ISSN 2005-8063
2017. 6. 30.
Vol.9 No.2
pp. 103-110

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

http://dx.doi.org/10.13064/KSSS.2017.9.2.103

잡음 환경에서 짧은 발화 인식 성능 향상을 위한 선택적 극점 필터링 기반의 특징 정규화*

Selective pole filtering based feature normalization for performance improvement of short utterance recognition in noisy environments

최 보 경 · 반 성 민 · 김 형 순**

Choi, Bo Kyeong · Ban, Sung Min · Kim, Hyung Soon

Abstract

The pole filtering concept has been successfully applied to cepstral feature normalization techniques for noise-robust speech recognition. In this paper, it is proposed to apply the pole filtering selectively only to the speech intervals, in order to further improve the recognition performance for short utterances in noisy environments. Experimental results on AURORA 2 task with clean-condition training show that the proposed selectively pole-filtered cepstral mean normalization (SPFCMN) and selectively pole-filtered cepstral mean and variance normalization (SPFCMVN) yield error rate reduction of 38.6% and 45.8%, respectively, compared to the baseline system.

Keywords: speech recognition, feature normalization, noisy environment, pole filtering

1. 서론

음성인식에서 훈련 환경과 인식 환경의 불일치 문제는 인식 성능 저하의 주된 요인이며, 이 문제의 해결을 위한 방법론은 크게 특징 영역 접근법과 모델 영역 접근법으로 분류할 수 있다[1]. 특징 영역 접근법은 모델 영역 접근법에 비해 계산량이 적고 인식엔진에 독립적이라는 장점이 있다. 특징 영역 접근법의 일종인 특징 정규화 방법은 음성 특징 파라미터들의 통계적 특성의 정규화를 통해 환경 불일치를 감소시키는 방법으로서, Cepstral Mean Normalization (CMN)[2], Cepstral Mean Variance Normalization(CMVN)[3], Cepstral Mean Scale Normalization (CMSN)[4], Histogram Equalization(HE)[5] 등 켑스트럼 정규화 방법들이 대표적인 예이다. 특히 CMN과 CMVN은 매우 적은 계산량으로 효과적인 환경 보상이 가능하기 때문에 음성인식 및 화자인식 분야에 널리 사용되고 있다.

CMN과 CMVN을 비롯한 켑스트럼 특징 정규화에 공통적으로 적용되는 평균 정규화(mean normalization) 과정은 훈련 및 인식 환경 발화들의 1 차 적률(moment)을 일치시키는 것으로서, 시불변(time-invariant) 채널 왜곡을 제거할 뿐 만 아니라 부가잡음에 대한 강인성도 높여서 음성인식 성능을 개선시킨다[1]. 다만 발화의 길이가 매우 짧은 경우, 정규화로 인한 음성 정보의 손실이 커서 성능 개선을 제한하거나 오히려 성능을 떨어뜨리는 문제점이 있다. 이 문제의 개선을 위해 본 논문의 선행연구에서는 음성인식에 가장 널리 사용되는 특징인 멜-주파수 켑스트럼 계수(Mel-Frequency Cepstral Coefficient, MFCC)를 기반으로 한 평균 정규화 방식에 극점 필터링 개념을 적용하여 잡음 환경

* 이 논문은 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음.
** 부산대학교, kimhs@pusan.ac.kr, 교신저자
Received 29 April 2017; Revised 15 June 2017; Accepted 25 June 2017

에서 음성인식 성능을 개선시켰다[6],[7]. 극점 필터링(Pole Filtering, PF)은 원래 화자인식 분야에서 선형예측 켑스트럼 계수(Linear Predictive Cepstral Coefficient, LPCC)에 평균 정규화를 적용 할 때 채널 성분 추정의 정확도 향상을 위해 제안된 방법인데[8], 이 아이디어를 MFCC 특징에도 적용할 수 있고 결과적으로 잡음 환경의 보상에도 효과적임을 확인하였다.

다만 선행 연구에서는 음성 구간과 비음성 구간을 모두 포함한 발화 전체에 대해 극점 필터링을 적용하였는데, 극점 필터링이 켑스트럼 평균의 제거 과정에서 음성 성분의 제거를 줄이는 역할을 하는 것을 감안하여, 본 논문에서는 비음성 구간은 제외하고 음성 구간에 대해서만 선택적으로 극점 필터링을 적용함으로써 특징 정규화의 성능을 향상시키는 방안을 제안한다.

본 논문의 구성은 다음과 같다. 서론에 이어 2 장에서는 짧은 발화에 대한 기존의 특징 정규화 방법들의 문제점과 극점 필터링을 통해 이 문제점을 어떻게 완화시킬 수 있는지에 관한 검토 내용을 다루고, 3 장에서는 극점 필터링 기반 특징 정규화의 추가적인 성능 향상을 위해 음성 구간에 대해서만 선택적으로 극점 필터링을 적용하는 방법을 제안한다. 4 장에서는 실험 및 결과에 대해 기술하고, 마지막으로 5 장에서 결론을 맺는다.

2. 짧은 발화에 대한 특징 정규화의 문제점 및 극점 필터링의 역할

2.1. 짧은 발화에 대한 특징 정규화의 문제점

켑스트럼 평균 정규화(CMN)[2]는 켑스트럼 시간열에서 장구간 켑스트럼 평균을 빼주는 가장 간단한 정규화 방법이다. 평균 정규화 이후 켑스트럼 벡터열의 모든 차수별 평균은 0이 되는데, 이는 1차 적률을 정규화 하는 특성뿐만 아니라 발화에 존재하는 미지의 채널 성분을 제거하는 특성도 가지고 있다[1].

