creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학박사학위논문

NMF-based Compositional Models for Audio Source Separation

음향 신호 분리를 위한 비음수 행렬 인수분해 기반 조합 모델

2017년 2월

서울대학교 대학원

전기 · 컴퓨터공학부

권 기 수

NMF-based Compositional Models for Audio Source Separation

음원 분리를 위한
비음수 행렬 인수분해 기반 조합 모델

지도교수 김 남 수

이 논문을 공학박사 학위논문으로 제출함
2017년 1월

서울대학교 대학원
전기·컴퓨터공학부
권 기 수

권기수의 공학박사 학위논문을 인준함
2016년 12월

위 원 장 김 성 철 (인)

부위원장 김 남 수 (인)

위 원 조 남 익 (인)

위 원 장 준 혁 (인)

위 원 신 종 원 (인)

Abstract

Many classes of data can be represented by constructive combinations of parts. Most signal and data from nature have nonnegative values and can be explained and reconstructed by constructive models. By the constructive models, only the additive combination is allowed and it does not result in subtraction of parts. The compositional models include dictionary learning, exemplar-based approaches, and nonnegative matrix factorization (NMF). Compositional models are desirable in many areas including image or visual signal processing, text information processing, audio signal processing, and music information retrieval. In this dissertation, we choose NMF for compositional models and NMF-based target source separation is performed for the application.

The target source separation is the extraction or reconstruction of the target signals in the mixture signals which consists with the target and interfering signals. The target source separation can be thought as blind source separation (BSS). BSS aims that the original unknown source signals are extracted without knowing or with very limited information. However, in these days, much of prior information is frequently utilized, and various approaches have been proposed for single channel source separation.

NMF basically approximates a nonnegative data matrix $\mathbf{V}$ with a product of

nonnegative basis and encoding matrices $\mathbf{W}$ and $\mathbf{H}$, i.e., $\mathbf{V} \approx \mathbf{WH}$. Since both $\mathbf{W}$ and $\mathbf{H}$ are nonnegative, NMF often leads to a part based representation of the data. The methods based on NMF have shown impressive results in single channel source separation The objective function of NMF is generally presented Euclidean distant, Kullback-Leibler divergence, and Itakura-saito divergence. Many optimization methods have been proposed and utilized, e.g., multiplicative update rule, projected gradient descent and NeNMF. However, NMF-based audio source separation has some issues as follows: non-uniqueness of the bases, a high dependence to the prior information, the overlapped subspace between target bases and interfering bases, a disregard of the encoding vectors from the training phase, and insufficient analysis of sparse NMF. In this dissertation, we propose new approaches to resolve the above issues.

In section 4, we propose a novel speech enhancement method that combines the statistical model-based enhancement scheme with the NMF-based gain function. For a better performance in time-varying noise environments, both the speech and noise bases of NMF are adapted simultaneously with the help of the estimated speech presence probability. In section 5, we propose a discriminative NMF (DNMF) algorithm which exploits the reconstruction error for the interfering signals as well as the target signal based on target bases. In section 6, we propose an approach to robust bases estimation in which an incremental strategy is adopted. Based on an analogy between clustering and NMF analysis, we incrementally estimate the NMF bases similar to the modified k-means and Linde-Buzo-Gray algorithms popular in the data clustering area. In Section 7, the distribution of the encoding vector is modeled as a multivariate exponential PDF (MVE) with a single scaling factor for each source. In Section 8, several sparse penalty terms for NMF are analyzed and compared in terms of signal to distortion ratio, sparseness of encoding vectors, reconstruction error, and entropy of basis vectors. The new objective function which applied sparse representation and discriminative NMF (DNMF) is also proposed.

Keywords: audio source separation, nonnegative matrix factorization (NMF), on-line bases update, discriminative NMF, incremental approach, modified k-means clustering, encoding vectors, exponential distribution, sparse NMF, anti-sparsity

Student number: 2011-20788

Contents

viii

List of Figures

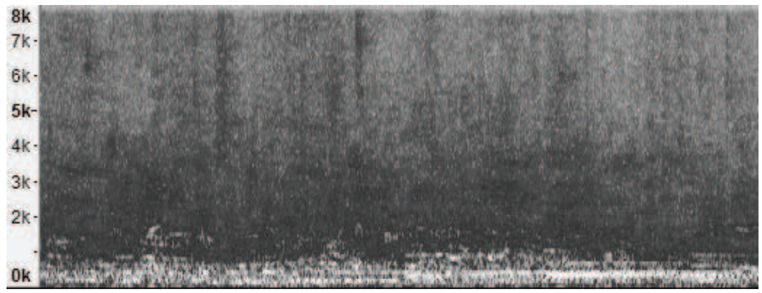
List of Tables

Chapter 1

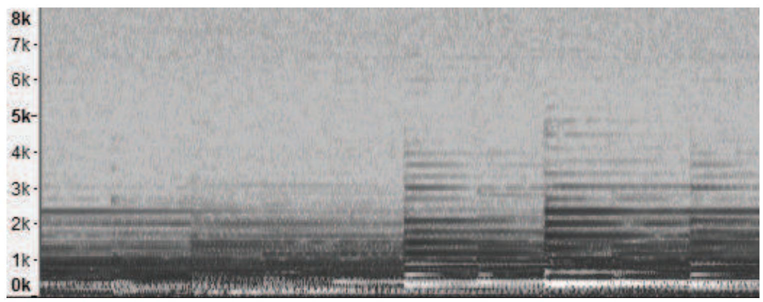
Introduction

1.1 Audio source separation

Over several years, audio source separation has been one of the interesting topics in the audio signal processing such as speech enhancement, speech recognition, music signal processing, and so on [1]- [10]. Audio source separation is the recovering one or several source signals from a given mixture or observed signal, for example, when piano and babble source signals are mixed, if piano is the target source, audio source separation reconstructs the piano signal and erases the babble signal from the mixture signals. Fig. 1.1 shows this example. Data-representation methods and template-based approaches have been widely applied to audio source separation, which make the representation models or statistics from a priori information possibly available from a training database (DB). Generally, audio source separation is the same or similar to blind source separation (BSS). The aim of BSS is to process these observations (acquired by sensors or sensor array) in such a way that the original unknown source signals are extracted by, e.g., an adaptive system, or separated



(a) Magnitude spectra of mixture signals (piano and babble sources)



Magnitude spectra of piano signals from the mixture (a)

Figure 1.1: Magnitude spectra of the mixture and reconstructed piano signals from audio source separation.

simultaneously using, e.g., a block (or batch)-based algorithm, without knowing or with very limited information about the characteristics of the transmission channels through which the sources propagate to the sensor [11]. Independent component analysis (ICA) is one of the most widely used and cited techniques for BSS and audio source separation by revealing the hidden factors that underlie the sets of measurement s of the observed signal. In these days, the prior information of the audio sources have been easily utilized for audio source separation as the enhancing of the computing power. As this flow and aspect, a number of new techniques have been proposed in audio source separation, such as latent variable analysis, dictionary learning, independent vector analysis, factor analysis, matrix completion, sparse component analysis, nonnegative matrix factorization (NMF), and complex-valued adaptive methods [11]. Audio source separation can be divided into two classes, multi-channel and single-channel audio source separation. The single-channel audio source separation utilizes the only one channel signal, and NMF and deep neural network algorithms have been widely applied for that in recent years [12]- [23].

