한국음향학회지 제38권 제1호 pp. 51~55 (2019)
The Journal of the Acoustical Society of Korea Vol.38, No.1 (2019)
https://doi.org/10.7776/ASK.2019.38.1.051

pISSN : 1225-4428
eISSN : 2287-3775

잡음 환경에 효과적인 음성인식을 위한 특징 보상 이득 기반의 음성 향상 기법

Speech enhancement method based on feature compensation gain for effective speech recognition in noisy environments

배아라,[1] 김우일[1†]

(Ara Bae[1] and Wooil Kim[1†])

[1]인천대학교 컴퓨터공학부

(Received November 9, 2018; accepted January 25, 2019)

초 록: 본 논문에서는 잡음 환경에 강인한 음성 인식 성능을 위해 특징 보상 이득을 이용한 음성 향상 기법을 제안한다. 본 논문에서는 변분 모델 생성 기법을 채용한 병렬 결합된 가우스 혼합 모델(Parallel Combined Gaussian Mixture Model, PCGMM) 기반의 특징 보상 기법으로부터 계산할 수 있는 특징 보상 이득을 이용하는 음성 향상 기술을 제안한다. 불일치 환경 음성 인식 시스템 적용 환경에서 본 논문에서 제안하는 기법이 실험 결과에서 기존의 전처리 기법 및 이전 연구에서 제안된 특징 보상 기반의 음성 향상 기법에 비해 다양한 잡음 및 SNR(Signal to Noise Ratio) 조건에서 월등한 인식 성능을 나타내는 것을 확인한다. 또한 잡음 모델 선택 기법을 적용함으로써 음성 인식 성능을 유사한 수준으로 유지하면서 계산량을 대폭적으로 감축할 수 있다.
핵심용어: 음성 향상, 특징 보상 이득, 변분 모델 생성, 음성 인식, 잡음 환경

ABSTRACT: This paper proposes a speech enhancement method utilizing the feature compensation gain for robust speech recognition performances in noisy environments. In this paper we propose a speech enhancement method utilizing the feature compensation gain which is obtained from the PCGMM (Parallel Combined Gaussian Mixture Model)-based feature compensation method employing variational model composition. The experimental results show that the proposed method significantly outperforms the conventional front-end algorithms and our previous research over various background noise types and SNR (Signal to Noise Ratio) conditions in mismatched ASR (Automatic Speech Recognition) system condition. The computation complexity is significantly reduced by employing the noise model selection technique with maintaining the speech recognition performance at a similar level.
Keywords: Speech enhancement, Feature compensation gain, Variational model composition, Speech recognition, Noisy environment

PACS numbers: 43.72.Bs, 43.72.Ne

I. 서 론

실제 시스템을 적용하는 환경과 인식 시스템에 장착되는 음향 모델을 훈련하는 환경이 음향학적 측면에서 불일치한다는 점은 인식 성능이 저하되는 가장 큰 원인 중 하나이다. 음성 인식 성능 향상을 위해 이러한 음향학적 불일치를 줄이는 다양한 연구가 진행되어왔다.[1-6] 이러한 연구는 두 가지 접근 방법으로 나눌 수 있다. 첫째는 음성 인식 시스템의 전처리 단계에서 잡음을 제거하고 음성을 향상하거나 잡음에 강인한 음성 특징을 추출하는 방법이다. 두 번째 접근 방법은 이미 훈련된 음향 모델을 새로운 잡음 환경과 일치하도록 적응해주는 기법이다. 최근에는 심층

†**Corresponding author:** Wooil Kim (wikim@inu.ac.kr)
Department of Computer Science and Engineering, Incheon National University, 119 Academy-ro, Yeonsu-gu, Incheon 22012, Republic of Korea
(Tel: 82-32-835-8459, Fax: 82-32-835-0780)

신경망을 활용한 기법들이 소개되었다.[7,8]