T개의 프레임으로 구성된 발화의 켑스트럼 특징벡터열 $\boldsymbol{X}=[\boldsymbol{x_1}\ \cdots\ \boldsymbol{x_t}\ \cdots\ \boldsymbol{x_T}]$ 에서 $x_t(i)$가 t번째 프레임의 특징벡터인 $\boldsymbol{x_t}$ 의 i번째 성분이라고 할 때, 평균 정규화를 통해 켑스트럼 벡터 열에서 채널 성분이 제거되는 과정을 간략히 설명하면 다음과 같다. 부가잡음이 없고 시불변 채널 왜곡만 있는 상황에서 채널의 임펄스 응답의 길이가 프레임 길이에 비해 충분히 짧다면, 켑스트럼 영역에서 원음성과 채널 특성을 거친 음성의 관계는 다음과 같이 표현할 수 있다.

$$x_t(i) = d_t(i) + h(i) \tag{1}$$

여기서 $d_t(i)$와 $x_t(i)$는 t번째 프레임에 해당하는 원음성과 채널 통과 음성의 i번째 켑스트럼 계수, $h(i)$는 시불변 채널 특성의 i번째 켑스트럼 계수이다. 식 (1)에 CMN을 적용하면

$$\begin{aligned} x_{t,\,CMN}(i) &= x_t(i) - \frac{1}{T}\sum_{t=1}^{T} x_t(i) \\ &= d_t(i) + h(i) - \{\mu_d(i) + h(i)\} \end{aligned} \tag{2}$$

와 같고, 이 때

$$\mu_d(i) = \frac{1}{T}\sum_{t=1}^{T} d_t(i) \tag{3}$$

는 원음성의 켑스트럼 평균 벡터의 i번째 성분이다. 최종적으로 식 (2)는 다음과 같이 정리된다.

$$x_{t,\,CMN}(i) = d_t(i) - \mu_d(i) \tag{4}$$

식 (2)와 (4)에서 보는 바와 같이, CMN을 통해 미지의 채널 성분 $h(i)$이 제거됨과 동시에 음성의 켑스트럼 평균 $\mu_d(i)$도 추가적으로 차감됨을 알 수 있다. 만약 모든 발화의 켑스트럼 평균이 동일하다고 가정하면 $\mu_d(i)$는 상수가 되어 음성인식 성능에 아무런 부정적 영향을 미치지 않으며, 실제로 길이가 긴 발화의 경우 많은 음소들이 비교적 골고루 분포되므로 이 가정이 근사적으로 성립한다. 그러나 발화 길이가 짧을 경우 발화내 음소 분포가 제한적이고, 따라서 발화마다 켑스트럼 평균 $\mu_d(i)$의 차이가 커져서 결과적으로 음성인식 성능을 저하시키는 새로운 변이 요인으로 작용하게 된다.

2.2. 극점 필터링의 역할

서론에서 언급한 것처럼 극점 필터링은 화자인식 분야에서 선형예측 켑스트럼 계수(LPCC)에 평균 정규화를 적용 할 때, 채널 성분 추정의 정확도 향상을 위해 제안된 방법이다[8].

LPCC의 값을 결정하는 전극 모델(all-pole model)의 협대역 극점(pole)은 스펙트럼 상에서 현저한 포먼트(formant)를 나타내고, 이는 유용한 음성 성분의 특성을 가지기 때문에 LPCC를 기반으로 추정된 켑스트럼 평균에는 채널 성분과 음성 성분이 함께 존재하는 문제점이 있다. 이러한 문제 해결을 위해 제안된 극점 필터링은 전극 모델에 속한 협대역 극점의 대역폭을 확장, 포먼트를 평활화 함으로써 유용한 음성 성분의 영향을 감쇠시켜 이후 LPCC 켑스트럼 평균에 포함되는 음성 성분의 비중을 줄여주는 역할을 한다.

극점 필터링의 방식에는 선별된 협대역 극점의 대역폭만을 변경하는 방식과 모든 극점의 대역폭을 일률적으로 변경하는 방식의 2 가지가 있다[8]. 선별된 협대역 극점의 대역폭만을 변경하는 방식은 단위원에 가까이 위치한 협대역 극점의 주파수는 유지하되 그 크기를 1 보다 작은 문턱값 α로 변경시킴으로써 해당 극점의 대역폭을 넓게 보정하는 방식이다.

이에 비해 모든 극점의 대역폭을 일률적으로 변경하는 방식은, 단위원 내부에 존재하는 모든 극점의 대역폭을 동일하게 확장하는 방식이다. 전극 모델과 LPCC의 관계식으로부터 i번째 LPCC, $c_{LPCC}(i)$는 다음과 같이 표현된다[9].

$$c_{LPCC}(i) = \frac{1}{i}\sum_{k=1}^{P} z_k^i \tag{5}$$

여기서 z_k는 전극모델의 k번째 극점을 나타내며, P는 전극모델의 차수이다. 이로부터 모든 극점 z_k의 크기를 γ $(0<\gamma<1)$ 만큼의 비율로 감쇠시켜 대역폭을 일률적으로 확장시키는 형태로 극점 필터링을 거친 $c_{PFCC}(i)$의 수식은 다음과 같다. (여기서 PFCC는 Pole-Filtered Cepstral Coefficient의 약자임.)

$$c_{PFCC}(i)=\frac{1}{i}\sum_{k=1}^{P}\gamma^i z_k^i \tag{6}$$

따라서, $c_{PFCC}(i)$와 $c_{LPCC}(i)$의 관계식은

$$c_{PFCC}(i)=\gamma^i c_{LPCC}(i) \tag{7}$$

이 되고, 즉, LPCC의 i번째 차수에 단지 γ^i를 곱해줌으로써 구현된다. 본 논문에서의 극점 필터링은 이 방식으로 구현되었으며, 이 방식은 협대역 극점의 대역폭만을 선별적으로 변경하지는 못하지만, 켑스트럼 영역에서 간단하게 구현 가능하여 계산량이 적게 소요되며, 무엇보다도 LPCC가 아닌 다른 켑스트럼, 즉, MFCC 등에도 그대로 적용할 수 있다는 장점이 있다.