1.2 Speech enhancement

In the audio source separation, speech source is the most important and frequently applied as the target source signals. If the target source of the audio source separation is speech, it is the same to the speech enhancement. Speech enhancement means that denosing from the observed audio signals. The aim of speech enhancement is the enhancing the speech quality for human or the speech recognition of machine.

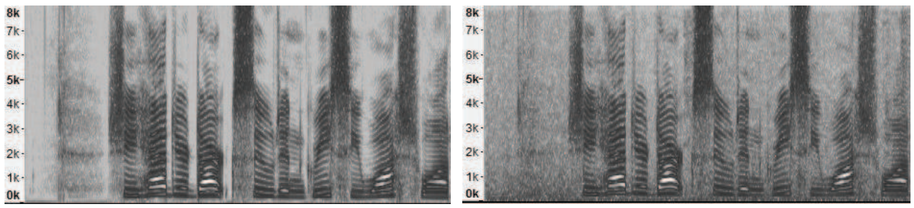
Two major classes of single channel speech enhancement techniques may be the statistical model-based and template-based approaches [24]- [36]. In the methods falling in the former category, speech and noise are assumed to have separate parametric distributions for which the parameters are estimated from the input signal [24]- [27]. In most of the cases, these approaches perform voice activity detection (VAD) implicitly or explicitly and compute the gains based on the assumed statistical models and estimated parameters. One of the significant advantages of the statistical model-based techniques is that the models do not need to be trained *a priori*. Since, however, the statistical models are constructed based on a stationarity assumption, the performance deteriorates when the background noise is highly non-stationary.

On the other hand, the template-based techniques utilize specific types of the *a priori* information of speech or noise [15]- [33]. *A priori* information can be typical patterns or statistics obtained from a speech or noise database (DB). One of the predominant approaches in this category is non-negative matrix factorization (NMF) or dictionary learning [15]- [30]. There have also been other attempts such as finding the longest matching segments in the corpus segments [31], sparse combinations of the training data [32], and graph-based processing for transient noises [33]. These approaches are more robust to non-stationary noise environments since there is no strict assumption made on the nature of the noise in contrast to the statistical model-based methods.

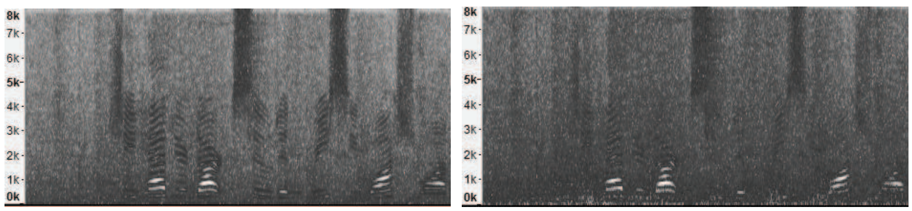
1.3 Measurements

For the check and compare the audio source separation performance, source-to-distortion ratio (SDR), source-to-interfering ratio (SIR), source-to-noise ration (SNR) and source-to-artifact ratio (SAR) are widely applied [42]. For example, SDR utilizes the target, interfering, noise, and artifact parts, SIR utilizes the target and interfering parts of the observed signal. Experiments involving typical mixtures and existing algorithms in [42] showed that these measures were relevant for algorithm evaluation and comparison. With respect to other existing performance measures, the main improvement is that it is not assume a particular separation algorithm nor a limited set of allowed distortions.

For the measurement of speech quality, perceptual evaluation of speech quality (PESQ) has been an important measurement [43] for a long time. It is a family of



(a) PESQ score of left: 4.50, PESQ score of right: 3.16



(b) PESQ score of left: 1.96, PESQ score of right: 1.11

Figure 1.2: Clean speech and speech contaminated by factory noise spectra (PESQ scores 4.5 and -0.5 mean ‘Excellent’ and ‘Bad’, respectively.).

standards comprising a test methodology for automated assessment of the speech quality as experienced by a user of a telephony system [43]. PESQ is a full-reference algorithm and analyzes the speech signal sample-by-sample after a temporal alignment of corresponding excerpts of reference and test signal. Except for that, [Y. Hu, 2008] proposed the three subjective rating scales (SIG, BAK, and OVAL) [44]. These measurement is consider the correlation between the several measurement, e.g., segment SNR, weighted spectral slope, PESQ, log-likelihood ratio, Itakura-Saito distance, cepstrum distance, and fwSNRseg. In this dissertation, we compared and checked each performance of audio source separation and speech enhancement by SDR and PESQ.

1.4 Outline of the dissertation

This dissertation proposes the diverse approaches and techniques of NMF-based audio source separation. NMF is belong to compositional model-based approaches [13], and it is a central algorithm of this dissertation. In chapter II, the compositional model is introduced and NMF is fully explained. Since the whole components of NMF analysis are nonnegative, NMF often leads to a part-based representation of the data, which may be desirable in many areas including image or visual signal processing, text information processing, audio signal processing, and music information retrieval [29], [45]- [47].

In chapter III, the process of NMF-based audio source separation is presented, and the several issues of the NMF-based approaches are discussed, e.g., the high dependency to the prior information of the sources, the over-lapped subspace between different source's bases, the non-uniqueness of the basis and encoding matrices, and the discarded information of the encoding vectors from the training phase. From chapter IV to VII, the diverse proposed methods for the above issues are introduced.

In chapter IV, we propose a cascaded structure that combines a statistical model-based enhancement and a template-based approach with simultaneous update of speech and noise bases. In virtue of the bases update considering the speech presence probability (SPP), the proposed approach can deal with the speech and noise patterns which were not included in the training database, and consequently is less vulnerable on the incomplete *a priori* information. Experimental results showed that the proposed algorithm outperformed not only the statistical model-based and NMF-based methods but also the combination of them.

In chapter V, we propose discriminative NMF that makes the higher reconstruction error to the other source than standard NMF does. Namely, the proposed DNMF gives the constraint to have a high reconstruction error of the other source, not target source. There are many candidates for basis vectors, because of non-convexity of NMF. The proposed DNMF finds a proper basis matrix that makes a high reconstruction error to the other source from the above candidates. In this dissertation, the application for the performance evaluation is speech enhancement, and perceptual evaluation speech quality (PESQ) [43] and signal-to-distortion (SDR) [42] are used for the measurement.

In chapter VI, we propose a novel approach to estimate the basis and encoding matrices for the NMF analysis. Exploiting the analogy between NMF analysis and data clustering, a systematic method for estimating the NMF basis matrix is proposed by combining the standard NMF basis training procedure and an efficient codebook learning algorithm. The proposed method borrows an idea from the modified k-means algorithm [48]. One of the prominent features of this algorithm is that it estimates the parameters incrementally, i.e. increases the number of bases at each iteration. In order to evaluate the performance of the proposed technique, we carried out an experiment on target source separation. In the experimental result, we can see that the proposed method outperformed the other bases initialization methods.

