Neurocomputing 284 (2018) 107–118

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Training and compensation of class-conditioned NMF bases for speech enhancement[☆]

Hanwook Chung[a,*], Roland Badeau[b], Eric Plourde[c], Benoit Champagne[a]

[a] *Department of Electrical and Computer Engineering, McGill University, Montreal, Quebec, Canada*
[b] *LTCI, Télécom ParisTech, Université Paris-Saclay, Paris 75013, France*
[c] *Department of Electrical and Computer Engineering, Sherbrooke University, Sherbrooke, Quebec, Canada*

ARTICLE INFO

Article history:
Received 4 July 2017
Revised 28 October 2017
Accepted 2 January 2018

Communicated by Prof. R. Capobianco Guido

Keywords:
Single-channel speech enhancement
Non-negative matrix factorization
Probabilistic generative model
Classification
Variational Bayesian expectation-maximization

ABSTRACT

In this paper, we introduce a training and compensation algorithm of the class-conditioned basis vectors in the non-negative matrix factorization (NMF) model for single-channel speech enhancement. The main goal is to estimate the basis vectors of different signal sources in a way that prevents them from representing each other, in order to reduce the residual noise components that have features similar to the speech signal. During the proposed training stage, the basis matrices for the clean speech and noises are estimated jointly by constraining them to belong to different classes. To this end, we employ the probabilistic generative model (PGM) of classification, specified by class-conditional densities, as an a priori distribution for the basis vectors. The update rules of the NMF and the PGM parameters of classification are jointly obtained by using the variational Bayesian expectation-maximization (VBEM) algorithm, which guarantees convergence to a stationary point. Another goal of the proposed algorithm is to handle a mismatch between the characteristics of the training and test data. This is accomplished during the proposed enhancement stage, where we implement a basis compensation scheme. Specifically, we use extra free basis vectors to capture the features that are not included in the training data. Objective experimental results for different combination of speaker and noise types show that the proposed algorithm can provide better speech enhancement performance than the benchmark algorithms under various conditions.

1. Introduction

The general objective of speech enhancement algorithms is to remove the background noise from a noisy speech signal to improve its quality or intelligibility. They have been an attractive research area for decades and find various applications including mobile telephony, hearing aid and speech recognition. Numerous single-channel speech enhancement algorithms have been proposed in the past, such as: spectral subtraction [1], minimum mean-square error (MMSE) estimation [2,3] and subspace decomposition [4]. However, these algorithms tend to provide limited performance in adverse noisy environments, e.g., low input signal-to-noise ratio (SNR) or non-stationary noise conditions, since they use a minimal amount of *a priori* information about the speech and noise.

Recently, the non-negative matrix factorization (NMF) approach has been successfully applied to various problems, such as music transcription [5], source separation [6], speech enhancement [7] and image representation [8]. In general, NMF is a dimensionality reduction technique, which decomposes a given matrix into basis and activation matrices with non-negative elements [9,10]. In audio and speech applications, the magnitude or power spectrum of the (noisy) audio signal is interpreted as a linear combination of the NMF basis vectors, which play a key role in the enhancement process. Deep neural network (DNN) algorithms have also gained enormous interest lately. The DNN training aims at estimating the nonlinear mapping function, specified by the weights and biases of the hidden layers, that relates the input features to the output target features. Applications of DNN to speech enhancement and source separation have been introduced in [11–13]. The NMF and DNN algorithms differ significantly in terms of underlying modeling structure and training requirements; in this paper, we focus on a linear NMF model.