본 논문에서는 특징 보상 기법의 결과로 얻어지는 특징 보상 이득을 이용한 음성 향상 기술을 제안한다. 특징 보상 기술로는 변분 모델 생성(Variational Model Composition, VMC) 기법을 채용한 병렬 결합된 가우스 혼합 모델(Parallel Combined Gaussian Mixture Model, PCGMM) 기반의 특징 보상 기법을 사용한다.[5] 성능 평가를 위해 Aurora 2.0 평가 프레임워크와 데이터베이스를 사용하였다.[9]

II. VMC 기반 특징 보상 기법

VMC 기법은 입력된 음성의 지속 기간 동안 변화하는 시변 잡음을 효과적으로 모델링하기 위해 제안된 기법으로 오염된 입력 음성으로부터 잡음 모델을 예측하고 이를 기저모델로 사용하여 다중의 유사 잡음 모델을 생성하는 방식이다.[5] 예측된 기저 잡음 모델의 분산 요소 중 크기가 큰 것을 변분 요소로 결정하고 교란 인자 f_p를 변분 요소의 평균 파라미터에 다음 식과 같이 적용함으로써 다중의 모델을 생성한다.

$$\tilde{\mu}_i = \begin{cases} \mu_i(1+f_p), \text{if } i \in \{v_1, v_2, ..., v_V\} \\ \mu_i, \ otherwise \end{cases}. \quad (1)$$

위 식에서 $f_p = 0, -\alpha, \text{or } +\alpha$이며, α는 실험에 의해서 결정되는 값이다. 최종적으로 3^V개의 모델을 가지는 변분 모델 집합 $\{\tilde{\boldsymbol{\lambda}} = (\tilde{\boldsymbol{\mu}}, \boldsymbol{\sigma}^2)\}$이 생성된다. 각 모델의 분산벡터는 기저 모델과 동일한 분산 벡터를 가지는 것으로 가정한다.

VMC 기법으로 생성되는 복수개의 잡음 모델과 깨끗한 음성 모델을 병렬 결합함으로써 다중의 오염 음성 모델을 생성할 수 있다. 다중 모델 채용 기법은 입력되는 오염 음성에 대해 각 잡음 환경 모델에 대한 사후 확률을 계산하여 이를 최소 평균 제곱 오차 기반의 예측 과정에 적용함으로써 구현된다. Fig. 1은 이전 연구에서 제안된 VMC 기법을 채용한 병렬 결합된 PCGMM 기반의 특징 보상 기법을 나타낸다.[3]

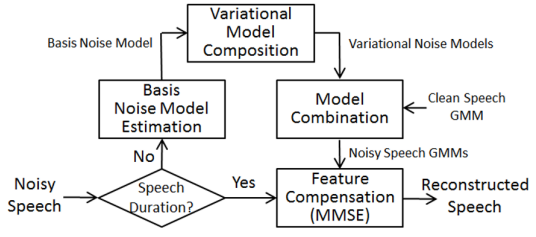

Fig. 1. Block diagram of the VMC-PCGMM-based feature compensation scheme.[5]

III. VMC-PCGMM 특징 보상 이득 기반의 음성 향상 기법

제안하는 음성 향상 기법에서는 II장에서 설명한 VMC-PCGMM 기반의 특징 보상 기법의 결과로 얻어지는 음성 특징 보상의 이득(gain)을 이용한다. VMC-PCGMM 기반의 특징 보상 기법은 켑스트럼 도메인에서 이루어지므로, 입력 음성 파형 도메인으로의 변환이 이루어져야 한다. 특징 보상 이득은 다음과 같이 DCT(Discrete Cosine Transform) 역변환을 통해 로그 스펙트럼 도메인에서 얻어진다.[10]

$$\boldsymbol{g} = \boldsymbol{C}^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{x}}). \quad (2)$$