In chapter VII, we propose the penalty terms based on the prior knowledge on $\mathbf{H}$ in the separation phase for NMF-based source separation. We also extend our study in [49] to address the problem of possible mismatch between the training and test data levels by introducing a new penalty function of parameter training. Assuming that the statistical characteristics of the encoding vector for a specific source are stationary except for the level of the signal, we model the distribution of the components of the encoding vector as a multivariate exponential PDF (MVE) with a single time-varying scaling factor for each source. The parameters of the MVE are initially estimated from $\mathbf{H}^{train}$ and then continuously adjusted with suitable scaling factors to match the current input level. The scaling factor is estimated according to the maximum likelihood criterion in conjunction with temporal smoothing. Experimental results on audio source separation in which the target signal was speech showed that the proposed method could enhance the separation performance in term of the signal-to-distortion ratio (SDR) [42] even in the presence of the signal level mismatch.

In chapter VIII, we analyze several sparse terms for NMF and propose a sparse NMF with discriminative NMF (DNMF). The principle of sparsity is representing a phenomenon with as few variables as possible, which can make the encoding easy to interpret with good predictive power [50]. In order to promote sparsity of the encoding matrix, the objective function for NMF parameter estimation is modified to have an additional penalty term. However, more than a few previous works reported that the sparse NMF is not helpful to the source separation and the other applications. From the analysis, we found that the source separation performance has a high correlation with an anti-sparsity of basis matrix, and the anti-sparse $\mathbf{W}$ from sparse NMF on $\mathbf{H}$ is more proper to the source separation than those from anti-sparse NMF on $\mathbf{W}$. From the analysis, we propose that the bases become anti-sparse and discriminative simultaneously for the source separation. Furthermore, we compare the sparse term of the source separation phase, and we showed that a prior model-based sparse term is more powerful than the other sparse constraints for the separation phase. Experimental results on audio source separation showed that the proposed method can enhance the separation performance in term of the signal-to-distortion ratio (SDR) [42].

Chapter 2

Compositional model and NMF

2.1 Compositional model

Many classes of data can be consider as the combinations of the proper or latent parts. For this assumption, nonnegative condition of the data components is the crucial point. The nonnegative components allow additive combination that does not result in subtraction of any of the parts [13]. Namely, in order to apply the compositional model, nonnegative data should be assumed. This compositional model and data can be explained by standard basis, i.e, if dimension of data is 3, then each data can be represented by 3-standard basis. The concept of the compositional model can be also explained bicycle. The bicycle is the combination of several parts, e.g., wheels, saddle, main frame, tires, steering part, alarm whistle, and so on. These models, in conformance with the nature of the data, represent them as nonnegative linear combinations of parts, which themselves are also nonnegative to ensure that such a combination does not result in subtraction [13]. During the last few years, the compositional models have provided new paradigms to solve audio pro-

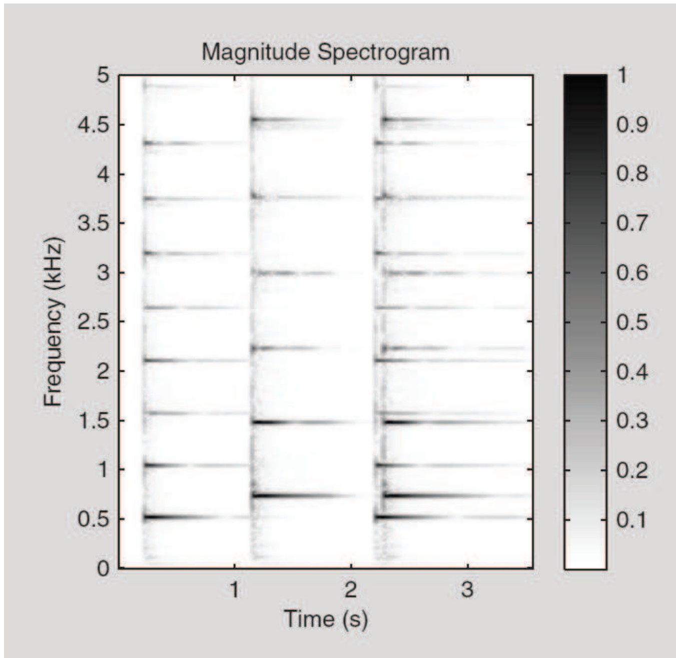


Figure 2.1: A magnitude spectrogram of a simple piano recording. Two notes are played in succession and then again in unison [13].

cessing problems, e.g., speech enhancement, target source separation, voice activity detection, audio event detection, robust recognition, and analysis of polyphonic music [4]- [21], [51]. Of course, the compositional models are firstly applied to nonsignal data such as counts of populations, but the compositional models have the flexibility to use diverse types of data if the data can be expressed by nonnegative values.

The basic premise underlying the application of compositional models to audio processing is that sound can be viewed as being compositional in nature [13]. This fact can be easily verified from the notes of piano. Fig. 2.1 shows a magnitude spectrogram of a simple piano recording [13]. Two of single note are played in series and these two notes are played at the same time. We can visually identify these notes using their unique harmonic structure. For the other example, we can consider the factory sound. The factory sound can be consider as the combination of several sound from the machines and human. The compositional framework for sound analysis builds upon these impressions: it characterizes the sounds from any source as a constructive composition of atomic sounds that are characteristic of the source and postulates that the decompositions of the signal into tis atomic parts may be achieved through the application of an appropriately constrained compositional model to an appropriate time-frequency representation of the signal [13].

Then, why is the compositional framework important for the sound analysis? [9] and [13] show the reason of this by the comparison with PCA and ICA or vector quantization (VQ). Fig. 2.2 and 2.3 show the reason why the compositional framework and the additive feature are important for the sound analysis. Fig. 2.2 shows each atom of PCA and ICA from the notes in Fig. 2.1. PCA and ICA discovers two bases that are actually combinations of the two notes, and their corresponding activations provide no indication of the actual composition of the sound. Namely, there is no physical interpretation. However, the atoms and activations from NMF seem to have a proper physical interpretation in Fig. 2.3. Each atom is presented each note of piano and each activation provides an exact activated information of the notes. Of course, we have assumed that the correct number of atoms, two, is known a priori, and this is generally not the case.