In a supervised NMF-based framework, the basis vectors are typically obtained *a priori* for each source by independently using isolated training data during the training stage. However, there

[☆] Funding for this work was provided by Microsemi Corporation (Ottawa, Canada) and a grant from NSERC (Govt. of Canada)

[*] Corresponding author.
E-mail addresses: hanwook.chung@mail.mcgill.ca (H. Chung), roland.badeau@telecom-paristech.fr (R. Badeau), eric.plourde@usherbrooke.ca (E. Plourde), benoit.champagne@mcgill.ca (B. Champagne).

https://doi.org/10.1016/j.neucom.2018.01.013
0925-2312/© 2018 Elsevier B.V. All rights reserved.

are two main problems in such a framework. The first one is that the basis vectors of the different signal sources, e.g., speech and noise, may share similar characteristics. For example, the basis vectors of the speech spectrum can represent the noise spectrum and hence, the enhanced speech may contain residual noise components which have features similar to the speech signal. One possible remedy is to train the basis vectors of each source in a way that prevents them from representing other sources. In [14], the cross-coherence of the basis vectors is added as a penalty term to the NMF cost function, whereas the cross-reconstruction error terms are considered in [15]. The authors in [16–18] propose to use additional training data which are generated by mixing, e.g., adding or concatenating, the isolated training data of each source. However, the approaches in [16,17] are based on heuristic multiplicative update (MU) rules which do not guarantee the convergence of the NMF in general [10,19]. Moreover, the basis vectors in [17,18] are obtained indirectly by means of the activation matrix estimated from the mixed training data and hence, lack an explicit interpretation in terms of discrimination.

The second problem in a supervised framework is the existence of a mismatch between the characteristics of the training and test data. A common approach to overcome this problem is to add explicit regularization terms to the NMF cost function that incorporate some prior knowledge, such as the temporal continuity [20] or statistical characteristics of the magnitude spectra [21]. In these algorithms, however, the basis vectors are fixed during the enhancement stage, which limits the performance when there is a large mismatch between the training and test data. One alternative approach is to use a basis adaptation scheme during the enhancement stage. In [22], the basis vectors are adapted based on prior distributions modeled by Gamma mixtures. The authors in [23] employ extra validation data for speaker adaptation in a speech-music separation task. In [24], the basis vectors are adapted by using a combination of the original and pre-processed noisy speech samples, the latter being obtained via a classical MMSE-based speech enhancement algorithm. In these algorithms, however, the basis vectors are adapted from the mixtures of multiple sources, e.g., noise and speech, such that the resulting basis vectors may still exhibit features of different sources. Consequently, the enhanced speech may contain some residual noise components and hence, adapting the complete set of basis vectors may limit the enhancement performance.

In this paper, to overcome these limitations, we introduce a training and compensation algorithm of the class-conditioned basis vectors in the NMF model for single-channel speech enhancement, which is an extension of our previous works on training class-conditioned basis vectors in [25], and basis compensation in [26]. In the proposed framework herein, we consider the probabilistic generative model (PGM) of classification specified by class-conditional densities [27], along with the NMF model [28]. Specifically, the PGM of classification is used as an explicit *a priori* distribution for the basis vectors. During the proposed training stage, the basis matrices for all the clean speech and noise sources are estimated jointly by constraining them to belong to one of several speech and noise classes. Previously in [25], we used a traditional Gaussian-distributed class-conditional density [27], and the model parameters were obtained through a maximum *a posteriori* (MAP) estimator using the expectation-maximization (EM) algorithm. In this paper, we make two key modifications. First, we employ a Gamma-distributed class-conditional density to bring more coherence into the NMF model. Second, the update rules of the NMF model and the PGM parameters for classification are jointly obtained via the variational Bayesian expectation-maximization (VBEM) algorithm, which can be considered as an extension of the EM algorithm [27–29].

The proposed enhancement stage consists of two steps. First, we perform noise classification based on the posterior class probability (PCP), in order to determine which type of noise is included in the noisy speech. Second, we implement a basis compensation algorithm by adopting the approach in [26]. That is, we use extra free basis vectors for both the clean speech and noise to capture the features which cannot be explained by the limited set of basis vectors due to the hard decision on the noise type as well as features which are not included in the training data. The PGM parameters for classification are employed while inferring the free basis vectors as well as during the noise classification. Previously in [26], the free basis vectors were estimated by using the MU rules, whereas we use the VBEM algorithm in this paper. Experimental results of perceptual evaluation of speech quality (PESQ) [30], source-to-distortion ratio (SDR) [31] and segmental SNR (SSNR) show that the proposed algorithm provides better enhancement performance than the benchmark algorithms under various conditions.