2.3. 극점 필터링 기반의 특징 정규화 방법

본 연구의 선행연구에서 음성인식에 가장 널리 사용되는 MFCC 특징에 극점 필터링을 적용함으로써, 잡음 환경에서의 짧은 발화 특징 정규화의 성능이 개선됨을 확인하였다[6],[7]. 실제로 켑스트럼 특징 정규화 방법인 CMN 및 CMVN 각각의 평균 정규화 과정에 극점 필터링을 적용하는 방식으로 구현되었으며, 이들은 각각 Pole-Filtered CMN(PFCMN)과 Pole-Filtered CMVN(PFCMVN)이라고 명명되었다. 그 수식은 다음과 같다.

$$x_{t,PFCMN}(i) = x_t(i)-\mu_{PF}(i), \quad 1\le t\le T \tag{8}$$

$$x_{t,PFCMVN}(i)=\frac{x_t(i)-\mu_{PF}(i)}{\sigma_{PF}(i)}, \quad 1\le t\le T \tag{9}$$

여기서

$$\mu_{PF}(i)=\frac{1}{T}\sum_{t=1}^{T}\gamma^i x_t(i)=\gamma^i\mu(i) \tag{10}$$

$$\sigma_{PF}(i)=\sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_t(i)-\mu_{PF}(i))^2} \tag{11}$$

이다. 이 방식에 사용되는 평균은 식 (10)에서 보는 바와 같이 장구간 켑스트럼 평균, $\mu(i)$에 바로 극점 필터링을 적용해줌으로써 구현가능한데, 이 때 γ 값$(0<\gamma<1)$은 실험적으로 결정된다. 참고로 만약 $\gamma=1$이 되면 식 (8), (9)는 기존의 CMN, CMVN 수식과 동일하다.

상기 수식과 같이, 기존 특징 정규화 방식들의 평균 정규화 과정에 극점 필터링을 적용하는 것이 짧은 발화에 대해 더 효과적인지 검토하기 위하여, <그림 1>에 극점 필터링 적용 여부에 따른 켑스트럼 평균의 역변환 결과인 멜-필터뱅크(Mel-filterbank) 출력 값들을 발화별로 중첩하여 나타내었다. <그림 1>의 (a)와 (b)는 각각 AURORA 2 데이터베이스(database, DB)[10]에 포함된 7 자리 연속숫자와 1 자리 숫자들의 일부에 대해 발화별 평균들의 편차를 보여 주며, 그림으로부터 발화 길이가 짧을 때 편차가 더 크다는 것을 알 수 있다. 이때 사용된 발화들이 모두 clean 음성이므로, 이 편차들은 전적으로 발화 내 음성정보의 차이에 기인한 것이다. <그림 1>의 (c)와 (d)를 통해 상대적으로 긴 발화나 짧은 발화 모두 극점 필터링을 통해 발화별 평균들의 편차가 줄어들고, 특히 짧은 발화에 대해 상대적인 감쇠 폭이 더 크다는 것을 관찰할 수 있다. 이로부터 극점 필터링이 짧은 발화에 대해 더 효과적임을 예상할 수 있으며, 실제 음성인식 실험을 통한 검증은 4.2 절에서 다룬다.

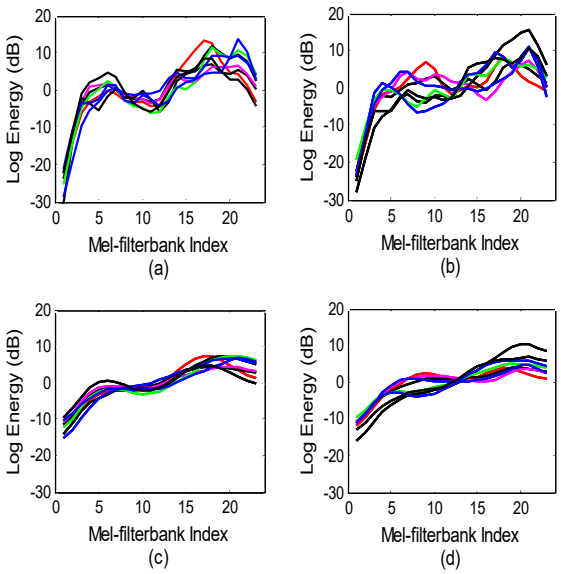

그림 1. 발화 길이에 따른 극점 필터링의 효과
(a) 7자리 숫자 발화들의 멜-필터뱅크 출력 평균
(b) 1자리 숫자 발화들의 멜-필터뱅크 출력 평균
(c) 7자리 숫자 발화들의 극점 필터링된 멜-필터뱅크 출력 평균
(d) 1자리 숫자 발화들의 극점 필터링된 멜-필터뱅크 출력 평균

Figure 1. Effects of pole-filtering according to utterance length
(a) Mean of Mel-filterbank output for 7-digit utterances
(b) Mean of Mel-filterbank output for 1-digit utterances
(c) Mean of pole-filtered Mel-filterbank output for 7-digit utterances
(d) Mean of pole-filtered Mel-filterbank output for 1-digit utterances

3. 음성/비음성 구분을 통한 극점 필터링 기반의 특징 정규화

극점 필터링 기반의 특징 정규화 시, 추정된 켑스트럼 평균에 포함된 음성 성분을 보다 정확하게 감쇠시켜 주기 위해서는, 음성 구간과 비음성 구간을 모두 포함하는 발화 전체에서의 켑스

트럼 평균에 극점 필터링을 적용해주는 것보다, 음성 구간에 대한 켑스트럼 평균에만 극점 필터링을 적용해주는 것이 더 효과적이라고 판단된다. 본 논문에서는 발화 내 음성 구간과 비음성 구간을 구분하여, 음성 구간의 켑스트럼 평균에만 선택적으로 극점 필터링을 적용 해주는 특징 정규화 방식을 제안한다. 이와 관련된 연구로, 음성/비음성 평균을 따로 추정하여 특징을 정규화 하는 augmented CMN 방식이 제안된 바 있다[11].