The compositional model-based approach can be expressed by several algorithms, e.g., dictionary learning, exemplar based algorithms, NMF, and probabilistic latent component analysis (PLCA) [9], [51]- [53]. The NMF models treat nonnegative time-frequency representations of the signal as matrices, which are decomposed into products of nonnegative component matrices. One of the matrices represents spectral patterns of the atomic parts and the other represents their activation to the

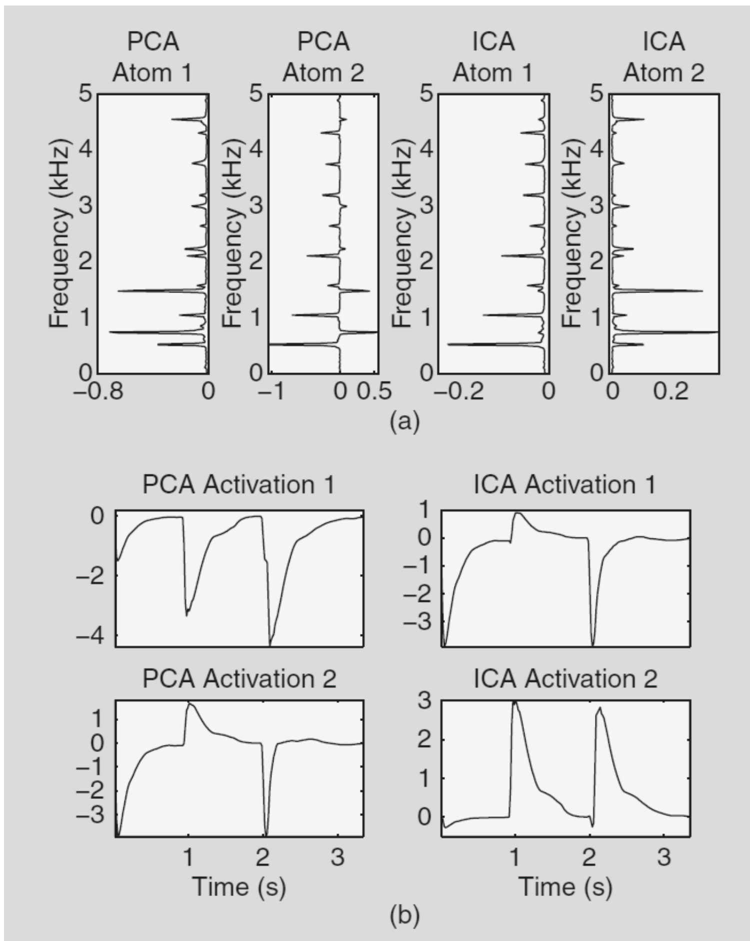


Figure 2.2: The PCA and ICA analysis of the data Fig. 2.1: (a) the learned PCA and ICA atoms and (b) their corresponding activations. [13]

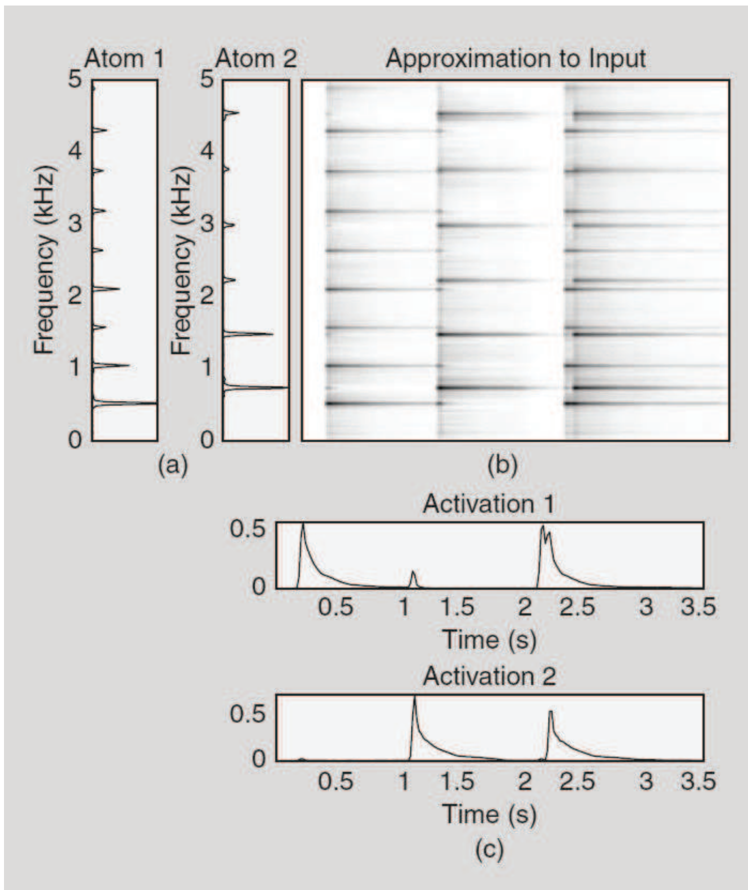


Figure 2.3: The NMF analysis of the data Fig. 2.1: (a) the discovered atoms and (c) their corresponding activations and (b) is the approximation to the input. [13]

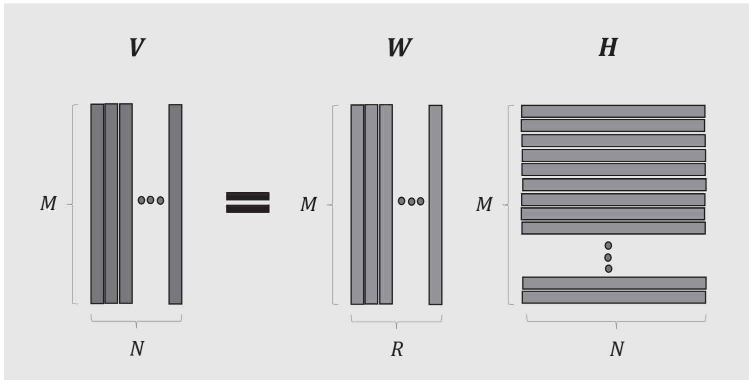


Figure 2.4: Nonnegative matrix factorization

signal over time [9], [13]. The PLCA models treat the nonnegative time–frequency representations as histograms drawn from a mixture of multivariate multinomial random variables representing the atomic parts [13], [54]. The two approaches can be shown to be equivalent as well as arithmetically identical under some circumstances [13], [55]. The motivation and approach are different between NMF and PLCA, but the update equations from the optimization are the same and the concept of NMF is more intuitive than PLCA. For this reason we choose the NMF model for the compositional model-based approach. In the next section, NMF is precisely explained.

2.2 NMF

NMF is one of the most popular methods for dictionary learning in which a nonnegative data matrix **V** is approximated by a product of a nonnegative basis matrix **W** and a nonnegative encoding matrix **H**. Fig. 2.4 shows the relationship between **V** and **WH**. Each column of **V** is a data vector (a magnitude spectrum of one time frame) and each column of **W** is basis vector. **H** indicates that how each basis vector is used for the reconstruction of $\mathbf{V}$. Generally, R is smaller than M, but R can be larger than M if the overcomplete basis set is needed. The objective function of NMF is given as the discrepancy between $\mathbf{V}$ and $\mathbf{WH}$, i.e.,

$$f(\mathbf{W}, \mathbf{H}) = D(\mathbf{V}|\mathbf{WH}) \tag{2.1}$$

where $\mathbf{V} \in \mathbb{R}_{+}^{M \times N}$, $\mathbf{W} \in \mathbb{R}_{+}^{M \times R}$, $\mathbf{H} \in \mathbb{R}_{+}^{R \times N}$ with R and $\mathbb{R}_{+}$ indicating the number of basis vectors and the set of nonnegative real numbers, respectively, and $D(a \mid b)$ denotes the divergence between a and b. The popular choices for the discrepancy measures are Euclidean distance (EuD), Kullback-Leibler divergence (KL), and Itakura-Saito divergence (IS). The most commonly used divergence in matrix factorization or matrix decomposition problem is EuD (squared error): $D(\mathbf{V}|\mathbf{WH}) = |\mathbf{V} - \mathbf{WH}|_{\mathbf{F}}^{\mathbf{2}}$. However, in the previous works, other divergence measures, KL and IS, have been found more appropriate for the audio signal [4]. The generalized KL and the IS are

$$D_{KL}(y|\hat{y}) = ylog(y/\hat{y}) - y + \hat{y}, \tag{2.2}$$

$$D_{IS}(y|\hat{y}) = y/\hat{y} - log(y/\hat{y}) - y - 1$$

where $\hat{y}$ is estimate of y. Fig. 2.5 shows the value of each discrepancy measurements according to the input value and the estimate [13]. The scale of the input affects the scale of the divergence or distance. The various discrepancy measures scale differently with their arguments. The squared error scales quadratically, the IS is scale invariant, while the KL scales linearly [13]. The scale invariant of IS may be a good feature as the purpose or the data, but it fails to distinguish between the noise floor and higher-energy target source signals. From the analysis in Fig. 2.5, KL provides a good compromise between the two [4]. In this dissertation, we choose two discrepancy measures, EuD and IS, for the audio source separation based on NMF. The objective