Eq. (2)에서 $\boldsymbol{y}$와 $\tilde{\boldsymbol{x}}$는 각각 오염된 입력 음성 특징 벡터, VMC-PCGMM 특징 보상 기법으로 얻을 수 있는 깨끗한 음성 특징 벡터이고, $\boldsymbol{C}$는 DCT 변환 행렬을 나타낸다. 이와 같이 얻어진 보상 이득 벡터 $\boldsymbol{g}$는 입력된 오염 음성 파형의 스펙트럼에 다음과 같이 적용되어 깨끗한 음성을 복구한다.

$$\tilde{X}(k,t) = \frac{Y(k,t)}{e^{g_i}}, \ \text{if} \ k \in i \ \text{in Mel-filterbank}. \quad (3)$$

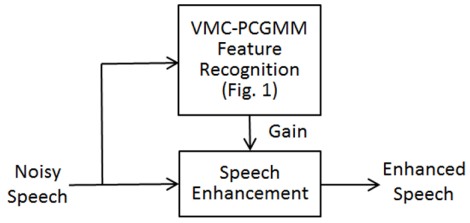

Fig. 2. Block diagram of the proposed speech enhancement scheme employing the VMC-PCGMM-based feature compensation method.

위 식에서 $Y(k,t)$와 $\widetilde{X}(k,t)$는 각각 입력 오염 음성스펙트럼과 향상된 깨끗한 음성 스펙트럼의 시간 t에서의 k번째 주파수 요소를 나타낸다. Fig. 2는 본 논문에서 제안하는 VMC-PCGMM 기법의 특징 보상 이득을 기반으로 하는 음성 향상 기법의 블록 다이어그램을 나타낸다.

IV. 잡음 모델 선택을 이용한 계산량 감소

VMC 기법에서 V개의 변분 요소를 채택할 경우 총 3^V개의 잡음 모델이 생성되고, 이를 깨끗한 음성 GMM 모델과 결합할 경우 상당한 계산량을 필요로 한다. 이에 따라 본 논문에서는 VMC 기법을 통해 생성된 잡음 모델 중에 음성 향상 성능에 기여할 수 있는 잡음 모델만을 선택함으로써 대폭적으로 계산량을 감축하고자 한다. 제안하는 모델 선택 기법에서는 Eq. (4)과 같이 기저 잡음 모델과 생성된 잡음 모델과의 유클리드(Euclidean) 거리를 측정하여 특징 보상에 사용될 잡음 모델을 선택한다.

$$d_k = \sqrt{(\boldsymbol{\mu}-\widetilde{\boldsymbol{\mu}}_k)^T(\boldsymbol{\mu}-\widetilde{\boldsymbol{\mu}}_k)}\,. \tag{4}$$

위 식에서 $\boldsymbol{\mu}$는 기저 잡음 모델의 평균 벡터를 나타내고, $\widetilde{\boldsymbol{\mu}}_k$는 VMC 과정을 통해 생성된 k번째 잡음 모델의 평균 벡터를 나타낸다. Eq. (4)에 의해 선택된 거리가 가장 큰 K개의 잡음 모델과 기저 잡음 모델을 최종적으로 모델 결합에 사용하여 VMC-PCGMM 특징 보상 과정에 적용한다.

V. 실험 및 결과

Aurora 2.0에서 제공하는 평가 방식을 사용하여 객관적인 성능 평가를 진행하였다.[9] 본 논문에서는 시간에 따라 변하는 잡음 환경을 반영하기 위해 Aurora 2.0의 SetA에 포함되어 있는 지하철, 자동차, 웅성거림(speech babble) 외에 배경 음악을 잡음 환경으로 사용하였다. 배경 음악은 빠르기와 비트가 다양한 유명 한국 가요 10곡의 전주 부분에서 샘플링 하였다.

