베이지안 비음수 행렬 인수분해 기반의 음성 강화 기법에서 최적의 latent source 개수에 대한 연구

이혜인, 서지훈, *이영한, *김제우, **이석필
상명대학교 컴퓨터과학과, *전자부품연구원, **상명대학교 미디어소프트웨어학과
hilee0907@gmail.com, JHseo890@gmail.com, *{yhlee, jwkim}@keti.re.kr,
**esprit@smu.ac.kr

Study on optimal number of latent source in speech enhancement based Bayesian nonnegative matrix factorization

Hye In Lee, Ji Hun Seo, *Young Han Lee, *Je Woo Kim, **Seok Pil Lee
Sangmyung University Computer Science, *Korea Electronics Technology Institute,
**Sangmyung University Media Software

요 약

본 논문은 베이지안 비음수 행렬 인수분해 (Bayesian nonnegative matrix factorization, BNMF) 기반의 음성 강화 기법에서 음성과 잡음 성분의 latent source 수에 따른 강화성능에 대해 서술한다. BNMF 기반의 음성 강화 기법은 입력 신호를 서브 신호들의 합으로 분해한 후, 잡음 성분을 제거하는 방식으로 그 성능이 기존의 NMF 기반의 방법들보다 우수한 것으로 알려져 있다. 그러나 많은 계산량과 latent source 의 수에 따라 성능의 차이가 있다는 단점이 있다. 이러한 단점을 개선하기 위해 본 논문에서는 BNMF 기반의 음성 강화 기법에서 최적의 latent source 개수를 찾기 위한 실험을 진행하였다. 실험은 잡음의 종류, 음성의 종류, 음성과 잡음의 latent source 의 개수, 그리고 SNR 을 바꿔가며 진행하였고, 성능 평가 방법으로 PESQ (perceptual evaluation of speech quality) 를 이용하였다. 실험 결과, 음성의 latent source 개수는 성능에 영향을 주지 않지만, 잡음의 latent source 개수는 많을수록 성능이 좋은 것으로 확인되었다.

1. 서론

모바일 통신이 가지는 이동성으로 인해 여러 환경의 다양한 잡음들이 음성신호에 섞이며 통화음질을 악화시킨다. 이로 인해 음성 강화의 필요성이 인식되며 수년간 많은 연구들이 이루어지고 있다 [1-5]. 기존의 음성 강화 기법에서는 원하지 않는 잡음을 제거함으로써 음성신호를 향상시키는 방식을 이용하였다. 앞의 방식과는 조금 다르게 오디오 소스 분리에 이용되는 비음수 행렬 인수분해 (NMF)를 음성강화에 적용한 기법들도 연구되었다 [6]. 그 중 음성신호와 잡음신호의 시간 종속특성을 베이즈 추론의 사전 분포를 만드는데 이용하여 NMF 에 적용하는 방식인 BNMF 는 기존의 NMF 보다 성능이 더 우수하다는 연구 결과가 있다 [7]. 또한 BNMF 의 파라메터 중 latent source 의 개수는 그 수에 따라 연산 시간과 음성이 향상되는 정도가 다르다.

따라서 본 논문에서는 BNMF 를 이용하여 음성강화를 할 때, 그 성능과 연산량이 최적이 되도록 하는 latent source 의 개수를 결정하기 위해 음성과 잡음의 latent source 개수를 바꿔가며 실험하였다. 실험은 다섯 종류의 잡음과 여성, 남성의 음성 신호를 SNR 별로 합성하여 사용했으며, 성능 평가지표로는 PESQ 를 이용하였다.

2. NMF 기반의 음성 강화 알고리즘

2.1 NMF 를 이용한 음성 강화

기존의 NMF 는 식 (1)을 만족하는 행렬 W 와 H 를 구하는데 이용되는 기법으로, 모든 행렬의 원소는 비음수로 이루어져있다[8]. 음성강화에서 NMF 는 식 (1)과 같이 잡음과 음성이 섞인 신호의 스펙트럼을 V_{fn} 으로 나타내고, 이 V 를 주파수의 특성을 나타내는 basis 행렬 W_{fk} 와 시간에 따른 기저 행렬의 이득 값을 가지는 activation 행렬 H_{kn} 의 곱으로 나타낼 수 있도록 행렬 W 와 H 를 구하는데 이용된다.