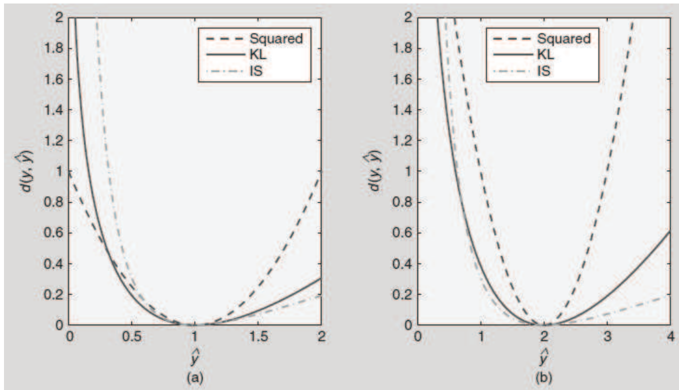


Figure 2.5: An illustration of the typical divergence functions used in NMF. The divergences are calculated for an observation (a) $y = 1$ and (b) $y = 2$ as the function of the model $\hat{y}$ (Squared denotes EuD.). [13]

function with KL is given as:

$$D(\mathbf{V}|\mathbf{WH}) = \sum_{m,n} \mathbf{V}_{m,n} log \frac{\mathbf{V}_{m,n}}{(\mathbf{WH})_{m,n}} - \mathbf{V}_{m,n} + (\mathbf{WH})_{m,n} \tag{2.3}$$

where $\mathbf{A}_{m,n}$ denotes the m-th row and n-th column component of the matrix $\mathbf{A}$.

2.2.1 Update rules: MuR, PGD

Since this is a jointly non-convex function of the basis and encoding matrices, alternating updates of $\mathbf{W}$ and $\mathbf{H}$ are usually performed [9]. The alternative update rule means that $\mathbf{W}$ is fixed when the components of $\mathbf{H}$ are updated and $\mathbf{H}$ is fixed when the components of $\mathbf{W}$ are updated. Other issue of the parameter update is an initialization of each components. Since the objective function of NMF is non-convex, the optimized results of $\mathbf{H}$ and $\mathbf{W}$ are differ according to the initialization. Generally, nonnegative random values are applied for the initialization and it has shown a comparatively proper performance. In order to resolve the non-unique factorization problem, it is needed to impose some constraints on the structures of $\mathbf{W}$ or $\mathbf{H}$. In our work, all the column vectors of $\mathbf{W}$ are constrained to have a unit L_1-norm or L_2-norm at each iteration of the update rule.

A well-known approach to estimate $\mathbf{W}$ and $\mathbf{H}$ is the multiplicative update rule (MuR) [9] which is simple to implement and shown to yield good results. The general gradient descent method needs a learning rate, but the decision of the learning rate may be heuristic and trouble. MuR resolve this issue by the decision of the learning rate from its parameters, $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{H}$. First of all, in order to update $\mathbf{H}$, $\mathbf{W}$ in the objective function is fixed. We calculate each update equation from the objective function by KL. Taking the gradient with respect to $\mathbf{H}$ gives:

$$\frac{\partial}{\partial \mathbf{H}_{r,n}} D(\mathbf{V}|\mathbf{W}\mathbf{H}) = \sum_{i=1}^{M} \mathbf{W}_{i,r} - \sum_{i=1}^{M} \frac{\mathbf{V}_{i,n}\mathbf{W}_{i,r}}{\sum_{k=1}^{R} \mathbf{W}_{i,k}\mathbf{H}_{k,n}}. \tag{2.4}$$

The gradient algorithm then states:

$$\mathbf{H}_{r,n} \Longleftarrow \mathbf{H}_{r,n} - \eta_{r,n} \frac{\partial}{\partial \mathbf{H}_{r,n}} D(\mathbf{V}|\mathbf{W}\mathbf{H}), \tag{2.5}$$

$$\mathbf{H}_{r,n} \Longleftarrow \mathbf{H}_{r,n} + \eta_{r,n}[\sum_{i=1}^{M} \frac{\mathbf{V}_{i,n}\mathbf{W}_{i,r}}{\sum_{k=1}^{R} \mathbf{W}_{i,k}\mathbf{H}_{k,n}} - \sum_{i=1}^{M} \mathbf{W}_{i,r}] \tag{2.6}$$

where η indicates the learning rate. In order to erase η in (2.6), we can force η as

$$\eta_{r,n} = \frac{\mathbf{H}_{r,n}}{\sum_{i=1}^{M} \mathbf{W}_{i,r}}. \tag{2.7}$$

The equation (2.7) gives the MuR:

$$\mathbf{H}_{r,n} \leftarrow \mathbf{H}_{r,n} \frac{\sum_{i=1}^{M} \frac{\mathbf{W}_{i,r}\mathbf{V}_{i,n}}{\sum_{k=1}^{R} \mathbf{W}_{i,k}\mathbf{H}_{k,n}}}{\sum_{i=1}^{M} \mathbf{W}_{i,r}}. \tag{2.8}$$

The update rule of $\mathbf{W}$ is obtained like to those of $\mathbf{H}$. Taking the gradient with respect to $\mathbf{W}$ gives:

$$\frac{\partial}{\partial \mathbf{W}_{m,r}} D(\mathbf{V}|\mathbf{W}\mathbf{H}) = \sum_{j=1}^{N} \mathbf{H}_{r,j} - \sum_{j=1}^{N} \frac{\mathbf{V}_{m,j}\mathbf{H}_{r,j}}{\sum_{k=1}^{R} \mathbf{W}_{m,k}\mathbf{H}_{k,j}}. \tag{2.9}$$

The gradient algorithm then states:

$$\mathbf{W}_{m,r} \Longleftarrow \mathbf{W}_{m,r} - \theta_{m,r}\frac{\partial}{\partial \mathbf{W}_{m,r}} D(\mathbf{V}|\mathbf{WH}), \tag{2.10}$$

$$\mathbf{W}_{m,r} \Longleftarrow \mathbf{W}_{m,r} + \theta_{m,r}[\sum_{j=1}^{N} \frac{\mathbf{V}_{m,j}\mathbf{H}_{r,j}}{\sum_{k=1}^{R} \mathbf{W}_{m,k}\mathbf{H}_{k,j}} - \sum_{j=1}^{N} \mathbf{H}_{r,j}] \tag{2.11}$$

where θ indicates the learning rate for $\mathbf{W}$. In order to erase θ in (2.11), we can force θ as

$$\theta_{m,r} = \frac{\mathbf{W}_{m,r}}{\sum_{j=1}^{N} \mathbf{H}_{r,j}}. \tag{2.12}$$