대표적인 전처리 알고리즘인 주파수 차감법(Spectral Subtraction, SS), 켑스트럼 정규화(Cepstral Mean Normalization, CMN) 기법, VTS(Vector Taylor Series) 기반 알고리즘을 이용하여 성능 비교를 수행하였다.[4] 또한 ETSI에서 개발한 AFE(Advanced Front-End) 알고리즘도 평가하였다.[11] 본 논문에서는 단어 오인식율(Word Error Rate, WER)을 음성 인식 성능의 지표로 사용하였다.

5.1 일치 환경 음성 인식 시스템 조건에서의 성능 평가

Table 1은 음향학적으로 일치하는 자동 음성 인식(Automatic Speech Recognition, ASR) 시스템에 대한 성능 평가 결과이다. 본 연구에서는 음성 인식 시스템과 동일한 특징 추출 기법을 적용할 수 있는 경우를 일치 환경 음성 인식 시스템이라 가정하였다. 일치 환경시스템 조건에서는 ASR 시스템의 음향 모델(즉, HMM) 훈련에 사용된 것과 동일한 음성 데이터베이스를 사용할 수 있다고 가정하여, VTS, PCGMM, VMC-PCGMM과 같이 음향 모델을 사용하는 특징 보상 기법에서 음향 모델 훈련에 동일한 음성 데이터를 사용하는 것이 가능한 것을 가정하였다.

Table 1의 결과는 본 논문에 사용한 Aurora 2.0 데이터베이스의 4개의 잡음 환경에 대해 모든 SNR (Signal to Noise Ratio) 조건 (0 dB, 5 dB, 10 dB, 15 dB, 20 dB)을 평균한 성능이다. 이전 연구에서 제안한 PCGMM 특징 보상 이득 기반의 음성 향상 기법(SE-PCGMM)은 평균 13.51 %의 오인식률을 나타냈고,[10] 본 논문에서 제안한 VMC-PCGMM 특징 보상 이득을 이용한 음성

Table 1. Speech recognition performance with the matched ASR system condition (WER, %).

	Subway	Car	Babble	Music	Avg.
No processing	31.23	35.65	46.98	34.22	37.02
SS+CMN	13.82	15.41	20.18	24.62	18.51
VTS	14.15	14.93	21.10	25.73	18.98
ETSI-AFE	7.70	7.18	16.36	21.14	13.10
PCGMM	8.58	7.97	18.23	15.17	12.49
VMC-PCGMM	7.34	7.26	13.44	14.12	10.54
SE-PCGMM	9.24	8.87	19.97	15.98	13.51
SE-VMC	**9.27**	**7.00**	**14.66**	**14.56**	**10.87**
SE-VMC-Top5	**7.45**	**7.11**	**14.29**	**15.05**	**10.98**

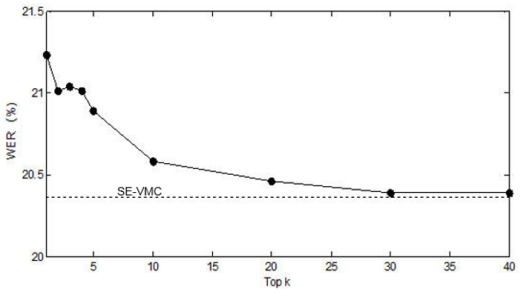

Fig. 3. Recognition performance for music noise in 5 dB SNR as change of the number of selected noise models (WER, %).

향상 기법(SE-VMC)은 10.87 %로서 기존의 SE-PCGMM에 비해 상당한 성능 향상을 보였다. 이와 같은 결과는 다중의 모델을 생성하는 VMC 기법이 본 논문에서 제안하는 음성 향상 기법에서도 다양한 잡음 환경에서 효과적으로 적용될 수 있음을 입증하는 결과이다.

