한국음향학회지 제39권 제1호 pp. 64~68 (2020)
The Journal of the Acoustical Society of Korea Vol.39, No.1 (2020)
https://doi.org/10.7776/ASK.2020.39.1.064

pISSN : 1225-4428
eISSN : 2287-3775

심층신경망을 이용한 짧은 발화 음성인식에서 극점 필터링 기반의 특징 정규화 적용

Applying feature normalization based on pole filtering to short-utterance speech recognition using deep neural network

한재민,[1] 김민식,[1] 김형순[1†]

(Jaemin Han,[1] Min Sik Kim,[1] and Hyung Soon Kim[1†])

[1]부산대학교 전자공학과

(Received December 2, 2019; accepted December 26, 2019)

초 록: 가우스 혼합 모델-은닉 마코프 모델(Gaussian Mixture Model-Hidden Markov Model, GMM-HMM)을 이용하는 전통적인 음성인식 시스템에서는, 극점 필터링 기반의 켑스트럼 특징 정규화 방식이 잡음 환경에서 짧은 발화의 인식 성능을 향상시키는데 효과적이었다. 본 논문에서는 심층신경망(Deep Neural Network, DNN)을 이용하는 최신의 음성인식 시스템에서도 이 방식의 유용성이 있는지 검토한다. AURORA 2 DB에 대한 실험 결과, 특히 훈련 및 테스트 환경 사이의 불일치가 클 때에, 극점 필터링 기반의 켑스트럼 평균 분산 정규화 방식이 극점 필터링을 사용하지 않는 방식에 비해 매우 짧은 발화의 인식 성능을 개선시킴을 보여 준다.
핵심용어: 음성인식, 심층신경망, 특징 정규화, 극점 필터링

ABSTRACT: In a conventional speech recognition system using Gaussian Mixture Model-Hidden Markov Model (GMM-HMM), the cepstral feature normalization method based on pole filtering was effective in improving the performance of recognition of short utterances in noisy environments. In this paper, the usefulness of this method for the state-of-the-art speech recognition system using Deep Neural Network (DNN) is examined. Experimental results on AURORA 2 DB show that the cepstral mean and variance normalization based on pole filtering improves the recognition performance of very short utterances compared to that without pole filtering, especially when there is a large mismatch between the training and test conditions.
Keywords: Speech recognition, Deep neural network, Feature normalization, Pole filtering

PACS numbers: 43.72.Ne, 43.72.Bs

I. 서 론

심층신경망(Deep Neural Network, DNN) 또는 딥러닝 기술의 도입으로 인해 최근 음성인식은 획기적인 성능개선이 이루어져, 많은 분야에 실제적으로 사용되고 있다. 특히 심층신경망 기반의 음성인식 기술은 다양한 환경의 대용량 음성 데이터가 제공될 경우 별다른 환경 보상 기술을 적용하지 않더라도 우수한 인식성능을 나타낸다. 그러나 훈련 환경과 테스트 환경의 불일치에 따른 성능저하 문제가 해결된 것은 아니어서 이를 위한 많은 시도들이 진행되고 있다. 음성인식에서 환경 불일치 문제를 해결하기 위한 방법론은 크게 특징 영역 방식과 모델 영역 방식으로 나눌 수 있고,[1] 최근에는 이들 각각의 영역, 그리고 통합 영역에서 딥러닝 기술을 통해 문제를 해결하려는 것이 전반적인 추세이다.[2]

특징 영역에서의 전통적인 환경 불일치 극복 방안들은 기존의 Gaussian Mixture Model-Hidden Markov

†**Corresponding author:** Hyung Soon Kim (kimhs@pusan.ac.kr)
Department of Electronics Engineering, Pusan National University, 2, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan 46241, Republic of Korea
(Tel: 82-51-510-2452, Fax: 82-51-515-5190)

Model(GMM-HMM) 기반의 음성인식 시스템에서는 우수한 성능개선 효과를 거둔 반면에, 심층신경망 기반의 음성인식 시스템에 적용했을 때는 성능개선이 미미하거나 오히려 성능이 저하되는 경우도 발생하고 있는데, 켑스트럼 영역에서의 최소평균제곱오차 음질향상 방식이 그 일례이다.[3]

