pISSN 2005-8063
eISSN 2586-5854
2020. 9. 30.
Vol.12 No.3
pp. 65-71

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2020.12.3.065

Compromised feature normalization method for deep neural network based speech recognition*

Min Sik Kim · Hyung Soon Kim**

Department of Electronics Engineering, Pusan National University, Busan, Korea

Abstract

Feature normalization is a method to reduce the effect of environmental mismatch between the training and test conditions through the normalization of statistical characteristics of acoustic feature parameters. It demonstrates excellent performance improvement in the traditional Gaussian mixture model-hidden Markov model (GMM-HMM)-based speech recognition system. However, in a deep neural network (DNN)-based speech recognition system, minimizing the effects of environmental mismatch does not necessarily lead to the best performance improvement. In this paper, we attribute the cause of this phenomenon to information loss due to excessive feature normalization. We investigate whether there is a feature normalization method that maximizes the speech recognition performance by properly reducing the impact of environmental mismatch, while preserving useful information for training acoustic models. To this end, we introduce the mean and exponentiated variance normalization (MEVN), which is a compromise between the mean normalization (MN) and the mean and variance normalization (MVN), and compare the performance of DNN-based speech recognition system in noisy and reverberant environments according to the degree of variance normalization. Experimental results reveal that a slight performance improvement is obtained with the MEVN over the MN and the MVN, depending on the degree of variance normalization.

Keywords: speech recognition, feature normalization, environmental mismatch, deep neural network

1. 서론

대규모 음성 데이터의 확보와 심층신경망(deep neural network, DNN)의 도입으로 최근 음성인식 성능이 크게 향상됨에 따라 많은 응용 분야에 활용되고 있다. 이러한 성능향상은 훈련 데이터가 테스트 데이터를 충분히 반영할 때에, 화자, 대역폭, 환경 차이와 같은 변화 요인에도 불구하고 상대적으로 안정적인 내부 표현을 학습하는 DNN의 장점에 기인한다(Yu et al., 2013). 그 결과로 DNN 기반의 음성인식 시스템은 시끄러운 장소나 화자의 거리가 멀 때와 같이 열악한 환경에서도 기존의 음성인식 방식들보다 훨씬 개선된 성능을 보여주고 있다. 그러나 훈련 환경과 테스트 환경의 불일치로 인한 성능 저하 문제가 충분히 해

* This work was supported by a 2-Year Research Grant of Pusan National University.
** kimhs@pusan.ac.kr, Corresponding author
Received 31 July 2020; Revised 18 September 2020; Accepted 18 September 2020

결된 것은 아니어서 이 문제를 극복하기 위한 많은 시도들이 여전히 진행되고 있다.

음성인식 분야에서 환경 불일치 문제를 해결하기 위한 방법은 크게 특징 영역 접근법과 모델 영역 접근법으로 나눌 수 있다(Li et al., 2014). 특징 정규화는 특징 영역 접근법의 한 가지 방식으로서, 음성 특징 파라미터들의 통계적인 특성의 정규화를 통해 환경 불일치의 영향을 감소시킴으로써 기존의 Gaussian mixture model-hidden Markov model(GMM-HMM) 기반의 음성인식 시스템에서 우수한 성능개선 효과를 입증한 바 있다(Li et al., 2014).

