한국음향학회지 제34권 제4호 pp. 316~320 (2015)
The Journal of the Acoustical Society of Korea Vol.34, No.4 (2015)
http://dx.doi.org/10.7776/ASK.2015.34.4.316

pISSN : 1225-4428
eISSN : 2287-3775

강인한 음성인식을 위한 극점 필터링 및 스케일 정규화를 이용한 켑스트럼 특징 정규화 방식

Cepstral Feature Normalization Methods Using Pole Filtering and Scale Normalization for Robust Speech Recognition

최보경, 반성민, 김형순†

(Bo Kyeong Choi, Sung Min Ban, and Hyung Soon Kim†)

부산대학교 전자공학과

(Received June 12, 2015; accepted July 16, 2015)

초 록: 본 논문에서는 Cepstral Mean Normalization(CMN)과 Cepstral Mean and Variance Normalization (CMVN) 프레임워크에서 극점 필터링(pole filtering) 개념을 Mel-Frequency Cepstral Coefficient(MFCC) 특징 벡터에 적용한다. 또한 분산 정규화를 대신하여 스케일 정규화를 사용하는 Cepstral Mean and Scale Normalization (CMSN)의 성능을 잡음 환경 음성인식 실험을 통해 평가한다. CMN과 CMVN은 보통 발화 단위로 수행되기 때문에 짧은 발화의 경우 특징에 대한 평균과 분산의 추정 신뢰도가 보장되지 않는 문제점을 가지는데, 극점 필터링과 스케일 정규화 방식을 적용함으로 이러한 문제점을 보완할 수 있다. Aurora 2 데이터베이스를 이용한 실험 결과, 극점 필터링과 스케일 정규화를 결합한 특징 정규화 방식의 성능이 가장 높은 성능 향상을 보인다.
핵심용어: 음성인식, 특징 정규화

ABSTRACT: In this paper, the pole filtering concept is applied to the Mel-frequency cepstral coefficient (MFCC) feature vectors in the conventional cepstral mean normalization (CMN) and cepstral mean and variance normalization (CMVN) frameworks. Additionally, performance of the cepstral mean and scale normalization (CMSN), which uses scale normalization instead of variance normalization, is evaluated in speech recognition experiments in noisy environments. Because CMN and CMVN are usually performed on a per-utterance basis, in case of short utterance, they have a problem that reliable estimation of the mean and variance is not guaranteed. However, by applying the pole filtering and scale normalization techniques to the feature normalization process, this problem can be relieved. Experimental results using Aurora 2 database (DB) show that feature normalization method combining the pole-filtering and scale normalization yields the best improvements.
Keywords: Speech recognition, Feature normalization
PACS numbers: 43.72.Ne, 43.72.Ar

I. 서 론

최근 음성인식 성능이 많이 향상되었지만, 훈련 환경과 테스트 환경의 불일치는 여전히 음성인식 성능 저하의 주요한 요인이다. 이러한 환경 불일치 문제를 해결하기 위해서 많은 연구가 진행되었는데, 특징 영역 및 모델 영역에서의 보상 방식들로 크게 분류할 수 있다.[1] 이 중 특징영역에서의 보상 방식은 비교적 계산량이 적고, 음성인식 엔진에 독립적으로 사용할 수 있다는 장점을 가진다. 본 논문에서는 다양한 특징 보상 방식 중 음성인식 분야에서 가장 널리 사용되고 있는 CMN과 CMVN 방식을 이용한다.[1] CMN은 음성 켑스트럼에서 장구간 켑스트럼 평균을 빼주는 경우, 미지의 채널 특성이 제거되는

†**Corresponding author:** Hyung Soon Kim (kimhs@pusan.ac.kr)
Department of Electronics Engineering, Pusan National University, Busandaehak-Ro, Geumjeong-Gu, Busan 609-836, Republic of Korea
(Tel: 82-51-510-2452, Fax: 82-51-516-4279)
"이 논문의 일부는 2015년 한국음성학회 춘계학술대회에서 발표되었습니다.[7]"

성질을 이용한다. 하지만 켑스트럼의 평균에는 채널 특성 이외에 음성 자체의 켑스트럼 평균도 포함되어 있고, 입력 발화가 포함하는 어휘에 따라 켑스트럼의 평균이 달라진다. 또한 켑스트럼의 분산 역시 발화의 길이가 짧을수록 추정치의 편차가 커져서 추정치의 신뢰도가 떨어지게 된다.