$$V_{fn} \approx W_{fk} H_{kn} \quad (1)$$

본 논문에서는 KL divergence 를 거리함수로 이용하며 식 (2)의 목적함수의 값이 최소가 되도록 최적화를 수행하여 음성과 잡음 각각의 행렬 W 와 H 를 구한다.

$$D(\mathrm{V}\,\mathrm{P}\mathrm{WH}) = \sum_{ij} \left(V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}\right) \quad (2)$$

$$(W^*, H^*) = \arg\min_{WH} D(V\,\mathrm{P}WH) \quad (3)$$

2.2 Bayesian NMF 를 이용한 음성 강화

BNMF 는 기존의 NMF 에 오디오 신호의 시간 종속성을 고려하여 확률적인 관점에서 도출된 개념이다 [9]. 따라서 식 (1)을 확률적 관점에서 다시 쓰면 다음과 같다.

$$V(f,n) = \sum_{k} Z(f,k,n), \qquad (4)$$
$$Z(f,k,n) : Po(Z(f,k,n); \Lambda(f,k,n))$$

[6] 에서 제안된 방법은 스펙트럼 행렬 V 가 확률 행렬이며 latent source 인 Z 의 합으로 이루어진다고 가정한다. 여기서 Z 는 포아송 분포를 가진다. 따라서 행렬 V 도 포아송 분포이며 정수값을 가진다는 결과를 도출해낼 수 있다. 본 논문에서는 [5]와 같은 방식으로 W 와 H 를 구하고, 이를 음성 강화에 이용하였다.

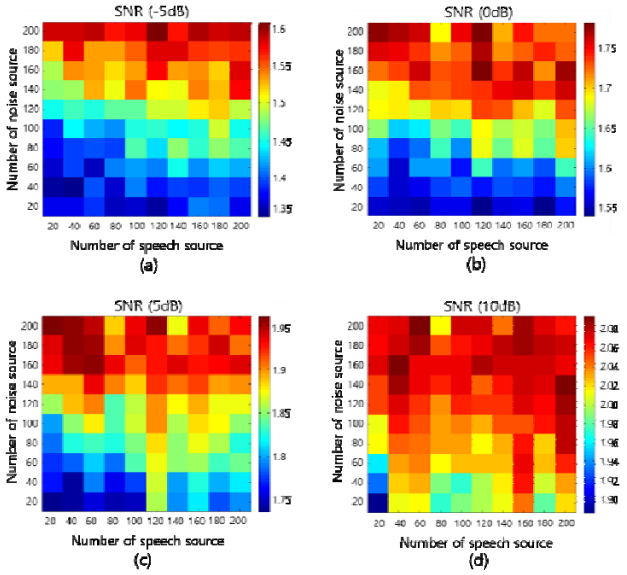
3. 실험방법

실험결과가 화자의 성별이나 잡음의 종류에 종속적이지 않도록 실험데이터로는 다섯 개의 환경잡음과 남자, 여자 화자의 음성신호가 이용되었다. 우선 잡음으로는 QUT-NOISE-TIMIT 의 데이터들을 사용하였다 [10]. 이 잡음 데이터는 사람들이 웅성거리는 babble 환경, TV 나 라디오 등을 통해 음악이 나오는 환경, 차도나 사람이 다니는 거리, 운전하는 차의 내부환경, 그리고 반향이 심한 수영장이나 주차장 환경의 잡음으로 이루어져있다. 음성 신호로는 남자와 여자가 같은 문장을 녹음한 것을 이용하였다. 실험에 쓰이는 합성신호는 음성과 잡음의 SNR 을 −5 dB, 0 dB, 5 dB, 10 dB 로 합성하여 SNR 에 독립적이도록 실험을 진행하였다. 모든 실험데이터는 샘플링 주파수가 16kHz 이며 모노 채널인 데이터를 이용하였다. Latent source 개수의 설정은 연산 시간과 음성 향상 성능을 고려하여 음성과 잡음신호 각각 20 개부터 200 개까지 20 의 간격으로 증가시키며 실험을 진행하였다.