GMM 기반의 음성인식 시스템에서는 특징 정규화를 통해 환경 불일치의 영향을 감소시킬수록 성능이 더 개선되는 반면, 로그 멜 필터뱅크 에너지(log Mel-filterbank energy, LMFE)를 입력 특징으로 사용하는 DNN 기반의 음성인식 시스템에서는 환경 불일치의 영향을 최소화 하는 것이 반드시 최고의 성능 개선으로 연결되지는 않는다는 것이 본 논문의 관심사이다. 대표적으로 평균 정규화(mean normalization, MN) (Atal, 1974)와 평균 및 분산 정규화(mean and variance normalization, MVN) (Viikki et al., 1998)를 예로 들면, 음성 특징 파라미터의 차원별 평균을 정규화 하는 MN 보다 평균과 분산을 함께 정규화 하는 MVN이 환경 불일치의 영향 감소에 더 효과적이지만, DNN 기반의 음성인식 시스템의 성능은 오히려 MN을 적용하였을 때 더 우수한 경우도 많이 관찰된다. 이를 통해, 과도한 특징 정규화는 오히려 DNN 기반의 음향모델을 훈련하는 데 유용한 정보의 손실을 유발할 수 있다고 생각해 볼 수 있다. 따라서 본 논문에서는 MN과 MVN을 기준으로 하여 그 사이에 음향모델을 훈련 하는데 유용한 정보는 보존하면서 환경 불일치의 영향은 적절히 감소시켜 음성인식 성능을 최대화 하는 절충점이 있을 것이라 보고, 분산에 대한 정규화의 정도에 따라 음성인식 성능을 비교해 보았다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 특징 정규화의 적용을 통한 환경 불일치의 영향 감소와 정보의 손실 및, 분산에 대한 정규화의 정도에 따라 음성인식 성능을 비교하기 위해 새로 도입한 특징 정규화 방식에 대해 설명하고, 3장에서는 실제 DNN 기반의 음성인식 시스템에서 분산에 대한 정규화의 정도에 따른 성능을 비교하며, 4장에서 결론을 맺는다.

2. 특징 정규화 방식의 적용

서론에서 언급한 바와 같이 특징 정규화는 음성 특징 파라미터들의 통계적인 특성의 정규화를 통해 환경 불일치의 영향을 감소시키는 방법으로서, 비교적 적은 연산량으로 간단하게 구현할 수 있다는 장점이 있다. 대표적인 특징 정규화에는 음성 특징 파라미터의 각 차원별 평균을 정규화 하는 MN, 평균과 더불어 분산을 함께 정규화 하는 MVN, 그리고 모든 통계적인 적률(moment)들을 정규화 하는 히스토그램 등화(histogram equalization, HEQ) (Molau et al., 2003) 등이 있으며, MN, MVN 그리고 HEQ 순서로 음성 특징 파라미터들의 통계적인 특성을 더 많이 정규화 한다. 특징 정규화는 스펙트럼 구조의 왜곡과 같은 부작용을 야기할 수 있으나, 기존 GMM 기반의 음성인식 시스템에서는 더 많은 통계적인 특성들의 정규화를 통해 환경 불일치의 영향을 최소화 하는 것이 인식성능 개선에 도움을 준다(De La Torre et al., 2005). DNN 기반의 음성인식 시스템에서도 음향모델의 효과적인 훈련을 위해서는 입력 특징의 정규화가 요구되는데 (Ioffe & Szegedy, 2015), DNN 기반의 음성인식 시스템에서는 더 많은 통계적인 특성들의 정규화를 통해 환경 불일치의 영향을 최소화 하는 것이 반드시 성능개선을 보장하지는 않는다.