본 논문에서는 이러한 문제를 완화시키기 위해서 평균 추정 개선용으로 극점 필터링[2]을, 분산 정규화 대신 스케일 정규화[3]를 적용한다. 극점 필터링은 선형예측 켑스트럼 계수(Linear Predictive Cepstral Coefficient, LPCC)에 CMN을 적용 시, 채널 성분 추정의 정확도 향상을 위해 제안되었고, 본 논문에서는 이를 MFCC에 적용하고 CMVN을 포함한 다른 정규화 방식에도 도입한다. 스케일 정규화 방식은 화자인식의 특징 정규화를 위해 제안되었는데, 본 논문에서는 이를 잡음 환경 음성인식의 특징 정규화에 도입하여, 짧은 발화에서 발생하는 분산 추정 문제를 해결하고자 한다.

II. 극점 필터링된 켑스트럼 정규화

LPCC에서 CMN으로 채널 특성을 추정할 때, 극점 필터링은 켑스트럼 평균에 포함된 음성 성분을 감쇠시킴으로써 채널 특성 추정의 정확도를 향상시키는 것으로 알려져 있다.[2] 음성에 대한 전극(all-pole) 모델에서의 극점을 Fig. 1과 같이 극좌표로 나타냈을 때, 단위원에 근접한 협대역 극점은 현저한 포먼트(formant)를 나타내는 음성 성분의 주요 특성이다. 발

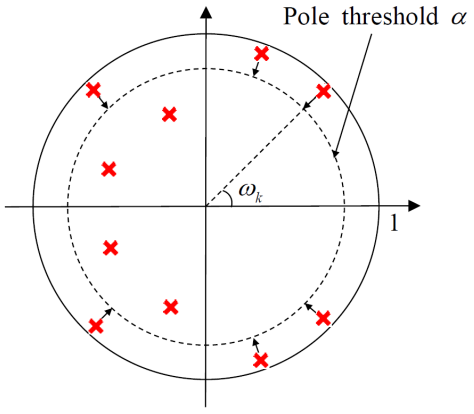

Fig. 1. Pole thresholding process on the unit circle[2].

화의 길이가 짧은 경우, 발화에 포함된 음소의 종류가 적기 때문에 켑스트럼 평균에 대해 특정 모음의 영향이 커지는 문제가 발생한다. 극점 필터링은 협대역 극점의 대역폭을 확장시킴으로써 해당 극점의 주파수 대역의 스펙트럼을 평활화하여 채널 특성 추정 시 음성 성분의 영향을 감쇠시킨다. 극점 필터링의 구체적인 구현 방식은 2가지가 있는데, 먼저 Fig. 1에서 보는 바와 같이 1보다 작은 임계값 α 보다 크기가 큰 극점들에 대해서만 주파수는 유지하면서 그 크기를 α 로 변경시켜서 해당 극점의 대역폭을 넓게 보정하는 방식이 있다.[2] 극점 필터링의 또 다른 구현 방법으로는 모든 극점의 대역폭을 일률적으로 넓히는 방법으로, 모든 극점에 대해서 주파수는 유지하면서 그 크기에 1보다 작은 γ를 곱해서 모든 극점의 대역폭을 넓게 보정한다($0<\gamma<1$).[2] p차 LPCC는 극점 z_k를 이용하여 다음과 같이 구할 수 있고,[4]

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

극점 필터링된 극점 $\hat{z}_k=\gamma z_k$를 Eq.(1)의 z_k에 대입하면 극점 필터링된 결과는

$$c_{PFLPCC}(i)=\gamma^i c_{LPCC}(i) \tag{2}$$

이 된다. 이 방법은 계산량 면에서 매우 유리하고, 더욱이 동일한 아이디어를 LPCC가 아닌 다른 켑스트럼 계수들에도 적용할 수 있다는 장점을 지닌다. 일반적으로 i번째 켑스트럼 계수에 γ^i를 곱해주는 것은 켑스트럼 리프터(lifter) 기술의 일종으로서, 고차 켑스트럼을 더 많이 감쇠시켜 스펙트럼을 평활화시키는 효과를 얻는다.