4. 실험결과

Latent source 의 수에 따른 성능을 평가하기 위해 평가지표로는 사람의 청각 인식 특성을 고려한 음성통화 품질을 객관적으로 평가해주는 PESQ 를 사용하였다 [11]. 성능의 추이를 보기 위해 가로축을 음성의 latent source 개수, 세로축을 잡음의 latent source 개수로 정하고, 이를 도식화하였다. 다음의 <그림 1>은 SNR 별 PESQ 를 나타낸 그림이다.

실험결과를 보면 음성의 latent source 개수가 늘어남에 따른 성능향상은 보이지 않지만, 잡음의 latent source 개수가 늘어남에 따라 성능이 향상된다는 것을 볼 수 있다. 또한 10dB 을 제외하고 음성의 latent source 의 개수가 120 개, 잡음의 latent source 의 개수가 200 개일 때 성능이 좋은 것으로 나타난다.



<그림 1> SNR 별 PESQ 측정 결과
(a) −5 dB, (b) 0 dB, (c) 5 dB, (d) 10 dB

5. 결론 및 향후 계획

본 논문에서는 BNMF 기반의 음성 향상 기법에서 음성과 잡음 각각 최적의 latent source 개수를 결정하기 위한 실험을 진행하였다. 실험데이터로는 화자와 잡음의 종류에 강인하도록 다섯 종류의 환경잡음과 남성과 여성의 음성신호를 이용하였다. 음성 향상 성능의 평가지표로는 PESQ 를 이용하였고, latent source 수에 따른 성능의 추이를 보기 위해 실험결과를 도식화하였다. 실험 결과 음성 latent source 개수의 증가는 성능에 영향을 미치지 않지만, 잡음 latent source 개수의 증가는 성능을 향상시킨다는 것을 확인하였다.

향후 BNMF 기반의 음성강화 기법을 이용할 때, 본 논문의 결과를 토대로 잡음의 latent source 수는 늘리고 상대적으로 음성의 latent source 를 줄여 불필요한 연산은 줄이면서 성능은 향상시킬 수 있다.

Acknowledgement

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [R0101-15-0061, 2D 스테레오 콘텐츠를 3D 입체 음향 콘텐츠로 변환하기 위한 음원 객체 분리/위치 추정 및 3D 렌더링 소프트웨어 기술 개발]

참 고 문 헌

[1] Ephraim, Y., and Malah, D., "Speech enhancement using a minimum-mean square error short time spectral amplitude estimator," IEEE Trans., Acoustics, Speech and Signal Processing, vol. 32, Issue. 6, pp. 1109-1121, Dec. 1984.

[2] V. Grancharov and J. S. B. Kleijn, "On causal algorithms for speech enhancement," IEEE Trans. Audio, Speech,

Lang. Process., vol. 4, no.3, pp.764-773, May 2006.

[3] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," IEEE Trans., Audio, Speech, Lang. Process., vol. 14, no. 1, pp. 163-176, Jan. 2006.

[4] H. Sameti, H. Sheikhzadeh, L. Deng, and R. Brennana, "HMM-based strategies for enhancement of speech signals embedded in nonstationaryh noise," IEEE Trans. Speech Audio Process., vol. 6, no. 5, pp.445-455, Sep. 1998.

[5] Nasser Mohammadiha, Paris smaragdis, and Aren Leijon, "Supervised and unsupervised speech enhancement using NMF," IEEE Trans. Audio, Speech, and Lang. Process., vol. 21, no.10, Oct. 2013.

[6] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorization models," Computat. Intell. Neuroscience, vol. 2009, no. Article ID 785152, p. 17, 2009.

[7] Nasser Mohamadiha, Jalil Taghia, and Arne Leijon, "Single channel speech enhancement using bayesian NMF with recursive temporal updates of prior distributions," in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), pp. 4561-4564, March, 2012.

[8] D.D. Lee, and H. S. Seung, "Algorithms for non-negative matrix factorization," in Proc. Neural Inf. Process. Syst. Conf. (NIPS), pp. 556-562, 2000.

[9] Tuomas Virtanen, A. Taylan Cemgil, and Simon Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in Proc. IEEE int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), pp. 1825 - 1828, April, 2008.

[10] Dean, David B., Sridharan, Sridha, Vogt, Robert J., Mason, and Michael W., "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," 11th Annual conference of the International Speech Communication Association, Sept. 2010.

[11] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Tech. Rep., ITU-T P.862, 2001.