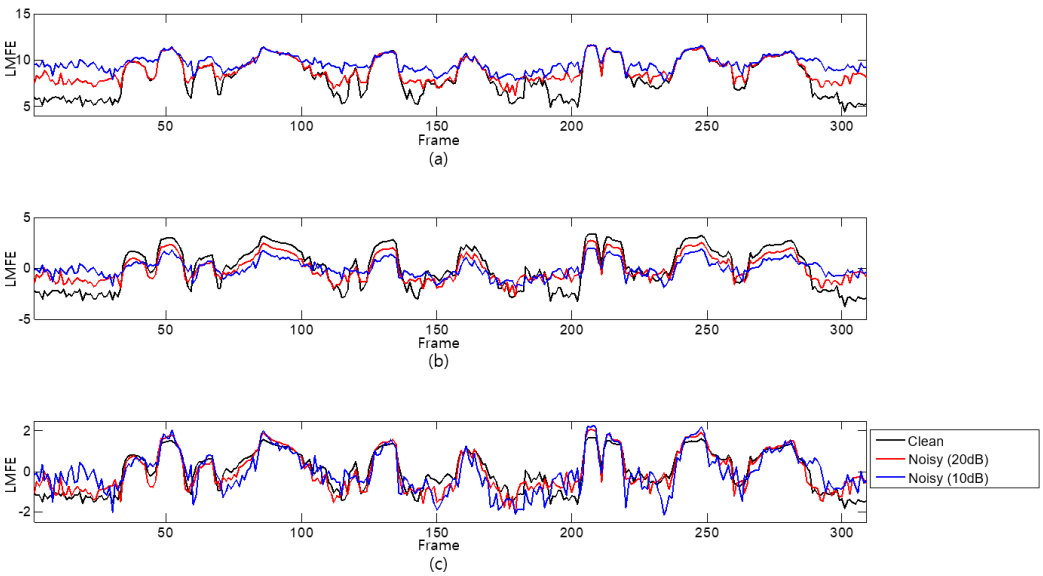

그림 1. 잡음환경에서의 특징 정규화 적용에 따른 LMFE 시간열의 예
(a) 특징 정규화 미적용, (b) MN 적용, (c) MVN 적용
Figure 1. An example of LMFE sequence according to applying feature normalization in noisy environments
(a) Not applying feature normalization, (b) applying MN, (c) applying MVN

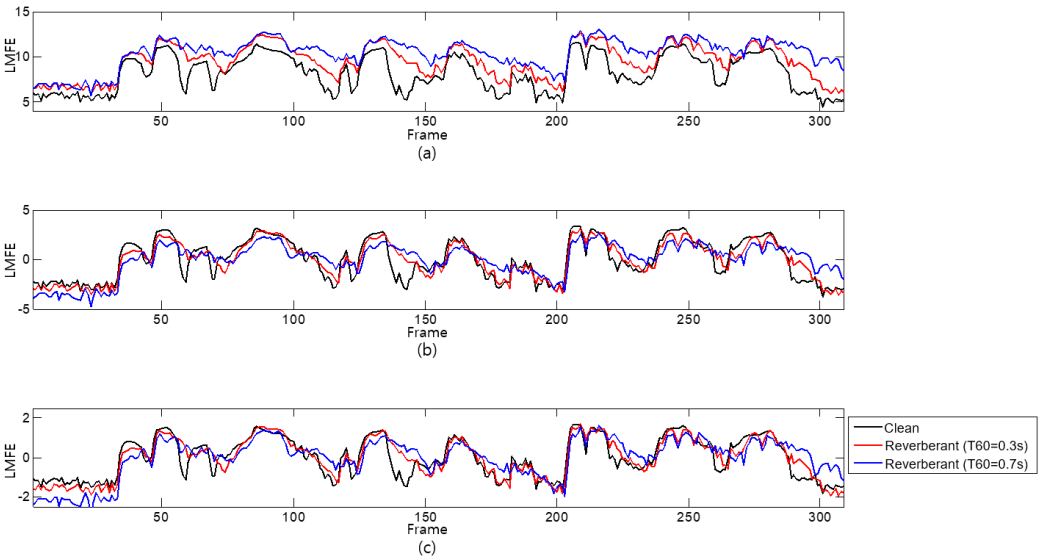

그림 2. 잔향환경에서의 특징 정규화 적용에 따른 LMFE 시간열의 예
(a) 특징 정규화 미적용, (b) MN 적용, (c) MVN 적용
Figure 2. An example of LMFE sequence according to applying feature normalization in reverberant environments
(a) Not applying feature normalization, (b) applying MN, (c) applying MVN

그림 1과 2는 잡음 환경과 잔향 환경에서의 음성으로부터 LMFE 특징 파라미터를 추출하고, 여기에 MN과 MVN을 각각 적용하여 특징 정규화 적용에 따른 LMFE 특징의 10번째 차원의 값들을 시간 순서대로 나타낸 것이다. 이들 그림으로부터 특징 정규화를 통해 환경 불일치에 의한 영향이 감소되는 것을 확인 할 수 있다. 특히 그림 1에서 보는 잡음 환경의 경우, 잡음이 심할수록 LMFE 값들의 동적 범위(dynamic range)가 줄어드는 차이가 있는데, 이러한 차이는 MN을 적용하여도 여전히 남아 있으나, MVN 적용 시에는 동적 범위와 관련된 통계적 적률인 분산의 정규화로 인해 거의 사라지게 된다. 따라서 환경 불일치에 의한 영향 감소 측면에서 MVN이 MN보다 효과적이라고 볼 수 있다. 하지만 잡음 환경에서 DNN 기반의 음성인식시스템에 MVN을 적용하면 MN을 적용할 때보다 오히려 인식성능이 저하되며, 이는 3장의 실험 결과에서 확인 할 수 있다.