The equation (2.12) gives the MuR:

$$\mathbf{W}_{m,r} \leftarrow \mathbf{W}_{m,r}\frac{\sum_{j=1}^{N} \frac{\mathbf{H}_{r,j}\mathbf{V}_{m,j}}{\sum_{k=1}^{R} \mathbf{W}_{m,k}\mathbf{H}_{k,j}}}{\sum_{j=1}^{N} \mathbf{H}_{r,j}}. \tag{2.13}$$

The MuR can be applied EuD, and the update rules are given as:

$$\mathbf{H}_{r,n} \leftarrow \mathbf{H}_{r,n}\frac{[\mathbf{W}^T\mathbf{V}]_{r,n}}{[\mathbf{W}^T\mathbf{WH}]_{r,n}}, \tag{2.14}$$

$$\mathbf{W}_{m,r} \leftarrow \mathbf{W}_{m,r}\frac{[\mathbf{VH}^T]_{m,r}}{[\mathbf{WHH}^T]_{m,r}} \tag{2.15}$$

where $[\cdot]_{a,b}$ indicates the a-th row and the b-th column component of the matrix in [], and ^{T} denote matrix transposition. The optimized solutions of $\mathbf{W}$ and $\mathbf{H}$ are obtained by the alternative iteration process of the set of (2.3) and (2.13) or the set of (2.14) and (2.15). However, the objective function of NMF is non-convex, the optimized solution may be different according to the initialization of the parameters. Namely, the NMF analysis does not guarantee the uniqueness. Furthermore, it is hard to exploit the prior knowledge of the basis. For these reason, the nonnegative random values are usually and generally applied for the initialization of $\mathbf{W}$ and $\mathbf{H}$, and this approach has shown a proper performance for the source separation. The iteration process of the above equations is usually performed until the convergence condition is satisfied or for the fixed iteration number.

There are diverse approaches for the update or obtain the parameters of NMF, e.g., MuR, projected gradient descent (PGD), NeNMF, Lasso-based approaches [9], [56]- [57]. In the diverse approaches, PGD is relatively easy to perform and shows a faster processing time than MuR. If the projected gradient descent (PGD) method and the Euclidean distance are used as an optimization method and a distance measure, the update rules for the encoding and basis matrices during the training phase are given as [56]

$$\mathbf{H} \leftarrow \mathbf{H} - \alpha_H(\mathbf{W}^T\mathbf{W}\mathbf{H} - \mathbf{W}\mathbf{W}^T\mathbf{V}), \tag{2.16}$$

$$\mathbf{W} \leftarrow \mathbf{W} - \alpha_W(\mathbf{W}\mathbf{H}\mathbf{H}^T - \mathbf{V}\mathbf{H}^T). \tag{2.17}$$

Each learning rate, α_H and α_W, is repeatedly increased if the below condition (2.18) is satisfied.

$$(1-\sigma)\triangledown f(\check{\mathbf{x}})^T(\hat{\mathbf{x}} - \check{\mathbf{x}}) + \frac{1}{2}(\hat{\mathbf{x}} - \check{\mathbf{x}})\triangledown^2 f(\check{\mathbf{x}})(\hat{\mathbf{x}} - \check{\mathbf{x}}) \leqq 0 \tag{2.18}$$

A common choice of σ is 0.001, and $\check{\mathbf{x}}$ and $\hat{\mathbf{x}}$ indicate the next $\mathbf{x}$ after the gradient descent method and the present $\mathbf{x}$ before the gradient descent method. This condition is can be applied for the NMF analysis and it shows a faster convergence than those of MuR. In this dissertation, we applied MuR or PGD for the NMF analysis. The detailed process will be explained each chapter.

2.2.2 Modified NMF

The cost function D can be obtained by β-divergence which for $\beta = 1$ yields KLD and for $\beta = 2$ yields EuD [58]. In this case, the MuRs are given as follows:

$$\mathbf{H}_{i,j} \leftarrow \mathbf{H}_{i,j} \otimes \frac{[\mathbf{W}^T(\mathbf{V} \otimes \mathbf{\Lambda}^{\beta-2})]_{i,j} + \lambda[\nabla f^-(\mathbf{H})]_{i,j}}{[\mathbf{W}^T\mathbf{\Lambda}^{\beta-1}]_{i,j} + \lambda[\nabla f^+(\mathbf{H})]_{i,j}}, \tag{2.19}$$

$$\mathbf{W}_{i,j} \leftarrow \mathbf{W}_{i,j} \otimes \frac{[(\mathbf{\Lambda}^{\beta-2} \otimes \mathbf{V})\mathbf{H}^T]_{i,j}}{[\mathbf{\Lambda}^{\beta-1}\mathbf{H}^T]_{i,j}} \tag{2.20}$$

where $\mathbf{\Lambda} = \mathbf{WH}$, and $\bigotimes$ denotes the element-wise multiplication of matrix. $\mathbf{H}$ and $\mathbf{W}$ are obtained by iterative application of the update rules (2.19) and (2.20) for a fixed number of iterations [14]- [18]. At each iteration, the columns of $\mathbf{W}$ are normalized to have unit L_1- or L_2-norm because the NMF analysis is not satisfied the uniqueness and it offers several benefits. We here denote this approach **Standard**.

The forcing normalization at each iteration makes a certain mismatch and it may change the value of the sparse term [59]. In [60], the cost of sparse term in **Standard** highly increased according to iteration of (2.19) and (2.20). The approach with a modification of the objective function in [61] solves the above problem, and the objective function is given as

$$D(\mathbf{V}|\tilde{\mathbf{W}}\mathbf{H}) + \lambda f(\mathbf{H}) \tag{2.21}$$

where $\tilde{\mathbf{W}}$ is the normalized version of $\mathbf{W}$. The MuR for this approach when $\mathbf{W}$ is normalized to have unit L_2-norm becomes

$$\mathbf{W}_{i,j} \leftarrow \mathbf{W}_{i,j} \otimes \frac{[(\mathbf{\Lambda}^{\beta-2} \otimes \mathbf{V})\mathbf{H}^T + \tilde{\mathbf{W}} \otimes (\mathbf{1}\mathbf{1}^T(\tilde{\mathbf{W}} \otimes (\mathbf{\Lambda}^{\beta-1}\mathbf{H}^T)))]_{i,j}}{[\mathbf{\Lambda}^{\beta-1}\mathbf{H}^T + \tilde{\mathbf{W}} \otimes (\mathbf{1}\mathbf{1}^T(\tilde{\mathbf{W}} \otimes ((\mathbf{\Lambda}^{\beta-2} \otimes \mathbf{V})\mathbf{H}^T)))]_{i,j}}. \tag{2.22}$$

The MuR for $\mathbf{H}$ is the same to (2.19), and $\mathbf{W}$ is normalized to have unit L_2-norm after each iteration. The MuR for this approach when $\mathbf{W}$ is normalized to have unit

L_1-norm becomes

$$\mathbf{W}_{i,j} \leftarrow \mathbf{W}_{i,j} \otimes \frac{[(\mathbf{\Lambda}^{\beta-2} \otimes \mathbf{V})\mathbf{H}^T + \mathbf{W} \otimes (\mathbf{1}\mathbf{1}^T(\tilde{\mathbf{W}} \otimes (\mathbf{\Lambda}^{\beta-1}\mathbf{H}^T)))]_{i,j}}{[\mathbf{\Lambda}^{\beta-1}\mathbf{H}^T + \mathbf{W} \otimes (\mathbf{1}\mathbf{1}^T(\tilde{\mathbf{W}} \otimes ((\mathbf{\Lambda}^{\beta-2} \otimes \mathbf{V})\mathbf{H}^T)))]_{i,j}}. \tag{2.23}$$

We denote this approach **Modified**. Interested readers are referred to [60] for more details.