음성 구간과 비음성 구간을 구분하는 방법으로는 여러 방식들이 제안되었지만, 본 논문에서는 프레임별 로그 에너지 값들의 분포를 음성과 비음성에 해당하는 정규분포들의 가중합 형태인 가우스 혼합 모델(Gaussian Mixture Model, GMM)로 모델링한 후[12], 이로부터 구한 각 프레임별 음성존재확률(Speech Presence Probability, SPP)을 이용하여 음성과 비음성을 구분하는 방식을 사용한다[12],[13]. 통상적으로 로그 에너지 값이 음성특징 벡터의 0 번째 차원이므로, $x_t(0),\ t=1,2,...,T$ 는 한 발화의 프레임별 로그 에너지 값들의 시간열을 나타낸다. 이로부터 2 개의 혼합(mixture)을 가지는 GMM을 훈련시키면, 음성 및 비음성 구간 각각에 대한 로그 에너지 분포의 가중치, 평균 및 분산, 즉, $\{w_S, \mu_S, \sigma_S^2\}$와 $\{w_{NS}, \mu_{NS}, \sigma_{NS}^2\}$이 추정된다. 이로부터 t번째 프레임이 음성 구간에 해당할 확률, 즉, 음성존재확률(SSP)은 다음과 같이 구할 수 있다.

$$P_S(t)=\frac{\omega_S N(x_t(0)|\mu_S,\sigma_S^2)}{\omega_S N(x_t(0)|\mu_S,\sigma_S^2)+\omega_{NS} N(x_t(0)|\mu_{NS},\sigma_{NS}^2)} \quad (12)$$

여기서 $N(x|\mu,\sigma^2)$은 평균이 μ이고 분산이 σ^2인 정규분포 함수이다. SPP 추정의 신뢰도를 높이기 위해 프레임별 로그 에너지의 시간열을 평활화(smoothing)하여 GMM 훈련에 사용한다. 본 논문에서는 창 크기가 11 인 이동 평균(moving average) 함수를 통해 평활화 과정을 수행하였다.

음성/비음성 구분은 SSP를 특정 문턱값(θ)과 비교하여 결정한다. 즉, $P_S(t)\geq\theta$이면, t번째 프레임을 음성구간으로 판정하고, 그렇지 않으면 비음성 구간으로 판정한다. <그림 2>는 발화별로 음성/비음성 구분을 위한 문턱값를 구하는 과정을 나타낸다. 이 그림은 한 발화에 대해 평활화를 거친 프레임별 로그 에너지의 히스토그램과 더불어 이들을 2 개의 가중 정규분포로 모델링한 결과를 보여준다. 그리고 문턱값 θ는 음성과 비음성에 해당하는 2 개의 가중 정규분포가 만나는 지점으로 결정한다. 즉, 다음 식을 만족하는 θ 값이 문턱치가 된다.

$$\omega_S N(\theta|\mu_S,\sigma_S^2)=\omega_{NS} N(\theta|\mu_{NS},\sigma_{NS}^2) \quad (13)$$

다만 신호대잡음비(signal-to-noise ratio, SNR)가 매우 낮은 경우에는, 잡음 섞인 신호의 로그 에너지의 분포를 음성과 비음성의 두 정규분포로 모델링하기 어려운 문제가 있어서[13], SNR 0 dB 및 −5 dB의 열악한 환경의 발화에 대해서는 European Telecommunications Standards Institute(ETSI)의 Advanced Front-End(AFE)[14] 전처리 방식, 즉, 2 단계 Mel-warped Wiener 필터를 통과하여 잡음이 보상된 음성의 로그 에너지를 사용하여 SPP를 추정하는 방식을 통해 음성/비음성 구간을 구분하도록 하였다. 물론 MFCC 특징 파라미터에는 AFE를 적용하지 않았다.

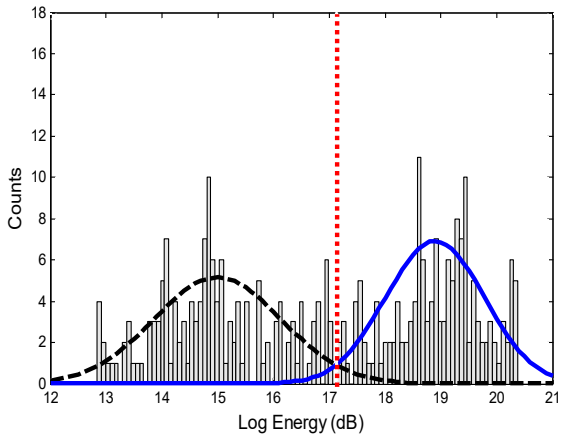

그림 2. 로그 에너지 분포 모델링과 문턱값 설정의 예
Figure 2. An example of modeling of log energy distribution and determination of threshold

각 발화에 대한 프레임별 SPP 값과 해당 발화의 문턱값(θ)이 정해지면, 음성 구간과 비음성 구간에 대한 각각의 켑스트럼 평균 벡터의 i번째 차원은 다음 식과 같이 구할 수 있다.

$$\mu_S(i)=\frac{\sum_{t=1}^{T} I_\theta(x_t(0))x_t(i)}{\sum_{t=1}^{T} I_\theta(x_t(0))} \quad (14)$$

$$\mu_{NS}(i)=\frac{\sum_{t=1}^{T}\{1-I_\theta(x_t(0))\}x_t(i)}{\sum_{t=1}^{T}\{1-I_\theta(x_t(0))\}} \quad (15)$$

여기서 $I_\theta(x)$는 지시(indicator) 함수로서 다음과 같이 정의된다.

$$I_\theta(x)=\begin{cases}1, & x\geq\theta\\ 0, & x<\theta\end{cases} \quad (16)$$