본 논문에서는 이러한 성능저하의 원인이 분산의 정규화로 인해 잡음에 따른 동적 범위 차이에 관한 정보의 손실 때문이라고 간주한다. 즉, DNN은 환경 불일치에 따른 차이에 대한 정보를 반영하여 보다 강인한 음향모델로 훈련 될 수 있으나, 과도한 특징 정규화로 인해 음향모델 학습에 유용한 정보가 손실되어 오히려 상대적 성능저하를 초래할 수 있다는 것이다.

따라서 본 논문에서는 평균에 대한 정규화는 MN과 동일하게 적용하되 분산에 대한 정규화는 그 정도를 조절할 수 있는 새로운 특징 정규화 방법인 평균 및 지수적 분산 정규화(mean and exponentiated variance normalization, MEVN)를 도입하여, 음성인식 성능 측면에서 MN과 MVN의 절충을 통한 성능개선 가능성을 검토하고자 한다.

$\boldsymbol{X}=[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_t,\cdots,\boldsymbol{x}_T]$이 T개의 프레임으로 구성된 발화의 특징벡터 열이고, $x_t(i)$는 t 번째 프레임의 특징벡터인 $\boldsymbol{x}_t$의 i 번째 차원의 값을 나타낸다고 할 때, MEVN 과정은 다음 식으로 표현할 수 있다.

$$x_{t,MEVN}(i)=\frac{x_t(i)-\mu(i)}{\{\sigma(i)\}^{\alpha}},\ 1\le t\le T \tag{1}$$

여기서 $\mu(i)$와 $\sigma(i)$는 각각 특징벡터의 i번째 차원의 평균과 표준편차이고, 각각

$$\mu(i)=\frac{1}{T}\sum_{t=1}^{T}x_t(i) \tag{2}$$

$$\sigma(i)=\sqrt{\frac{1}{T}\sum_{t=1}^{T}\{x_t(i)-\mu(i)\}^2} \tag{3}$$

이다. 식 (1)에서 $\alpha(0\le\alpha\le1)$의 값에 따라 분산의 정규화 정도가 결정된다. 특히 식 (1)에서 $\alpha=0$일 때 MEVN은 MN과 동일하고, $\alpha=1$일 때는 MVN과 동일해진다.

3. 실험 및 결과

본 논문에서는 DNN 기반의 음성인식 시스템으로 Kaldi 음성인식 toolkit(Povey, 2011)의 chain 모델을 사용하였다. 시간 지연 신경망(Time delay neural network, TDNN) 구조의 음향모델에 기반한 chain 모델의 음성인식 시스템에 대해 40차원 LMFE를 입력특징으로 발화 단위의 특징 정규화 방식을 적용하여 잡음 및 잔향 환경에서의 인식성능을 평가하였다. 통상적으로 음성인식에 사용되는 음향특징은 LMFE 이외에도 여러 가지 존재하

지만 DNN 기반 음성인식 시스템에서는 LMFE 특징을 사용하였을 때 성능이 가장 우수한 것으로 보고되고 있으며(Deng et al., 2013), 본 논문에서 수행한 실험들에서도 LMFE가 가장 우수하였기 때문에 LMFE를 입력특징으로 특징 정규화 방식에 따른 인식 성능을 평가하였다.