극점 필터링(pole filtering) 기반의 켑스트럼 특징 정규화 방식은 특징 영역 보상 방식의 일종으로서, 잡음 환경에서 GMM-HMM을 이용한 음성인식 시스템이 짧은 발화의 인식 성능 향상에 도움이 된다고 보고되었다.[4,5] 그러나 이 방식이 현재 주류를 이루는 심층신경망을 이용한 음성인식 시스템에서도 효용성이 있는지에 대해서는 검증된 바 없다. 본 논문은 극점 필터링 기반의 특징 정규화 방식이 심층신경망을 이용한 음성인식 시스템에서도 환경 불일치 여건에서 짧은 발화의 인식성능 개선에 도움을 주는지 여부를 확인하는 것을 목표로 한다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 극점 필터링 기반의 특징 정규화 방식에 대해 간단히 소개하고, 3장에서는 심층신경망을 이용한 음성인식 시스템에서 이 방식의 유용성을 평가하는 실험 내용을 다루며, 4장에서 결론을 맺는다.

II. 극점 필터링 기반의 특징 정규화

특징 정규화 방법은 음성 특징 파라미터들의 통계적 특성의 정규화를 통해 환경 불일치의 영향을 감소시키는 방법으로서, 켑스트럼 평균 정규화(Cepstral Mean Normalization, CMN), 켑스트럼 평균 분산 정규화(Cepstral Mean Variance Normalization, CMVN), 히스토그램 등화(Histogram Equalization, HE) 등 켑스트럼 정규화 방법들이 대표적인 예이다.[1] 이들에 공통적으로 적용되는 평균 정규화 과정은 시불변 채널 왜곡을 제거할 뿐 만 아니라 부가잡음에 대한 강인성도 높여서 음성인식 성능을 개선시킨다.[1] 다만 짧은 발화의 경우 정규화 과정에서의 음성 정보 손실로 인해 오히려 인식 성능을 떨어뜨리는 문제점이 있다.

극점 필터링은 원래 화자인식 분야에서 선형예측 켑스트럼 계수(Linear Predictive Cepstral Coefficient, LPCC)에 평균 정규화를 적용할 때 채널 성분 추정의 정확도 향상을 위해 제안된 방법이다.[6] 그런데 이 아이디어는 LPCC 이외에 멜-주파수 켑스트럼 계수(Mel-Frequency Cepstral Coefficient, MFCC)에도 적용할 수 있고, 잡음 환경에서 짧은 발화의 특징 보상에 효과적이라고 보고된 바 있다.[4,5]

극점 필터링 기반의 켑스트럼 특징 정규화는 CMN과 CMVN 모두에 적용 가능하지만, 본 논문에서는 이들 중 성능 면에서 더 우수한 CMVN에 적용하는 경우만 고려하기로 한다. T개의 프레임으로 구성된 발화의 특징벡터 열 $\boldsymbol{C}=[\boldsymbol{c}_1,\cdots,\boldsymbol{c}_t,\cdots,\boldsymbol{c}_T]$가 주어졌을 때 $c_t(i)$는 t 번째 프레임의 특징벡터 $\boldsymbol{c}_t$의 i번째 성분의 값을 나타낸다. 이때 극점 필터링을 적용한 켑스트럼 평균 분산 정규화(Pole-Filtered CMVN, PFCMVN) 과정은 다음 식으로 표현할 수 있다.

$$c_{t,PFCMVN}(i)=\frac{c_t(i)-\mu_{PF}(i)}{\sigma_{PF}(i)},\ 1\le t\le T, \qquad (1)$$

여기서 $\mu_{PF}(i)$와 $\sigma_{PF}(i)$는 각각 특징벡터의 i번째 성분에 극점 필터링을 수행했을 때의 평균과 표준편차이고, 각각

$$\mu_{PF}(i)=\frac{1}{T}\sum_{t=1}^{T}\gamma^i c_t(i), \qquad (2)$$

$$\sigma_{PF}(i)=\sqrt{\frac{1}{T}\sum_{t=1}^{T}\{c_t(i)-\mu_{PF}(i)\}^2} \qquad (3)$$

이다. Eq. (2)에서 $c_t(i)$에 $\gamma^i(0<\gamma<1)$를 곱해주는 것은 일종의 켑스트럼 리프터링(liftering) 기법으로, 고차 켑스트럼 성분을 더 많이 감쇄시켜 결과적으로 평균 스펙트럼을 평활화시키는 역할을 한다.[4] $\gamma=1$일 경우, Eq. (1)은 기존의 CMVN의 식과 동일해진다.