식 (14)와 (15)의 방법은 로그 에너지에 대한 문턱값을 기준으로 매 프레임을 음성 구간 또는 비음성 구간으로 경판정(hard decision)한 경우이다. 그런데 SPP를 잘 활용하면 경판정 대신에 연판정(soft decision)을 통해 음성 구간과 비음성 구간의 켑스트럼 평균을 구할 수도 있으며, 이는 다음 식과 같다.

$$\mu_S(i) = \frac{\sum_{t=1}^{T} P_S(t) x_t(i)}{\sum_{t=1}^{T} P_S(t)} \tag{17}$$

$$\mu_{NS}(i) = \frac{\sum_{t=1}^{T} P_{NS}(t) x_t(i)}{\sum_{t=1}^{T} P_{NS}(t)} \tag{18}$$

식 (18)에서 $P_{NS}(t)$는 t번째 프레임이 비음성 구간에 해당할 확률로서, $P_{NS}(t) = 1 - P_S(t)$이다.

상기 두 가지 방식으로 음성 구간과 비음성 구간에 대한 켑스트럼 평균을 구한 다음, 본 논문에서는 음성 구간의 켑스트럼 평균인 $\mu_S(i)$에만 선택적 극점 필터링(selective pole filtering, SPF)를 적용하여 CMN과 CMVN을 구현하는 방식을 제안하며, 이들을 각각 Selectively Pole-Filtered CMN(SPFCMN) 및 Selectively Pole-Filtered CMVN (SPFCMVN)이라 명명한다. 이들 각각을 수식으로 나타내면 다음과 같다.

$$\begin{aligned} &\text{If } (x_t(0) \geq \theta) \\ &\quad x_{t,SPFCMN}(i) = x_t(i) - \mu_{S,PF}(i) \\ &\text{else} \\ &\quad x_{t,SPFCMN}(i) = x_t(i) - \mu_{NS}(i) \end{aligned} \tag{19}$$

$$\begin{aligned} &\text{If } (x_t(0) \geq \theta) \\ &\quad x_{t,SPFCMVN}(i) = \frac{x_t(i) - \mu_{S,PF}(i)}{\sigma_{S,PF}(i)} \\ &\text{else} \\ &\quad x_{t,SPFCMVN}(i) = \frac{x_t(i) - \mu_{NS}(i)}{\sigma_{NS}(i)} \end{aligned} \tag{20}$$

여기서 $\mu_{S,PF}(i)$, $\sigma_{S,PF}(i)$와 $\sigma_{NS}(i)$의 수식은 다음과 같다.

$$\mu_{S,PF}(i) = \gamma^i \mu_S(i) \tag{21}$$

$$\sigma_{S,PF}(i) = \sqrt{\frac{\sum_{t=1}^{T} I_\theta(x_t(0))(x_t(i) - \mu_{S,PF}(i))^2}{\sum_{t=1}^{T} I_\theta(x_t(0))}} \tag{22}$$

$$\sigma_{NS}(i) = \sqrt{\frac{\sum_{t=1}^{T} \{1 - I_\theta(x_t(0))\}(x_t(i) - \mu_{NS}(i))^2}{\sum_{t=1}^{T} \{1 - I_\theta(x_t(0))\}}} \tag{23}$$

식 (21)에서의 γ 값$(0 < \gamma < 1)$도 식 (10)의 경우와 마찬가지로 실험적으로 결정한다.

4. 실험 및 결과

4.1. 실험 환경

다양한 특징 정규화 방식들의 성능을 평가하기 위해 잡음과 채널 왜곡의 영향이 반영된 AURORA 2 평가 환경을 그대로 사용하였다[10]. AURORA 2 DB는 미국인 화자가 발성한, 1~7자리의 연속 숫자로 구성된 clean TIdigit DB에 실제 환경의 SNR별 잡음을 더한 뒤, International Telecommunication Union(ITU)에서 정의한 두 개의 채널인 G.712 또는 Modified Intermediate Reference System(MIRS)를 통과시킨 데이터이다. AURORA 2 DB에 사용된 부가잡음으로는 열차, 군중소음(babble), 자동차, 전시장, 음식점, 거리, 공항, 기차역의 8가지 종류가 사용되었으며, SNR 범위는 -5 dB에서 20 dB까지이다.

특징 벡터는 스펙트럼 크기(magnitude spectrum)로부터 구해진 12 차 MFCC와 로그 에너지, 그리고 그 각각에 대한 델타, 델타-델타 파라미터를 포함한 총 39 차 특징을 사용하였다. 음향 모델은 AURORA 2 베이스라인 시스템과 동일하게 단어 단위의 은닉 마르코프 모델(Hidden Markov Model, HMM)로 16 개 상태의 left-to-right 모델을 사용하였고, 각 상태 당 가우스 혼합의 수는 3 개이다. 음향 모델 훈련에는 AURORA 2 DB에 정의된 clean-condition DB를 사용하였다.

4.2. 발화 길이에 따른 PFCMN 와 PFCMVN 의 성능평가

이전 논문[6],[7]에서 제안한 극점 필터링 기반의 특징 정규화 방식이 특히 짧은 발화에 대해 더 효과적인지를 확인하기 위해, 발화의 길이 별로 테스트 데이터를 분류하여 인식하는 실험을 수행하였다. 1~7 자리의 연속 숫자로 구성된 전체 발화를 음향 모델에 인식한 결과는 <표 1>의 (a)와 같고, 5~7 자리의 연속 숫자로 구성된 상대적으로 긴 발화, 3~4 자리의 연속 숫자로 구성된 중간 발화, 그리고 1~2 자리의 연속 숫자로 구성된 짧은 발화를 인식한 결과는 각각 <표 1>의 (b), (c), (d)에 나타내었다. <표 1>의 (e)에 (a)~(d) 결과의 평균 인식률을 요약하여 정리하였으며, 이로부터 극점 필터링 방식을 통한 개선효과가 크지는 않으나 일률적인 성능향상이 얻어지며, 특히 발화의 길이가 짧을수록 개선효과가 증대되는 것을 확인할 수 있다.