잡음 환경에서의 음성인식 성능평가를 위해 Aurora-4 DB (Pearce & Picone, 2002)를 사용하였다. Aurora-4 DB는 5,000단어급 영어 연속음성인식 DB인 Wall Street Journal(WSJ) DB에 여러 가지 잡음 및 채널특성을 부가한 것으로서, 평가 데이터는 clean 환경인 set A, 잡음 환경인 set B, 채널 불일치 환경인 set C, 그리고 잡음 및 채널 불일치 환경인 set D로 구성된다. 다만, Aurora-4 DB의 경우 다중 조건(multi-condition) 훈련 데이터에 부가된 잡음과 평가 데이터에 부가된 잡음의 종류가 동일하기 때문에, 훈련 환경에 존재하지 않는 잡음 환경에 대한 추가적인 음성인식 성능평가를 위해 Aurora-4 DB의 clean 환경 평가 데이터인 set A 데이터에 Noisex-92 DB로부터 획득한 잡음 1종(type 1)과 DEMAND(diverse environments multichannel acoustic noise database)로부터 획득한 잡음 3종(type 2, 3, 4)을 각각 부가하여 추가적인 평가 데이터를 생성하였다.

표 1. 동일 잡음 조건에서 특징정규화 방식에 따른 단어 오류율 (%)
Table 1. Word error rate according to the feature normalization methods in matched noise condition (%)

특징 정규화 방식(α)	Set A	Set B	Set C	Set D	평균
N/A	**2.37**	5.24	5.64	14.80	9.16
MN	2.75	5.19	4.88	**13.85**	**8.71**
MEVN (0.1)	2.91	5.20	5.14	14.34	8.95
MEVN (0.2)	2.86	5.33	5.23	14.07	8.89
MEVN (0.3)	2.50	5.17	4.65	14.12	8.78
MEVN (0.4)	2.75	5.21	4.54	14.25	8.86
MEVN (0.5)	2.58	5.09	5.03	14.25	8.83
MEVN (0.6)	2.80	5.05	4.48	14.25	8.79
MEVN (0.7)	2.73	**4.98**	4.78	14.28	8.79
MEVN (0.8)	2.60	5.01	4.75	14.17	8.74
MEVN (0.9)	2.47	5.34	**4.26**	14.35	8.92
MVN	2.93	5.30	5.45	14.60	9.13

N/A, not applicable; MN, mean normalization; MEVN, mean and exponentiated variance normalization; MVN, mean and variance normalization.

표 1과 2는 Aurora-4 DB의 다중 조건 훈련 환경에서 기존 Aurora-4 DB의 평가 데이터에 해당하는 일치 잡음 조건(matched noise condition)과 추가로 새로 생성한 미관측 잡음 조건(unseen noise condition)에 대해 특징 정규화 방식에 따른 인식 성능을 나타낸다. 앞서 언급한 바와 같이 두 경우 모두 MN을 적용하였을 때가 MVN을 적용하였을 때 보다 성능 면에서 우수함을 확인할 수 있다. 그리고 표 1의 일치 잡음 조건에서는 MN과 MVN의 절충방식인 MEVN보다도 MN의 성능이 평균적으로 더 우수한 반면, 일치 잡음 조건보다 현실적으로 더 실제적인 상황인 표 2의 미관측 잡음 조건에서는 성능 차이가 크지는 않으나 $\alpha = 0.4$일 때의 MEVN 성능이 평균적으로 가장 우수하였다.

표 2. 미관측 잡음 조건에서 특징정규화 방식에 따른 단어 오류율 (%)
Table 2. Word error rate according to the feature normalization methods in unseen noise condition (%)

특징 정규화 방식(α)	Type 1	Type 2	Type 3	Type 4	평균
N/A	5.08	4.48	7.88	6.24	5.92
MN	**3.96**	4.45	7.77	5.49	5.42
MEVN (0.1)	4.52	4.56	8.05	5.62	5.69
MEVN (0.2)	4.80	4.63	7.77	5.92	5.78
MEVN (0.3)	4.25	4.39	7.83	5.38	5.46
MEVN (0.4)	3.98	**4.17**	**7.42**	5.49	**5.27**
MEVN (0.5)	4.39	4.61	7.96	5.36	5.58
MEVN (0.6)	4.63	4.78	8.63	**5.25**	5.82
MEVN (0.7)	4.76	4.30	8.67	5.59	5.83
MEVN (0.8)	4.35	4.26	8.85	5.90	5.84
MEVN (0.9)	4.56	4.71	8.26	5.60	5.78
MVN	4.82	4.46	8.07	5.45	5.70