III. 실험 및 결과

본 논문에서는 극점 필터링 기반의 특징 정규화 방식이 심층신경망을 이용한 음성인식 시스템에서 환경 불일치 여건에서 짧은 발화의 인식성능 개선에

도움을 주는지 확인하기 위해, 잡음과 채널 왜곡의 영향이 반영된 AURORA 2 평가 환경을 사용하였다.[7] AURORA 2 DB는 미국인 화자가 발성한 1~7자리의 연속 숫자로 구성된 TI digit DB에 실제 환경의 잡음을 SNR별로 더하고, 이를 International Telecommunication Union(ITU)에서 정의한 두개의 채널을 통과시킨 데이터이다.

AURORA 2 평가 환경[7]에 기본 제공되는 음성인식 시스템은 GMM-HMM 방식이기 때문에, 본 논문에서는 Kaldi 음성인식 toolkit[8]의 nnet3를 기반으로 DNN-HMM 방식의 음성인식 시스템을 구성하였다. 특징 벡터로는 이전과 동일하게 12차 MFCC와 로그 에너지에 대한 각각의 델타, 델타-델타 파라미터를 포함하여 총 39차 특징을 사용하되, 전후 각각 4 프레임을 추가로 고려하여 총 351차원의 벡터를 입력 특징으로 사용하였다. 음향 모델에 사용한 DNN은 정류선형유닛 함수를 활성 함수로 사용하는 8개의 은닉층과 소프트맥스 함수를 사용하는 1개의 분류층으로 구성하였다. 은닉층은 각각 512개의 노드를 가지며, 분류층은 11개 단어(1~9, zero, oh) 각각에 대해 16개의 상태, 그리고 묵음 상태 5개를 포함하여 총 181개의 상태를 분류하도록 하였다. 음향 모델 훈련은 AURORA 2 DB에 정의된 무잡음 환경(clean-condition) DB와 다중 환경(multi-condition) DB를 사용하여 두 가지 모드로 훈련하였다.

Table 1은 무잡음 환경 훈련 및 다중 환경 훈련 각각에 대해 특징 정규화를 하지 않은 베이스라인 방식과 기존의 CMVN 방식, 그리고 네 가지 γ 값에 대한 PFCMVN 방식의 인식성능을 나타낸다. 표에서 Set A, B, C는 AURORA 2 평가 환경[7]에서 정한 3가지 테스트 셋으로서, Set A에는 열차, 군중소음(babble), 자동차, 전시장의 4가지 잡음이, 그리고 Set B에는 음식점, 거리, 공항, 기차역의 4가지 잡음이 부가되었다. Set C는 Set A 및 B에서 각각 1가지씩 선택한 총 2가지 잡음이 부가되며, Set A, B에 사용된 G.712 채널 대신 Modified Intermediate Reference System(MIRS) 채널 특성을 통과시켰다. Signal to Noise Ratio(SNR) 범위는 −5 dB에서 20 dB까지인데, 0 dB에서 20 dB 범위만 인식성능 평가에 반영된다.[7]

먼저 이 결과와 Reference [4]의 Tables 1과 2의 결과

Table 1. Word accuracy (%) according to test sets.