따라서 본 논문에서는 두 번째 극점 필터링 방법을 음성인식에 가장 널리 사용되는 MFCC 특징에 적용한다. 또한 CMN 이외에 CMVN에도 극점 필터링을 함께 도입하여 이들 각각을 Pole-Filtered CMN (PFCMN) 및 Pole-Filtered CMVN(PFCMVN)이라고 명명한다. M개의 프레임으로 구성된 발화의 특징벡터 열 $\boldsymbol{c}=[\boldsymbol{c}_1,\cdots,\boldsymbol{c}_m,\cdots,\boldsymbol{c}_M]$에서 m 번째 프레임의 특징벡터에 해당하는 $\boldsymbol{c}_m$의 i번째 성분의 값을 $c(m,i)$

라고 나타낼 때, PFCMN과 PFCMVN은 각각 아래 Eqs.(3)과 (4)와 같다.

$$\hat{c}_{PFCMN}(m,i) = c(m,i) - \mu_{PF}(i), \quad 1 \le m \le M, \tag{3}$$

$$\hat{c}_{PFCMVN}(m,i) = \frac{c(m,i) - \mu_{PF}(i)}{\sigma_{PF}(i)}, \quad 1 \le m \le M, \tag{4}$$

여기서 $\hat{c}_{PFCMN}(m,i)$과 $\hat{c}_{PFCMVN}(m,i)$은각각PFCMN과 PFCMVN 결과로 얻어진 m 번째 프레임 특징벡터의 i번째 성분이다. 그리고, $\mu_{PF}(i)$와 $\sigma^2_{PF}(i)$는 각각 원래 특징벡터의 i번째 성분에 극점 필터링을 수행했을 때의 평균과 분산이고, 각각

$$\mu_{PF}(i) = \frac{1}{M}\sum_{m=1}^{M} \gamma^i c(m,i) = \gamma^i \mu(i), \tag{5}$$

$$\begin{aligned}\sigma^2_{PF}(i) &= \frac{1}{M}\sum_{m=1}^{M} \{c(m,i) - \mu_{PF}(i)\}^2 \\ &= \sigma^2(i) + \{\mu(i)(\gamma^i - 1)\}^2\end{aligned} \tag{6}$$

이다. $\mu(i)$와 $\sigma^2(i)$는 각각 특징벡터의 i번째 성분에 해당하는 평균과 분산이다. $\gamma = 1$일 경우, Eqs.(3)과 (4)는 기존의 CMN 및 CMVN의 식과 동일해진다. Fig. 2는 MFCC 특징벡터의 평균에 해당하는 로그 멜-필터뱅크 출력과 이를 극점 필터링한 결과를 나타낸 것인데, γ가 작아질수록 극점에 해당하는 대역폭이 넓어지는 것을 관찰할 수 있다.

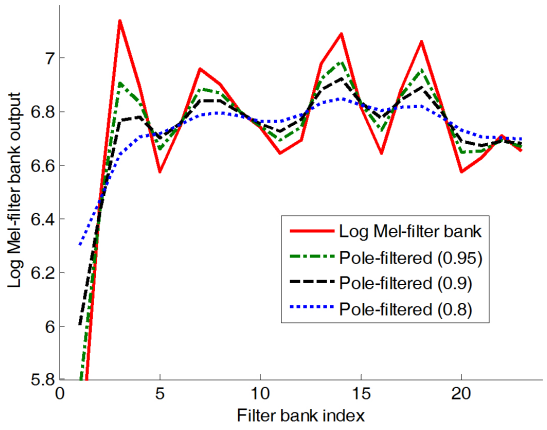

Fig. 2. Effect of pole filtering on spectra of cepstral mean.

III. 스케일 정규화

CMVN은 켑스트럼 특징의 동적 범위(dynamic range)를 정규화하기 위해 특징 차원별 분산 값을 1로 만들어주는 정규화 방법을 사용한다. 그러나 잡음 환경에 따라 켑스트럼 특징의 분포 특성이 많이 달라지기 때문에, 분산 정규화를 하더라도 피크-투-피크(peak-to-peak) 범위가 동일해지는 것은 아니다. 특히 입력 발화의 길이가 짧을 경우 분산 추정의 신뢰도가 떨어지기 때문에 더더욱 피크-투-피크 범위 관점에서의 정규화 능력은 떨어지게 된다. Alam *et al.*[3]은 분산 정규화 대신에 피크-투-피크 범위를 직접적으로 정규화하는 스케일 정규화 방식을 제안하여 화자인식 실험에 적용였으며, 켑스트럼 특징의 평균과 스케일을 함께 정규화하는 방식을 Cepstral Mean and Scale Normalization(CMSN)이라 명명하였다. CMSN의 수식은 다음과 같다.