잔향 환경에서의 음성인식 성능평가를 위해서는 Reverb 2014 DB(Kinoshita et al., 2013)를 사용하였다. Reverb 2014 DB의 평가데이터는 clean 음성 데이터에 3종류의 방에서 수집한 room impulse response(RIR)와 가산잡음을 인공적으로 부가하여 생성된 합성 데이터(SimData)와 실제 방에서 MC-WSJ-AV DB를 재생하여 녹음한 실제 데이터(RealData)로 구성된다. 표 3은 Reverb 2014 DB의 다중 조건 훈련 환경에서 특징 정규화를 적용하지 않은 기본 방식과 α값에 따른 MEVN을 적용한 방식의 인식성능을 나타낸다. 먼저 SimData와 RealData 모두 특징 정규화를 적용하지 않은 기본 방식보다 MEVN을 적용한 방식의 성능이 우수함을 확인할 수 있다. 그 차이는 SimData에 비해 RealData에서 두드러지는데 이는 RealData의 다중 조건 훈련 데이터와의 환경 불일치 정도가 SimData에 비해 더 심하기 때문으로 추정된다. 또한 RealData의 경우 표 1과 2의 잡음 환경에서와 달리 MN보다 MVN의 성능이 더 우수함을 확인할 수 있다. 추가적으로, α값에 따른 MEVN의 성능을 비교해보면 성능개선 폭이 크지는 않았지만 SimData는 $\alpha = 0.3$일 때, RealData는 $\alpha = 0.5$일 때 가장 우수한 성능을 얻을 수 있었다.

표 3. 잔향 환경에서 특징정규화 방식에 따른 단어 오류율 (%)
Table 3. Word error rate according to the feature normalization methods in reverberant environments (%)

특징 정규화 방식(α)	SimData	RealData
N/A	7.58	48.22
MN	6.34	18.90
MEVN (0.1)	6.19	18.32
MEVN (0.2)	6.26	18.47
MEVN (0.3)	6.13	17.94
MEVN (0.4)	6.25	17.76
MEVN (0.5)	6.27	17.51
MEVN (0.6)	6.23	18.14
MEVN (0.7)	6.44	17.84
MEVN (0.8)	6.36	18.19
MEVN (0.9)	6.42	17.82
MVN	6.53	18.46

Reverb 2014 DB의 경우 잡음이 부가되기는 하나, 잔향의 영향에 비해 잡음의 영향은 미미하여 잔향만 존재하는 환경으로 보아도 무방하다. 따라서 잔향과 잡음의 영향이 공존하는 환경에 대한 음성인식 성능평가를 위해 Reverb 2014 DB에 Aurora-4 DB의 잡음을 부가하여 음성인식 성능 평가에 사용하였다. 표 4는 잔향과 잡음의 영향이 공존하는 환경에서 특징 정규화를 적용하지 않은 기본 방식과 α 값에 따른 MEVN을 적용한 방식의 인식성능을 나타낸다. 잔향만 존재하는 환경의 경우와 마찬가지로 SimData와 RealData 모두 특징 정규화를 적용하지 않은 기본 방식보다 MEVN을 적용한 방식의 성능이 우수함을 확인할 수 있다. 또한, 잔향만 존재하는 환경에서처럼 RealData의 경우 MN보다 MVN의 성능이 더 우수하였다. α 값에 따른 MEVN의 성능을 비교해보면 성능개선 폭이 크지는 않았지만 SimData는 $\alpha = 0.2$일 때, RealData는 $\alpha = 0.6$일 때 가장 우수한 성능을 얻을 수 있었다.