(a) Clean-condition training

Algorithm (γ)	Set A	Set B	Set C	Average
Baseline	54.36	50.35	57.11	53.31
CMVN	83.82	84.98	83.69	84.26
PFCMVN (0.80)	84.54	85.36	82.35	84.43
PFCMVN (0.85)	85.22	86.23	83.24	85.23
PFCMVN (0.90)	85.49	86.46	84.06	85.60
PFCMVN (0.95)	85.16	86.26	84.73	85.51

(b) Multi-condition training

Algorithm (γ)	Set A	Set B	Set C	Average
Baseline	94.41	86.23	85.82	89.42
CMVN	94.95	92.51	93.69	93.72
PFCMVN (0.80)	95.02	92.40	93.30	93.63
PFCMVN (0.85)	95.08	92.51	93.55	93.74
PFCMVN (0.90)	95.09	92.49	93.82	93.80
PFCMVN (0.95)	95.00	92.60	93.87	93.81

Table 2. Word accuracy (%) according to the length of utterances.

(a) Clean-condition training

Algorithm (γ)	Short	Medium	Long	Average
Baseline	25.70	55.87	60.82	53.31
CMVN	82.80	84.90	84.40	84.26
PFCMVN (0.90)	85.98	86.06	85.21	85.60

(b) Multi-condition training

Algorithm (γ)	Short	Medium	Long	Average
Baseline	87.37	89.62	89.97	89.42
CMVN	92.18	94.07	94.02	93.72
PFCMVN (0.90)	92.50	94.13	94.02	93.80

를 비교하면, 예상대로 DNN-HMM 시스템의 성능이 기존의 GMM-HMM 시스템의 성능보다 전반적으로 월등히 우수함을 확인 할 수 있다. 그리고 훈련 환경과 테스트 환경의 불일치가 매우 큰 무잡음 환경 훈련의 경우 기존의 CMVN보다 PFCMVN의 성능이 우수한 반면, 환경 불일치 영향이 적은 다중 환경 훈련의 경우 이들의 성능 차이가 미미하였다.

Table 2는 역시 두 가지 훈련환경에서 베이스라인, CMVN 및 PFCMVN 방식($\gamma = 0.90$)의 성능을 발화 길이에 따라 비교한 것이다. 표에서 Short, Medium, 및 Long은 각각 1~2 자리, 3~4자리 및 5~7자리 숫자열

Table 3. Error rate reduction (%) of PFCMVN (γ = 0.90) over CMVN according to the length of utterances.

Training condition	Short	Medium	Long	Average
Clean-condition	18.5	7.7	5.2	8.5
Multi-condition	4.1	1.1	0.1	1.2

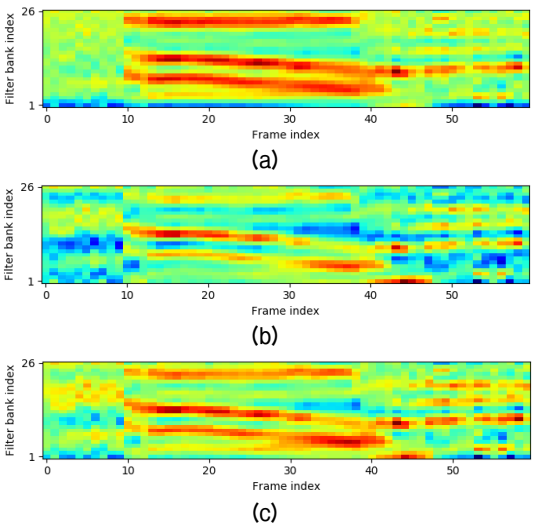

Fig. 1. (Color online) Effect of cepstral feature normalization methods on LMFE spectrogram of very short utterance (English single digit "oh"). (a) No processing (b) CMVN, (c) PFCMVN with $\gamma = 0.90$.

을 의미한다. Table 2의 결과로부터 PFCMVN 방식($\gamma = 0.90$)이 기존의 CMVN 방식에 비해 얼마나 성능이 개선되는지 오류감소율(Error Rate Reduction, ERR)로 정리한 결과를 Table 3에 나타내었다.

이 표에서 발화 길이가 짧을수록 PFCMVN 방식을 통한 인식오류 감소 효과가 커지는 것을 확인할 수 있고, 무잡음 환경 훈련일 때 매우 짧은 발화(1 ~ 2 자리 숫자열)의 오류감소율은 18.5 %로서 의미 있는 성능 개선이 이루어졌다. 또한 Table 3의 결과와 Reference [5]의 Table 1(e) 결과를 비교하면 무잡음 환경 훈련의 경우 극점 필터링으로 인한 효과가 GMM-HMM 시스템보다 DNN-HMM 시스템에서 더 크다는 것을 확인할 수 있다.