$$\hat{c}_{CMSN}(m,i) = \frac{c(m,i) - \mu(i)}{d(i)}, \quad 1 \le m \le M, \tag{7}$$

여기서 $d(i)$는 특징벡터의 i번째 성분의 피크-투- 피크 범위를 나타내며,

$$d(i) = \max_{1 \le m \le M} c(m,i) - \min_{1 \le m \le M} c(m,i) \tag{8}$$

이다. CMSN이 잡음 환경 음성인식의 특징 정규화 용도로 사용된 예는 있으나,[5] 사실 잡음 환경 음성인식에서 분산 정규화와 스케일 정규화 중 어느 쪽이 더 효과적인지를 직접적으로 비교한 연구는 지금까지 없었던 것으로 파악된다. 따라서 본 논문에서는 잡음 환경 음성인식 실험을 통해 CMVN과 CMSN의 성능을 비교하였다. 또한 극점 필터링은 CMSN에도 적용 가능하므로, 본 논문에서는 Eq.(7)에서 $\mu(i)$ 대신 Eq.(5)의 $\mu_{PF}(i)$를 이용하는 방식을 제안하며, 이를 Pole-Filtered CMSN(PFCMSN)이라 명명하기로 한다.

IV. 성능 평가

다양한 특징 정규화 방식들의 성능을 평가하기 위해 잡음과 채널 왜곡의 영향이 반영된 Aurora 2 평가 환경을 사용하였다.[6] Aurora 2 DB는 미국인 화자가 발성한 1~7자리의 연속 숫자로 구성된 Texas Instruments(TI) digit DB에 실제 환경의 잡음을 신호대잡음비 별로 더하고, 이를 International Telecommunication Union(ITU)에서 정의한 두개의 채널을 통과시킨 데이터이다. 특징 벡터는 12차 MFCC와 로그 에너지에 대한 각각의 델타, 델타-델타 파라미터를 포함하여 총 39차 특징을 사용하였다. 음향 모델로는 단어 단위의 은닉 마르코프 모델로 16개 상태의 left-to-right 모델을 사용하였고, 각 상태당 가우스 혼합의 수는 3개이다. 음향 모델 훈련은 Aurora 2 DB에 정의된 clean-condition과 multi-condition DB를 사용하여 두 가지 모드로 훈련하였다.

Fig. 3은 clean-condition 훈련 환경에서 극점 필터링을 적용했을 때, γ에 따른 음성인식률을 나타낸 것이다. PFCMN과 PFCMVN은 각각 $\gamma = 0.8$과 $\gamma = 0.85$에서 최적의 인식률을 보이고, 기존의 CMN과 CMVN에 비해서 나은 성능을 보였다. Table 1과 Table 2는 각각 clean-condition 및 multi-condition 훈련환경에서 다양한 특징 정규화 방식들의 성능을 비교한 결과이다. PFCMN과 PFCMVN은 기존의 CMN과 CMVN에 비해서 약간의 성능 향상을 나타내고, CMSN은 기존 CMVN에 비해서 비교적 큰 폭의 성능 향상을 보였다. 특히 CMSN을 극점 필터링과 결합한 PFCMSN은 기존의 CMVN에 비해서 clean-condition 및 multi-condition 훈련환경에서 각각 24.9 %와 10.3 %의 오류 감소율을 얻었다.

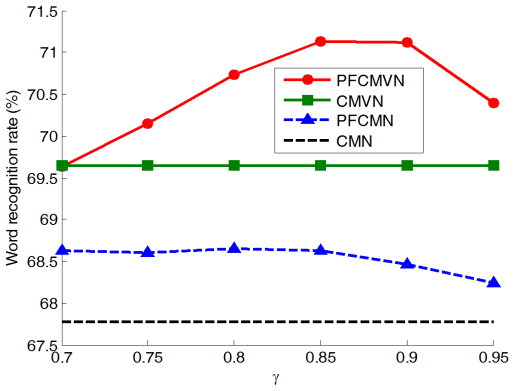

Fig. 3. Speech recognition rates of feature normalization using pole-filtering methods according to γ.