표 4. 잔향 및 잡음 환경에서 특징 정규화 방식에 따른 단어 오류율 (%)
Table 4. Word error rate according to the feature normalization methods in reverberant and noisy environments (%)

특징 정규화 방식(α)	SimData	RealData
N/A	13.02	67.29
MN	10.10	28.82
MEVN (0.1)	10.19	29.91
MEVN (0.2)	9.95	28.62
MEVN (0.3)	10.05	28.93
MEVN (0.4)	10.19	28.16
MEVN (0.5)	10.19	28.17
MEVN (0.6)	10.16	27.67
MEVN (0.7)	10.19	28.45
MEVN (0.8)	10.36	28.09
MEVN (0.9)	10.61	28.28
MVN	10.73	28.41

이상의 실험들을 통해 LMFE 특징을 이용한 DNN 기반의 음성인식 시스템에서는 특징 정규화 방식으로서 MN과 MVN의 어느 한 쪽이 성능 면에서 일관성 있게 우수하지 않음을 확인하였고, MN과 MVN의 절충방식인 MEVN에서 α 값을 잘 선택하면 많은 경우 이들 두 방식보다 개선된 성능을 얻을 수 있음도 살펴보았다. 다만 인식 환경에 따라 최적의 α 값이 다르고, 최적의 α 값을 자동적으로 결정하는 방법을 찾기가 쉽지는 않다는 문제점이 있다. 그러나 α 값의 변화에 따른 인식 성능 차이가 크지는 않으므로, MEVN에서 고정된 α 값을 사용하더라도 MN과 MVN 중 어느 하나를 사용할지 결정하는 고민을 해결하는 현실적인 수단이 될 수는 있다고 판단된다.

표 5는 앞에서 나타낸 음성인식 실험 결과들 중 MN과 MVN 그리고 $\alpha = 0.4$인 MEVN을 적용하였을 때의 인식성능만을 나타낸 것이다. 여기서 실험 1은 일치 잡음 조건에서 전체 평가 데이터의 단어 오류율, 실험 2는 미관측 잡음 조건에서 전체 평가 데이터의 단어 오류율, 실험 3은 잔향 환경에서 RealData의 단어 오류율 그리고 실험 4는 잔향 및 잡음 환경에서 RealData의 단어 오류율을 의미한다. 각 실험에 대해 특징 정규화 방식에 따른 인식성능을 비교하여 순위를 매겨보면 MN은 실험 3과 실험 4에서 최하위, MVN은 실험 1과 실험 2에서 최하위인 것에 반해 MEVN은 최하위인 경우가 없고 실험 1을 제외하면 모두 최상위인 것을 확인할 수 있다.

표 5. 여러 인식 환경에서 특징 정규화 방식에 따른 단어 오류율 (%)
Table 5. Word error rate according to the feature normalization methods for several recognition environments (%)

특징 정규화 방식(α)	실험 1	실험 2	실험 3	실험 4
MN	8.71	5.42	18.90	28.82
MEVN (0.4)	8.86	5.27	17.76	28.16
MVN	9.13	5.70	18.46	28.41

4. 결론

본 논문에서는 LMFE 특징을 이용한 DNN 기반 음성인식 시스템에서 특징 정규화 과정을 통해 환경 불일치의 영향을 최소화 하는 것이 반드시 음성인식 성능 개선으로 연결되지 않음을 확인하였다. 환경 불일치에 의한 영향을 감소시키는 측면에서는 평균만을 정규화 하는 MN보다 평균과 분산 모두 정규화 하는 MVN이 더 효과적이지만, 실제 음성인식 실험을 통해 일치 잡음 조건 및 미관측 잡음 조건의 잡음 환경에서는 MN을 적용하였을 때에 MVN을 적용하였을 때보다 인식 성능이 더 우수하였다. 또한 다양한 환경에 대한 실험을 통해 MN과 MVN의 절충 방식인 MEVN 방식이 분산 정규화의 정도에 따라, 비록 성능 개선 폭이 크지는 않으나, 많은 경우 상대적으로 더 우수한 성능을 얻을 수 있음을 확인하였다. 결론적으로, 특징 정규화를 위해 MN 또는 MVN 중 하나의 방식을 선택하여 적용하는 부담을 가지는 것보다는, 둘 사이의 절충 방식인 MEVN을 적용하는 것이 다양한 환경에 대한 음성인식 성능을 개선하는 데에는 도움이 될 수 있다.