The paper is organized as follows. In Section 2, we review the basic principles of supervised NMF-based single-channel speech enhancement. In Section 3, we introduce the PGMs of the NMF and classification models. The proposed training stage is derived in Section 4, and the proposed enhancement stage is explained in Section 5. Experimental results are presented in Sections 6 and 7 concludes the paper.

2. NMF-based speech enhancement framework

For a given matrix $\mathbf{V} = [v_{kl}] \in \mathbb{R}_+^{K\times L}$, NMF finds a local optimal decomposition of $\mathbf{V} \approx \mathbf{WH}$, where $\mathbf{W} = [w_{km}] \in \mathbb{R}_+^{K\times M}$ is a basis matrix, $\mathbf{H} = [h_{ml}] \in \mathbb{R}_+^{M\times L}$ is an activation matrix, $\mathbb{R}_+$ denotes the set of non-negative real numbers and M is the number of basis vectors, typically chosen such that $M < \min(K, L)$ [19]. The factorization is obtained by minimizing a suitable cost function, such as the Kullback-Leibler (KL) divergence. In this case, the solutions can be obtained iteratively using the following MU rules [9]

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{(\mathbf{V}/(\mathbf{WH}))\mathbf{H}^T}{\mathbf{1}_{KL}\mathbf{H}^T}, \mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T(\mathbf{V}/(\mathbf{WH}))}{\mathbf{W}^T\mathbf{1}_{KL}} \tag{1}$$

where the operation $\otimes$ denotes element-wise multiplication, / and the quotient line are element-wise division, $\mathbf{1}_{KL}$ is a $K \times L$ matrix with all entries equal to one, the superscript T is the matrix transpose, and $\leftarrow$ refers to an iterative overwrite.

In NMF-based single-channel speech enhancement, one commonly assumes that the magnitude spectrum of the noisy speech, obtained via short-time Fourier transform (STFT), can be approximated by the sum of the clean speech and noise magnitude spectra [6,7,32], i.e., $|Y_{kl}| \approx |S_{kl}| + |N_{kl}|$ where Y_{kl}, S_{kl} and $N_{kl,}$ respectively denote the STFT coefficients of the noisy speech, clean speech and noise at the frequency bin $k \in \{1, \ldots, K\}$ and time frame $l \in \{1, \ldots, L\}$. Hence, in this work, $\mathbf{V} = [v_{kl}]$ may contain the magnitude spectral values of the noisy speech, clean speech or noise, as indicated by subscripts or superscripts Y, S and N, respectively.

In a supervised framework, $\mathbf{W}_S$ and $\mathbf{W}_N$ are first obtained during the training stage, by applying (1) to the training data $\mathbf{V}_S$ and $\mathbf{V}_N$. In the enhancement stage, for an online application, the activation vector $\mathbf{h}_l^Y = [(\mathbf{h}_l^S)^T(\mathbf{h}_l^N)^T]^T \in \mathbb{R}_+^{(M_S+M_N)\times 1}$ is estimated for the l-th time frame by applying the activation update in (1) to $|\mathbf{y}_l| = [|Y_{kl}|] \in \mathbb{R}_+^{K\times 1}$, while fixing $\mathbf{W}_Y = [\mathbf{W}_S\mathbf{W}_N]$. In this work, we instead consider a *mini-batch* online application by concatenating several successive time frames of the noisy speech. That is, we construct a target matrix as $\mathbf{V}_{l_b}^Y = |\mathbf{Y}_{l_b}| \in \mathbb{R}_+^{K\times L_b}$, where $l_b = 1, 2, \ldots$ is the mini-batch index, $\mathbf{Y}_{l_b}$ is the noisy speech matrix consisting of the time frames $l \in \{(l_b - 1)L_b + 1, \ldots, l_bL_b\}$, L_b is the mini-batch size, and $|\cdot|$ denotes the element-wise magnitude computation. The merit of using a mini-batch approach is that we can not only