4.3. 제안한 SPFCMN 및 SPFCMVN 의 성능평가

본 논문에서 새롭게 제안한 선택적 극점 필터링을 이용한 특징 정규화 방식(SPFCMN 및 SPFCMVN)에서는 식 (14)와 (15) 또는 식 (17)과 (18)에서 본 바와 같이 음성/비음성 구분 과정에서 경판정 또는 연판정을 이용한 두 종류의 선택적 평균 추정방법을 사용할 수 있다. <표 2>에 이들 두 가지 평균 추정 방법에 따른 제안 방식들의 성능을 비교해서 나타내었다. 표로부터 경판정보다 연판정에 의한 음성/비음성 평균 추정 방법이 더 효과적임을 확인 할 수 있다. 따라서 이후 실험에서는 연판정에 의한 음성/비음성 평균 추정 방법을 사용하였다.

표 1. 발화 길이에 따른 PFCMN/PFCMVN 실험 결과
(a) 1~7자리 숫자로 구성된 전체 발화에 대한 인식 결과
(b) 5~7자리 숫자로 구성된 긴 발화에 대한 인식 결과
(c) 3~4자리 숫자로 구성된 중간 발화에 대한 인식 결과
(d) 1~2자리 숫자로 구성된 짧은 발화에 대한 인식 결과
(e) 발화 길이에 따른 CMN/CMVN 대비 오류 감소율(%)

Table 1. Experimental results of PFCMN/PFCMVN according to the length of utterance
(a) Recognition results for all utterances composed of 1-7 digits
(b) Recognition results for long utterances composed of 5-7 digits
(c) Recognition results for medium utterances composed of 3-4 digits
(d) Recognition results for short utterances composed of 1-2 digits
(e) Error rate reduction (%) against CMN/CMVN according to the length of utterance
(% Word Accuracy)

(a)

SNR	Baseline	CMN	PFCMN	CMVN	PFCMVN
clean	99.03	99.03	97.43	98.91	98.23
20 dB	94.07	96.91	96.99	95.90	95.98
15 dB	85.04	92.72	93.83	91.21	92.75
10 dB	65.51	78.19	84.32	79.53	85.06
5 dB	38.61	47.25	63.37	56.06	68.86
0 dB	17.09	23.80	24.37	25.57	28.12
-5 dB	8.53	13.08	13.54	10.59	11.33
Average btw 0 and 20 dB	60.06	67.77	68.65	69.65	71.13

(b)

SNR	Baseline	CMN	PFCMN	CMVN	PFCMVN
clean	99.04	99.01	99.06	98.97	99.00
20 dB	94.86	96.65	97.01	96.32	96.41
15 dB	87.49	91.67	92.46	91.87	92.38
10 dB	70.52	75.22	76.59	80.87	81.73
5 dB	45.00	40.62	41.57	57.60	59.07
0 dB	23.05	16.53	17.04	25.61	27.98
-5 dB	12.08	9.46	9.75	10.28	11.11
Average btw 0 and 20 dB	64.18	64.14	64.93	70.45	71.51

(c)

SNR	Baseline	CMN	PFCMN	CMVN	PFCMVN
clean	98.83	98.99	98.85	98.75	98.76
20 dB	94.26	96.82	97.27	95.99	96.50
15 dB	86.23	92.86	93.82	91.39	92.27
10 dB	67.93	78.58	80.27	79.68	81.19
5 dB	42.62	47.44	49.03	55.14	57.75
0 dB	20.59	24.23	24.43	25.27	28.53
-5 dB	11.12	13.98	14.24	11.17	11.71
Average between 0 and 20dB	62.33	67.99	68.96	69.49	71.25

(d)

SNR	Baseline	CMN	PFCMN	CMVN	PFCMVN
clean	99.33	99.13	99.24	99.03	99.09
20 dB	91.30	97.83	98.17	94.51	95.54
15 dB	75.44	95.62	96.29	88.94	90.12
10 dB	45.88	86.46	87.78	75.31	78.25
5 dB	11.94	66.88	68.02	53.12	56.19
0 dB	-7.27	44.88	49.29	25.93	28.46
-5 dB	-6.91	22.40	23.93	10.44	12.17
Average btw 0 and 20 dB	43.46	78.34	79.31	67.56	69.71

(e)

	All	Long	Medium	Short
CMN vs PFCMN	2.73	2.20	3.03	4.52
CMVN vs PFCMVN	4.88	3.59	5.76	6.63

표 2. 선택적 평균 추정 방법에 따른 SPFCMN/SPFCMVN 실험 결과

Table 2. Experiment results of SPFCMN/SPFCMVN according to selective mean estimation methods
(% Word Accuracy)

SNR	SPFCMN		SPFCMVN	
	Hard decision	Soft decision	Hard decision	Soft decision
clean	98.74	98.59	98.25	98.67
20 dB	96.49	97.39	94.67	96.58
15 dB	92.60	94.40	91.02	93.97
10 dB	82.60	85.28	83.59	87.32
5 dB	61.30	64.40	67.07	71.12
0 dB	34.01	35.98	40.27	42.79
-5 dB	15.63	16.17	17.01	18.10
Average btw 0 and 20 dB	73.40	75.49	75.32	78.36

그 다음으로 본 논문에서 제안한 선택적 극점 필터링 방식과 이전 논문[6],[7]에서 제안한 기존의 극점 필터링 방식에 의한 잡음 환경 특징 정규화 성능을 비교 평가하였다. <표 3>에 기존의 CMN/CMVN 평균 정규화 과정에 극점 필터링(PF)을 적용한 방식인 PFCMN/PFCMVN과 선택적 극점 필터링(SPF)을 적용한 방식인 SPFCMN/SPFCMVN 각각에 대한 성능을 나타내었다. PFCMN/PFCMVN 방식들과 SPFCMN/SPFCMVN 방식들 모두 clean 환경을 제외한 모든 잡음 레벨에 대해 CMN/CMVN 대비 일률적인 성능향상을 보이며, 특히 SPFCMN/SPFCMVN 방식들의 성능 향상 폭이 큰 것을 알 수 있다.