References

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

De La Torre, A., Peinado, A. M., Segura, J. C., Pérez-Córdoba, J. L., Benítez, M. C., & Rubio, A. J. (2005). Histogram equalization of speech representation for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, *13*(3), 355-366.

Deng, L., Li, J., Huang, J. T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., ... Gong, Y. (2013, May). Recent advances in deep learning for speech research at Microsoft. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8604-8608). Vancouver, BC.

Ioffe, S., & Szegedy, C. (2015, July). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of 32nd International Conference on Machine*

Learning (Vol. 37, pp. 448-456). Lille, France.

Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Habets, E., Haeb-Umbach, R., Leutnant, V., ... & Gannot, S. (2013, October). The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 1-4). New Paltz, NY.

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.

Molau, S., Hilger, F., & Ney, H. (2003, April). Feature space normalization in adverse acoustic conditions. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing 2003 Proceedings (ICASSP'03)* (Vol. 1, pp. I-I). Hong Kong.

Pearce, D., & Picone, J. (2002). *Aurora working group: DSR front end LVCSR evaluation AU/384/02* (Technical report). Mississippi State, MS; Institute for Signal and Information Processing at Mississippi State University.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., ... Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding (No. CONF)*. Hawaii, HI.

Viikki, O., Bye, D., & Laurila, K. (1998, May). A recursive feature vector normalization approach for robust speech recognition in noise. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98* (Vol. 2, pp. 733-736). Seattle, WA.

Yu, D., Seltzer, M. L., Li, J., Huang, J. T., & Seide, F. (2013, March). Feature learning in deep neural networks − studies on speech recognition tasks. *Proceedings of International Conference on Learning Representations(ICLR).* Scottsdale, AZ.

• **김민식 (Min Sik Kim)**
부산대학교 전기전자컴퓨터공학과 석박사통합과정
부산시 금정구 부산대학로 63번길 2 특공관 10707호
Tel: 051-510-1704
Email: fire9945@pusan.ac.kr
관심분야: 음성인식, 음성신호처리

• **김형순 (Hyung Soon Kim)** 교신저자
부산대학교 전자공학과 교수
부산시 금정구 부산대학로 63번길 2 기전관 409호
Tel: 051-510-2452
Email: kimhs@pusan.ac.kr
관심분야: 음성인식 및 합성, 음성신호처리

심층신경망 기반의 음성인식을 위한 절충된 특징 정규화 방식*

김 민 식 · 김 형 순

부산대학교 전자공학과

국문초록

특징 정규화는 음성 특징 파라미터들의 통계적인 특성의 정규화를 통해 훈련 및 테스트 조건 사이의 환경 불일치의 영향을 감소시키는 방법으로서 기존의 Gaussian mixture model-hidden Markov model(GMM-HMM) 기반의 음성인식 시스템에서 우수한 성능개선을 입증한 바 있다. 하지만 심층신경망(deep neural network, DNN) 기반의 음성인식 시스템에서는 환경 불일치의 영향을 최소화 하는 것이 반드시 최고의 성능 개선으로 연결되지는 않는다. 본 논문에서는 이러한 현상의 원인을 과도한 특징 정규화로 인한 정보손실 때문이라 보고, 음향모델을 훈련 하는데 유용한 정보는 보존하면서 환경 불일치의 영향은 적절히 감소시켜 음성인식 성능을 최대화 하는 특징 정규화 방식이 있는지 검토해보고자 한다. 이를 위해 평균 정규화(mean normalization, MN)와 평균 및 분산 정규화(mean and variance normalization, MVN)의 절충 방식인 평균 및 지수적 분산 정규화(mean and exponentiated variance normalization, MEVN)를 도입하여, 잡음 및 잔향 환경에서 분산에 대한 정규화의 정도에 따른 DNN 기반의 음성인식 시스템의 성능을 비교한다. 실험 결과, 성능 개선의 폭이 크지는 않으나 분산 정규화의 정도에 따라 MEVN이 MN과 MVN보다 성능이 우수함을 보여준다.

핵심어: 음성인식, 특징 정규화, 환경 불일치, 심층신경망

* 이 과제는 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음.