Chapter 3

NMF-based audio source separation and issues

3.1 NMF-based audio source separation

In order to separate audio source from the mixture signals using NMF, the training phase is essential for a beforehand operation. The whole basis matrices of the audio source which we can know beforehand should be trained from each proper database. The case of audio source separation can be divided into two-case according to the prior information of the sources, i.e., supervised and semi-supervised cases. We assume that the mixture signals consist two-class source, target and interfering sources in this dissertation. In the supervised case, we know the types of target and interfering sources in advance. On the other hand, the only target or interfering source type is known in the semi-supervised case. In this case, the basis matrix of the unknown type source is updated by the mixture with fixed the basis matrix of the known type source [62]. Of course, the performance of the semi-supervised case is lower than those of the supervised case. In this section, we explain the NMF-based audio source separation in the supervised case.

Some conditions of the NMF-based audio source separation are important to the separation performance, e.g., the number of bases (R) and iteration of the MuR, but the decision of these factors is difficult. In the previous work [21], the optimized number of bases is different as the type of the source and the experiment condition. Fortunately, NMF shows the stable separation performance in some range of R. In our experimental condition, $R \in [30, 128]$ has shown a proper performance when the FFT size is 257. If R is extremely small, the reconstruction error of the source is high but the discriminative to the other source may be increased. On the contrary, if R is extremely high, the reconstruction error of the source is small but the discriminativity may be decreased. These two factors, reconstruction error and the discriminativity, are trade-off relation in the audio source separation. The number of iteration of MuR, also, has a influence on the separation performance. If the iteration of MuR is close to 1 or 2, the update of $\mathbf{H}$ is not sufficient and the separation performance is lower than those of the optimized performance. On the contrary, if the iteration of MuR is higher than the specific number, the separation performance becomes low. Since the objective functions of the training phase and the separation phase are difference each other, the high iteration number of MuR can bring a performance degradation. In the previous work [21], the iteration of MuR $\in [10, 40]$ has shown a proper performance.

Fig. 3.1 is a block diagram of the general NMF-based audio source separation. Let $Y(t) \in \mathbb{C}^{K\times 1}$, $S(t) \in \mathbb{C}^{K\times 1}$ and $N(t) \in \mathbb{C}^{K\times 1}$ denote the short-time Fourier transform (STFT) coefficients of the observed signal, the target audio signal and the interfering signals, respectively, for the t-th frame where $\mathbb{C}$ indicates the set of complex numbers

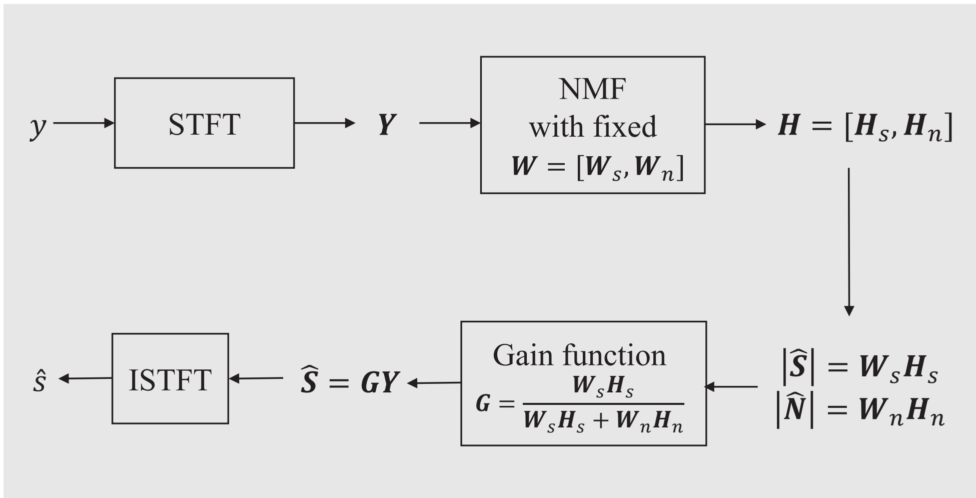


Figure 3.1: Block diagram of the general NMF-based audio source separation.

and K denotes the number of frequency bins. It is assumed that the interferences are additive, i.e., $Y(t) = S(t) + N(t)$. For each frame of the observed signal, the magnitude spectra $V(t) = |Y(t)| \in \mathbb{R}_+^{K\times 1}$ are approximated as $V(t) \approx \mathbf{W}H(t)$ with a fixed basis matrix $\mathbf{W} = [\mathbf{W}_S\ \mathbf{W}_N] \in \mathbb{R}_+^{K\times R}$ in which $\mathbf{W}_S \in \mathbb{R}_+^{K\times R_s}$ and $\mathbf{W}_N \in \mathbb{R}_+^{K\times R_n}$ are the basis matrices for the target signal and interferences, respectively, and $|\cdot|$ denotes element-wise magnitude. It is noted that the variables in boldface capital letters denote matrices while those which are not in the boldface represent vectors. $\mathbf{W}_S$ and each section of $\mathbf{W}_N$ are trained separately through the NMF analysis of the magnitude spectra of the target signal only and each interfering signal only, respectively. $H(t) = [H_S^T(t)\ H_N^T(t)]^T \in \mathbb{R}_+^{R\times 1}$ for $V(t)$ is initialized with nonnegative random numbers and estimated through the iteration given in (2.8) or (2.14) or (2.16) where ^{T} denotes matrix transposition. Regardless of how the bases are trained, $H(t)$ can be updated by any update equation. Once $H(t)$ is obtained, the target and interfering signals can be estimated as follows:

$$|\hat{S}(t)| = \mathbf{W}_S H_S(t), \tag{3.1}$$

$$|\hat{N}(t)| = \mathbf{W}_N H_N(t)$$

where $|\hat{S}(t)|$ and $|\hat{N}(t)|$ denote the magnitude spectra estimates of the target and interfering signals. Instead of directly using the estimated magnitude spectra in (3.1), a spectral gain function similar to the Wiener filter is adopted in [10], [17], and [49] where the final speech estimate is obtained through the filtering as given by

$$\hat{S}(t)^{final} = \frac{|\hat{S}(t)|}{|\hat{S}(t)| + |\hat{N}(t)|} \otimes Y(t) \tag{3.2}$$

where $\frac{A}{B}$ and $\otimes$ respectively denote element-wise division and multiplication of the vectors.