Table 1에서 SE-VMC-Top5는 본 논문에서 제안하는 잡음 모델 선택 기법을 채용하여 기저 잡음 모델과 가장 거리가 먼 5개의 잡음 모델을 선택했을 때의 결과를 나타낸다. 실험 결과 10.98 %의 오인식률을 나타냈으며, 이는 모든 잡음 모델 개수인 81개의 모델을 사용한 결과와 거의 유사한 성능 결과이다. 제안한 선택적 모델 채용 기법인 SE-VMC-Top5에서는 기저 모델을 포함하여 총 6개의 잡음 모델을 사용한다. 이에 따라 계산량은 모델 선택 기법을 채용하지 않는 기법과 비교하여 약 92.59 % (= 75/81*100) 감소하는 것을 알 수 있다. 모델 선택을 위한 유클리드 거리를 구하는 과정에서 추가적인 계산량을 요구하지만, 잡음 모델을 이용하여 모든 입력 특징 벡터에 대해 확률값을 계산하는 과정이 전체적인 계산량을 좌우하므로 모델 선택 과정에 의한 계산량 추가는 상대적으로 작다고 할 수 있다. Fig. 3은 배경 음악 잡음 5 dB SNR 환경에서 모델 선택 개수에 따른 WER 성능 변화를 나타낸다.

5.2 불일치 환경 음성 인식 시스템 조건에서의 성능 평가

Table 2는 불일치 환경 ASR 시스템 조건에서의 성능평가 결과이다. 본 연구에서 불일치 환경 ASR 시스템 조건은 ASR 시스템에 대한 정보가 알려져 있지 않은 상태로 전처리 기법을 개발하는 조건을 가정하였다. 이러한 조건을 실험에 적용하기 위해 음성 인식 시스템은 일치 환경 ASR 실험과 동일한 시스템을 사용하였다. 음성 인식 시스템에 대한 정보가 알려져 있지 않은 상황을 가정하므로, 전처리 기법에서는 음성 인식 시스템과 세부 처리 과정이 상이한 HTK[12]로 구현한 특징 추출 기법을 사용하였다. VTS와 PCGMM, VMC-PCGMM 전처리 기법과 제안한 기법에 필요한 음향 모델은 음성인식기 훈련에 사용된 것과 상이한 TIMIT 데이터베이스를 사용하여 훈련함으로써 불일치되는 상황을 의도적으로 구현하였다. 따라서 Table 2에 평가된 각 전처리 기법은 음성 특징 추출 기법과 사용된 음향 모델이 음성 인식기와 불일치되는 성질을 갖는다.

Table 2. Speech recognition performance with the mismatched ASR system condition (WER, %).

	Subway	Car	Babble	Music	Avg.
No processing	28.12	31.50	40.33	40.58	35.13
SS+CMN	20.09	22.34	29.27	41.25	28.24
VTS	19.06	18.47	31.67	45.31	28.63
PCGMM	14.27	15.18	26.97	31.11	21.88
VMC-PCGMM	13.53	13.56	23.09	30.47	20.16
SE-PCGMM	11.05	11.70	27.49	16.96	16.80
SE-VMC	**7.06**	**6.38**	**16.48**	**14.65**	**11.14**
SE-VMC-Top5	**7.22**	**6.48**	**16.23**	**14.81**	**11.18**

Table 2의 결과에서 알 수 있듯이 기존의 전처리 기법과 이전 연구에서 제안한 음성 향상 기법(SE-PCGMM)이 일치 환경 ASR 조건과 비교하여 대폭적인 성능 하락을 보이는 것에 비하여, 본 논문에서 제안하는 VMC-PCGMM 기반의 음성 향상 기법(SE-VMC)은 성능 하락이 상대적으로 매우 작은 것을 확인할 수 있다. 불일치 환경 ASR 조건에서 가장 우수한 인식 성능인 11.14 %의 오인식률을 나타내고, 해당 성능은 일치 환경 ASR 조건과 비교하여 0.27 %의 매우 낮은 성능 하락을 보인다. 모델 선택 기법을 채용한 경우에는 모든 모델을 사용한 경우와 매우 유사한 성능을 보이며 (11.14 % vs. 11.18 %), 일치 환경 ASR 조건과 비교하여 0.04 %의 성능 하락을 나타낸다. 이와 같은 결과는 본 논문에서 제안하는 VMC-PCGMM

기반의 특징 보상 이득을 이용한 음성 향상 기법이 사양이 알려져 있지 않은 음성 인식시스템의 전처리 기법으로서 효과적으로 사용될 수 있음을 입증하는 것이다.