<표 3>의 (a)와 (b) 각각의 맨 아래 행에는 CMN/CMVN, PFCMN/PFCMVN 및 SPFCMN/SPFCMVN 각각에 대한 베이스라인 방식 대비 오류 감소율, 그리고 PFCMN/PFCMVN 및 SPFCMN/SPFCMVN 각각에 대해 CMN/CMVN 대비 오류 감소율을 나타내었다. 베이스라인 방식과 비교해서 CMN/CMVN은 각각 19.30%, 24.01%의 오류 감소율, PFCMN/PFCMVN은 각각 21.51%, 27.72%의 오류 감소율, 그리고 SPFCMN/SPFCMVN은 각각 38.63%, 45.82%의 오류 감소율을 얻어, 본 논문에서 제안한 선택적 극점 필터링을 이용한 방식이 가장 뛰어난 성능을 나타냄을 확인 할 수 있다.

<그림 3>은 식 (10) 및 (21)에서의 γ 값에 따른 PF 및 SPF 기반의 특징 정규화 방식들의 평균 인식률(SNR 0~20 dB 범위)을 나타낸 것이다. 그림에서 PFCMN/PFCMVN은 각각 $\gamma = 0.8, 0.85$에서 최적의 인식률을 보이고, SPFCMN/SPFCMVN은 각각 $\gamma = 0.65, 0.85$에서 최적의 인식률을 보인다. 그리고 PFCMVN의 경우 제한된 γ 범위에 대해서만 기존의 CMVN보다 개선된 성능을 나타낸 반면에, SPFCMVN은 도시된 모든 γ 범위에 대해 CMVN보다 훨씬 우수한 성능을 보임을 알 수 있다.

표 3. 극점 필터링 방식들에 따른 CMN/CMVN 실험 결과
(a) PF, SPF 방식들을 CMN에 적용한 인식 결과
(b) PF, SPF 방식들을 CMVN에 적용한 인식 결과

Table 3. Experimental results of CMN/CMVN according to the pole-filtering methods
(a) Recognition results of applying PF/SPF methods to CMN
(b) Recognition results of applying PF/SPF methods to CMVN
(% Word Accuracy)

(a)

SNR	Baseline	CMN	PFCMN	SPFCMN
clean	99.03	99.03	97.43	98.59
20 dB	94.07	96.91	96.99	97.39
15 dB	85.04	92.72	93.83	94.40
10 dB	65.51	78.19	84.32	85.28
5 dB	38.61	47.25	63.37	64.40
0 dB	17.09	23.80	24.37	35.98
-5 dB	8.53	13.08	13.54	16.17
Average between 0 and 20 dB	60.06	67.77	68.65	75.49
ERR compared to baseline	N/A	19.30	21.51	38.63
ERR compared to CMN	N/A	N/A	2.73	23.95

(b)

SNR	Baseline	CMVN	PFCMVN	SPFCMVN
clean	99.03	98.91	98.23	98.67
20 dB	94.07	95.90	95.98	96.58
15 dB	85.04	91.21	92.75	93.97
10 dB	65.51	79.53	85.06	87.32
5 dB	38.61	56.06	68.86	71.12
0 dB	17.09	25.57	28.12	42.79
-5 dB	8.53	10.59	11.33	18.10
Average between 0 and 20 dB	60.06	69.65	71.13	78.36
ERR compared to baseline	N/A	24.01	27.72	45.82
ERR compared to CMVN	N/A	N/A	4.88	28.70

본 연구의 제한점은 다음과 같다. 본 논문에서는 음성/비음성 구분을 위해 프레임별 로그 에너지 특징을 사용하였는데, SNR 0 dB 이하에서는 로그 에너지로 음성/비음성 구분이 용이하지 않아서, 그런 경우에 대해서만 3 장에서 이미 언급한 바와 같이 ETSI AFE 전처리 방식을 통해 잡음 제거된 음성의 로그 에너지를 사용하였다. ETSI AFE의 계산량이 아주 많은 것은 아니나, 본 논문에서 다루고자 하는 보다 경량의 특징 정규화 방식에 도구로 사용되기는 적합하지 않으며, 잡음 환경에 강인하면서도 보다 경량의 음성검출 기술을 사용하는 것이 바람직하다고 판단된다. 또한, 본 본문에서는 음성인식 방식으로 전통적인 GMM-HMM 방식을 사용했으나, 현재 음성인식 방식의 주류는 심층신경망(Deep Neural Network, DNN) 기반 방식이다. 실제로 심층신경망 기반의 음성인식에서도 발화 단위의 특징 정규화가 널리 사용되고 있기 때문에[15], 본 논문의 아이디어가 짧은 발화의 인식 성능 개선에 도움이 될 수 있을 것이라 판단되며, 이 부분은 추후 연구를 통해 검증하고자 한다.