Fig. 1은 단일 숫자 발화 "oh"에 대해 MFCC 특징으로부터 역 이산 코사인 변환(Inverse Discrete Cosine Transform, IDCT)를 통해 재구성한 로그 멜-필터뱅크 에너지(Log Mel-Filter bank Energy, LMFE) 스펙트로그램을 보여준다. 그림에서 (a), (b), (c)는 각각 특징 정규화 이전의 MFFCC, CMVN을 거친 MFCC, 그리고 PFCMVN($\gamma = 0.90$)을 거친 MFCC로부터 구한 LMFE 스펙트로그램이다. CMVN은 그 특성 상 짧은 발화에서의 모음의 포먼트 정보를 상당히 유실시키는 반면, PFCMVN의 경우 훨씬 더 많은 정보를 유지함을 확인할 수 있다. 물론 발화 길이가 길어지면, CMVN의 경우에도 모음의 포먼트 정보가 대부분 그대로 유지된다.

IV. 결 론

본 논문에서는 GMM-HMM 음성인식에서 잡음 환경 짧은 발화 인식의 성능 개선에 효과적이었던 극점 필터링 기반의 켑스트럼 특징 정규화 방식이 최근 각광을 받고 있는 심층신경망을 이용한 음성인식에서도 효용성이 있는지 실험을 통해 검토하였다. AURORA 2 DB를 이용한 잡음 환경 연결숫자인식 실험 결과, 훈련 환경과 테스트 환경의 불일치가 큰 경우에는 극점 필터링을 적용한 특징 정규화 방식이 적용하지 않은 방식에 비해 유의미한 성능 개선이 있음을 확인하였다. 다만 환경 불일치가 크지 않은 경우에는 극점 필터링의 적용을 통한 성능 개선이 미미하였다. 향후 극점 필터링의 아이디어를 MFCC 특징 대신에 LMFE 특징에 직접 반영하는 방법을 개발하여 이를 통한 추가 개선 효과를 확인해 보고자 한다.

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. Z. Zhang, J. Geiger, A. Mousa, J. Pohjalainena, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: an overview of

recent developments," ACM Trans. Intell. Syst. Tech. **9**, 1-12 (2018).
3. M. L. Seltzer, D. Yu, and Y. Wang, " An investigation of deep neural networks for noise robust speech recognition," Proc. IEEE Int. Conf. Acoust. Speech, Signal Process, 7398-7402 (2013).
4. B. K. Choi, S. M. Ban, and H. S. Kim, "Cepstral feature normalization methods using pole filtering and scale normalization for robust speech recognition" (in Korean), J. Acoust. Soc. Kr. **34**, 316-320 (2015).
5. B. K. Choi, S. M. Ban, and H. S. Kim, "Selective pole filtering based feature normalization for performance improvement of short utterance recognition in noisy environments" (in Korean), Phonetics and Speech Sciences, **9**, 103-110 (2017).
6. D. Naik, "Pole-filtered cepstral mean subtraction," Proc. IEEE Int. Conf. Acoust. Speech, Signal Process, 157-160 (1995).
7. H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," Proc. ISCA ITRW ASR2000, 181-188 (2000).
8. *Kaldi Speech Recognition Toolkit*, https://kaldi-asr.org/, (Last viewed January 06, 2020).

저자 약력

▸ 한 재 민 (Jaemin Han)

2018년 8월 : 부산대학교 전자공학과 학사
2018년 9월 ~ 현재 : 부산대학교 전자전기컴퓨터공학과 석사과정

▸ 김 민 식 (Min Sik Kim)

2015년 2월 : 부산대학교 전자공학과 학사
2015년 3월 ~ 현재 : 부산대학교 전자전기컴퓨터공학과 석박사통합과정

▸ 김 형 순 (Hyung Soon Kim)

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