비교
Table 4. Averaged recognition performance comparisons for all methods.

Method	Set A	Set B	Set C	Avg.
baseline	52.44	47.73	57.03	52.40
CMVN	53.74	53.61	58.59	55.31
SFN-I	62.25	63.95	53.10	59.76
SFN-II	63.09	65.27	53.07	60.48
CSFN	63.79	62.77	59.34	61.97
CLSFN	65.85	65.66	58.46	63.33

V. 결 론

훈련 환경과 인식 환경의 차이가 음성인식 성능저하의 주요 요인이다. 이러한 환경의 불일치를 줄이기 위한 방법으로 다양한 묵음특징 정규화 방법이 제안되고 있다. 하지만 기존의 묵음특징 정규화방법은 로그 에너지 특징 정보만을 이용한 방법으로서 낮은 SNR에서 음성/묵음 분류의 정확도가 떨어지므로 인식성능도 같이 현저하게 저하되는 문제점이 있다. 이를 해결하기 위해 본 논문에서는 로그 에너지와 음성/묵음 (또는 잡음)의 켑스트럼 특징의 분포 특성의 차이를 나타내는 켑스트럼 유클리디언 거리를 결합해서 음성/묵음 분류에 사용하여 묵음특징을 정규화하는 CLSFN 방법을 제안하였다. 인식실험결과 기존의 SFN-I/II, CSFN 방법에 비해 전반적으로 향상된 인식성능을 얻을 수 있어 그 유효성을 확인할 수 있었다.

감사의 글

이 연구는 2008학년도 영남대학교 학술연구조성비에 의한 것임.

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. 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.
3. 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.
4. 신광호, 구엔딘쿵, 정현열, "강인한 음성인식을 위한 켑스트럼 유

클리디언 거리 기반 묵음 특징 정규화," *제26회 음성통신 및 신호처리 학술대회 논문집*, vol. 26, no. 1, pp. 265-268, 2009. 8.
5. S.E. Bou-Ghazale and K. Assaleh, "A Robust Endpoint Detection of Speech for Noisy Environments with Application to Automatic Speech Recognition," *Proc. ICASSP*, 2002.
6. H.-G Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *ISCA ITRW ASR* 2000, France, Sep. 2000.

저자 약력

•신 광 호 (Guanghu Shen)
한국음향학회지 제27권 제4호 참조

•정 현 열 (Hyun-Yeol Chung)
한국음향학회지 제27권 제4호 참조