VI. 결 론

본 논문에서는 잡음 환경에 강인한 음성 인식 성능을 위해 특징 보상 이득을 이용한 음성 향상 기법을 제안하였다. 본 논문에서는 변분 모델 생성 기법을 채용한 PCGMM 기반의 특징 보상 기법으로부터 계산할 수 있는 특징 보상 이득을 이용하는 음성 향상 기술을 제안하였다. 불일치 환경 음성 인식 시스템 적용 환경에서 본 논문에서 제안하는 기법이 실험 결과에서 기존의 전처리 기법 및 이전 연구에서 제안된 특징 보상 기반의 음성 향상 기법에 비해 다양한 잡음 및 SNR 조건에서 월등한 인식 성능을 나타내는 것을 확인하였다. 또한 잡음 모델 선택 기법을 적용함으로써 음성 인식 성능을 유사한 수준으로 유지하면서 계산량을 대폭적으로 감축할 수 있었다.

감사의 글

본 논문은 인천대학교 2014년 자체연구비 지원에 의하여 연구되었음.

References

1. S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” Proc. IEEE Trans. on Acoustics, Speech and Signal, **27**, 113-120 (1979).
2. P. J. Moreno, B. Raj, and R. M. Stern, “Data-driven environmental compensation for speech recognition: a unified approach,” Speech Communication, **24**, 267-285 (1998).
3. W. Kim and J. H. L. Hansen, “Variational noise model composition through model perturbation for robust speech recognition with time-varying background noise,” Speech Communication, **53**, 451-464 (2011).
4. J. L. Gauvain and C. H. Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains,” Proc. IEEE Trans. on Speech and Audio, **2**, 291-298 (1994).
5. C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density HMMs,” Computer Speech and Language, **9**, 171-185 (1995).
6. M. J. F. Gales and S. J. Young, “Robust continuous speech recognition using parallel model combination,” Proc. IEEE Trans. on Speech and Audio, **4**, 352-359 (1996).
7. J. Du, L.-R. Dai, and Q. Huo, “Synthesized stereo mapping via deep neural networks for noisy speech recognition,” ICASSP 2014, 1764-1768 (2014).
8. K. Han, Y. He, D. Bangchi, E. F. -Lussifer, and D. L. Wang, “Deep neural network based spectral feature mapping for robust speech recognition,” Interspeech 2015, 2484-2488 (2015).
9. H. G. Hirsch and D. Pearce, “The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions,” ISCA ITRW ASR2000 (2000).
10. W. Kim, “Speech enhancement based on feature compensation for independently applying to different types of speech recognition systems” (in Korean), J. Korea Institute of Information and Communication Engineering, **18**, 2367-2374 (2014).
11. ETSI ES 201 108, *ETSI Standard Document*, v1.1.2 (2000-04), 2000.
12. http://htk.eng.cam.ac.uk

저자 약력

▸ 배 아 라 (Ara Bae)

2015년 3월 ~ 현재: 인천대학교 컴퓨터공학부 학사과정

▸ 김 우 일 (Wooil Kim)

1996년 2월: 고려대학교 전자공학과 학사
1998년 8월: 고려대학교 전자공학과 석사
2003년 8월: 고려대학교 전자공학과 박사
2012년 8월 ~ 2016년 8월: 인천대학교 컴퓨터공학부 조교수
2016년 9월 ~ 현재: 인천대학교 컴퓨터 공학부 부교수