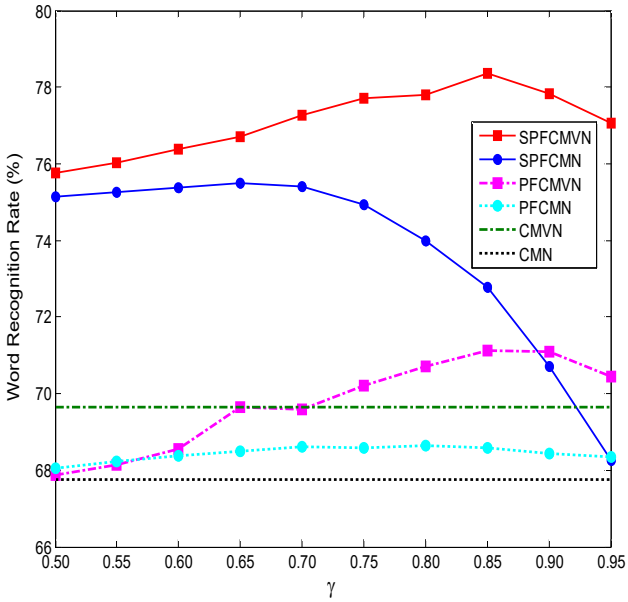

그림 3. γ 값에 따른 여러 특징 정규화 방식들의 평균 인식률
Figure 3. Average recognition rates of various feature normalization methods according to γ value

5. 결론

본 논문에서는 MFCC 특징의 발화 단위 평균 정규화에 극점 필터링 개념을 적용하여 짧은 발화에 대한 특징 정규화의 문제점을 완화시킴으로써, 잡음 환경에서의 음성인식 성능을 개선하고자 하였다.

먼저 극점 필터링을 이용한 특징 정규화 방식이 특히 짧은 발화에 대해 더 효과적인지를 확인하기 위해, 발화의 길이 별로 테스트 데이터를 분류하여 인식 실험을 수행한 결과, 발화의 길이가 짧을수록 성능이 향상되는 것을 확인하였다. 그 다음으로 극점 필터링에 의한 성능개선 효과를 극대화하기 위해 음성 구간에 대해서만 선별적으로 극점 필터링을 적용하는 새로운 방식을 제안하였다. AURORA 2 DB의 clean-condition 훈련 환경에서의 평가 결과, 본 논문에서 제안한 SPFCMN 및 SPFCMVN 방식이 기존의 CMN 및 CMVN 방식 대비 각각 24.0% 및 28.7%의 오류 감소율을 보였고, 선행 연구 결과인 PFCMN 및 PFCMVN 방식에 비해서도 각각 21.8% 및 25.0%의 성능향상을 얻었다.

앞으로 심층신경망 기반의 음성인식에 제안된 방식의 아이디어를 적용하여 추가적인 성능 향상을 도모하는 연구를 계속 진행할 예정이다.

참고문헌

[1] Li, J., Deng, L., Gong, Y., & Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4), 745-777.

[2] Atal, B. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55(6), 1304-1312.

[3] Viikki, O., Bye, D., & Laurila, K. (1998). A recursive feature vector normalization approach for robust speech recognition in noise. *Proceedings of the IEEE ICASSP* (pp. 733-736).

[4] Alam, J., Ouellet, P., Kenny, P., & O'Shaughnessy, D. (2011). Comparative evaluation of feature normalization techniques for speaker verification. *Proceedings of the International Conference on Nonlinear Speech Processing* (pp. 246-253).

[5] Molau, S., Hilger, F., & Ney, H. (2003). Feature space normalization in adverse acoustic conditions. *Proceedings of the IEEE ICASSP* (pp. 656-659).

[6] Choi, B., Ban, S., & Kim, H. (2015). Pole-filtered cepstral normalization methods for robust speech recognition. *Proceedings of the 2015 Spring Conference of the Korean Society of Speech Sciences* (pp. 101-102). (최보경·반성민·김형순 (2015). 강인한 음성인식을 위한 극점 필터링된 켑스트럼 정규화 방식. *한국음성학회 2015 봄학술대회 논문집*, 101-102.)

[7] Choi, B., Ban, S., & Kim, H. (2015). Cepstral feature normalization methods using pole filtering and scale normalization for robust speech recognition. *The Journal of the Acoustical Society of Korea*, 34(4), 316-320. (최보경·반성민·김형순 (2015). 강인한 음성인식을 위한 극점 필터링 및 스케일 정규화를 이용한 켑스트럼 특징 정규화 방식. *한국음향학회지*, 34(4), 316-320.)

[8] Naik, D. (1995). Pole-filtered cepstral mean subtraction. *Proceedings of the IEEE ICASSP* (pp. 157-160).

[9] Schroeder, M. R. (1981). Direct (nonrecursive) relations between cepstrum and predictor coefficients. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2), 297-301.

[10] Hirsch, H. G., & Pearce, D. (2000). The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions. *Proceedings of the ASR2000-Automatic Speech Recognition: Challenges for the New Millenium ISCA Tutorial and Research Workshop* (pp. 181-188).

[11] Acero, A., & Huang, X. (1995). Augmented cepstral normalization for robust speech recognition. *Proceedings of t*he *IEEE Automatic Speech Recognition Workshop* (pp. 146-147).

[12] Compernolle, D. V. (1989). Noise adaptation in a hidden Markov model speech recognition system. *Computer Speech and Language*, 3(2), 151-167.

[13] Ying, D., Yan, Y., Dang, J., & Soong, F. K. (2011). Voice activity detection based on an unsupervised learning framework. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(8), 2624-2633.

[14] ETSI Standard (2003). Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced frontend feature extraction algorithm; compression algorithms. *ETSI Technical Report ES 202 050*, 1.1.3.

[15] Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10), 1533-1545.

• **최보경 (Choi, Bo Kyeong)**
부산대학교 전자전기컴퓨터공학과
부산시 금정구 부산대학로63번길 2
Tel: 051-510-1704 Fax: 051-515-5190
Email: choibok15@pusan.ac.kr
관심분야: 음성인식

• **반성민 (Ban, Sung Min)**
SK텔레콤 AI사업단 음성인식기술팀
서울시 중구 을지로 100
Tel: 010-4441-1003
Email: sungmin.ban@sk.com
관심분야: 음성인식, 음성신호처리

• **김형순 (Kim, Hyung Soon)** 교신저자
부산대학교 전자공학과
부산시 금정구 부산대학로63번길 2
Tel: 051-510-2452 Fax: 051-515-5190
Email: kimhs@pusan.ac.kr
관심분야: 음성인식 및 합성, 음성신호처리