Table 1. Word recognition rates (%) in clean-condition training.

Algorithm (γ)	Set A	Set B	Set C	Average
Baseline	61.34	55.75	66.14	60.06
CMN	66.18	70.82	64.88	67.77
CMVN	70.17	70.77	66.37	69.65
CMSN	77.16	77.80	74.71	76.93
PFCMN (0.8)	67.62	71.08	65.84	68.65
PFCMVN (0.85)	71.91	72.52	66.79	71.13
PFCMSN (0.95)	77.47	78.14	74.78	77.20

Table 2. Word recognition rates (%) in multi-condition training.

Algorithm (γ)	Set A	Set B	Set C	Average
Baseline	80.31	79.40	77.00	79.28
CMN	81.02	80.44	79.75	80.53
CMVN	81.99	80.08	78.73	80.57
CMSN	83.35	81.60	80.52	82.08
PFCMN (0.8)	81.01	80.44	79.13	80.40
PFCMVN (0.85)	82.57	80.60	79.36	81.14
PFCMSN (0.95)	83.81	82.11	81.06	82.58

V. 결 론

본 논문에서는 기존의 CMN과 CMVN을 이용한 특징 정규화 방식이 짧은 발화에서 가지는 문제점을 보완하기 위해서 기존의 극점 필터링 방식과 스케일 정규화 방식을 도입하였다. Aurora 2 평가 환경에서 극점 필터링은 소폭의 성능 향상을 보였고, 스케일 정규화는 극점 필터링과 결합하여 의미 있는 성능 향상을 보임을 확인하였다. 현재 음성/비음성 구간을 구분하여 극점 필터링을 적용한 켑스트럼 정규화를 통해 잡음 환경에서 추가적인 성능을 향상시키는 연구를 진행 중이다.

감사의 글

이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음.

References

1. J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” IEEE/ACM Trans. Audio, Speech, Language Process., **22**, 745-777 (2014).
2. D. Naik, “Pole-filtered cepstral mean subtraction,” in Proc. IEEE Int. Conf. Acoust. Speech, Signal Process., 157-160 (1995).
3. M. Alam, P. Ouellet, P. Kenny, and D. O'Shaughnessy, “Comparative evaluation of feature normalization techniques for speaker verification,” Adv. Nonlinear Speech Process., 246-253 (2011).
4. M. R. Schroeder, “Direct (nonrecursive) relations between cepstrum and predictor coefficients,” IEEE Trans. Acoust., Speech, Signal Process. **29**, 297-301 (1981).
5. M. J. Alam, P. Kenny, P. Dumouchel, and D. O'Shaughnessy, “Robust feature extractors for continuous speech recognition,” in Proc. Eur. Signal Process. Conf., 944-948 (2014).
6. H. G. Hirsch and D. Pearce, “The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in Proc. Int. Conf. on Spoken Language Process., 29-32 (2000).
7. B. K. Choi, S. M. Ban, and H. S. Kim, “Pole-filtered cepstral normalization methods for robust speech recognition” (in Korean), in Proc. the 2015 Spring Conf. of the Korean Society of Speech Sciences, 101-102 (2015).

저자 약력

▸ 최 보 경 (Bo Kyeong Choi)

2013년 2월: 부산대학교 정보컴퓨터공학과 학사
2013년 3월 ~ 현재: 부산대학교 전자전기컴퓨터 공학과 석사과정

▸ 반 성 민 (Sung Min Ban)

2008년 2월: 부산대학교 전자전기공학과 학사
2010년 2월: 부산대학교 전자전기공학과 석사
2015년 2월: 부산대학교 전자전기공학과 박사
2015년 3월 ~ 현재: 부산대학교 컴퓨터 및 정보 통신연구소 전임연구원

▸ 김 형 순 (Hyung Soon Kim)

1983년 2월: 서울대학교 전자공학과 학사
1984년 2월: KAIST 전기및전자공학과 박사과정 조기진학
1989년 2월: KAIST 전기및전자공학과 박사
1987년 1월 ~ 1992년 6월: 디지콤 정보통신 연구소 선임연구원
1992년 7월 ~ 현재: 부산대학교 전자공학과 교수