3.2 Problems of NMF in audio source separation

3.2.1 A high dependency to the prior knowledge

The papers related audio source separation generally assume that we know the information of the target and interfering sources, i.e., the experiment is performed in the supervised case. However, in practically, it is impossible to know the whole types of the source before the audio source separation. We can think that the semi-supervised case is reasonable, that is, we can easily designate the target source type. In this case, the target bases are trained before the separation phase and these are fixed during the separation phase. The interfering bases are generally estimated from the observation signals [29], [36], but its separation performance severely degraded. Fig. 3.2 shows why the interfering bases from the semi-supervised case make a performance degradation. Generally, the target bases can not reconstruct the target

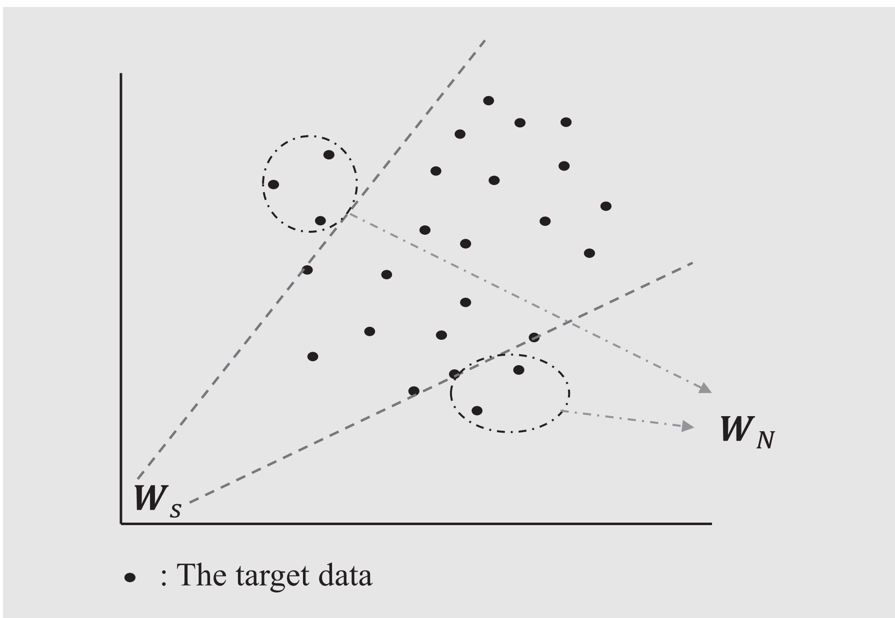


Figure 3.2: The reason why the interfering bases from the semi-supervised case make a performance degradation (The target bases perfectly represent the target signals, and the data points in a circle are covered by the interfering bases.).

signals perfectly. Then, the interfering bases learn to cover the remain part of the target signals from the update phase of the semi-supervised case. Consequently, the interfering bases are resemble to the target signals as time goes on. Furthermore, the interfering signals can consist of more than two types of sources. It needs a high complexity to applied many types of bases to the separation phase. Namely, the proper update and estimate the interfering bases is essential for the NMF-based audio source separation.

3.2.2 A overlapped subspace between the target and interfering basis matrices

Generally, the bases of each type are trained separately to the other source. Namely, $\mathbf{W}_S$ is estimated by only the target source signals without the interfering source signals. This means that each source's bases can reconstruct and represent one's own self, but each bases may not be discriminative to the other source signals. This occurs the overlapped subspace between the target and interfering basis matrices. The audio signals on the magnitude spectrum can be similar to the other source' signals, that is, each subspace from the linear combination of the bases can be overlapped. In order to reduce this overlapped subspace, discriminative NMF (DNMF) has been proposed in many previous works [6]- [8]. DNMF utilizes more than two types of source data for the bases estimation. The goal of DNMF is presented in Fig. 3.3. We want to reduce the subspace of each source bases. However, DNMF may occur some problem. If the target bases are extremely trained by DNMF, a residual noise decreases but target distortion increases. If the interfering bases are extremely trained by DNMF, the target distortion decreases but the residual noise increases. Namely, DNMF has a trade-off between the target distortion and residual

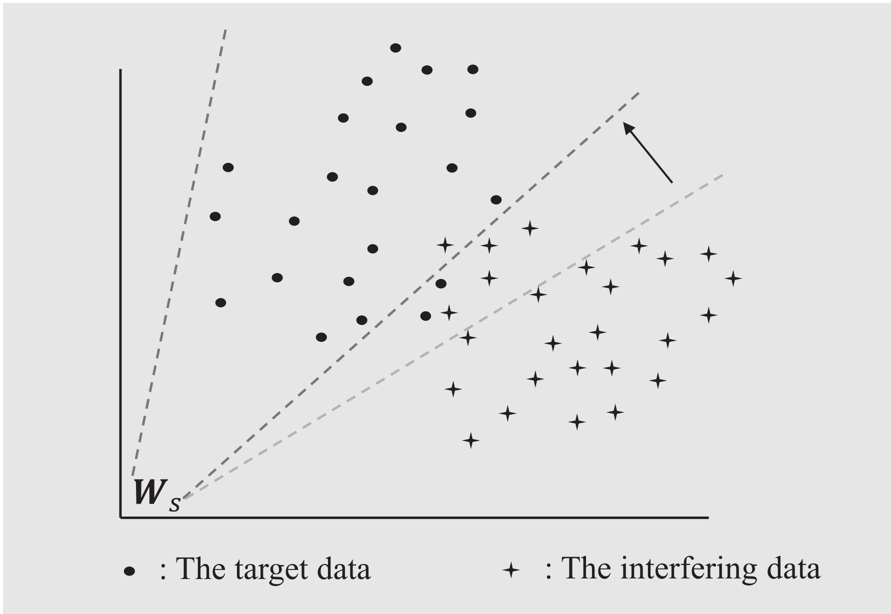


Figure 3.3: DNMF controls the subspace of the bases.

noise. However, the previous works show a little performance enhancement.

3.2.3 A non-uniqueness of the bases

The objective function of NMF is non-convex. It means that the bases from the optimization phase are varied as the initial values of the components. It is hard to decide the initial values of the bases and encoding. For this reason, random nonnegative values are usually applied for the initial values and it shows a proper performance. However, if the random initialization can be stuck to a local minimum, which implies that the overall performance may significantly depend on the initial parameter values. In order to resolve this issue, some previous works utilize the centroids of the clustering [63]- [68]. This approach obtains the centroid which number is the same to the number of bases and these centroids are used for the initialization of the bases.

In [69], singular value decomposition (SVD) is utilized for the initialization. However, these approaches may not be proper to the NMF analysis because NMF has the part-based feature but these approaches do not utilize the part-based feature. For the proper initialization of the bases, we need the method based on part-based analogy.

3.2.4 A prior knowledge of the encoding vectors

In the NMF analysis, the basis matrix is the most important, but the encoding matrix is relatively disregarded. After the training phase, the basis matrix is saved and fixed, and the encoding matrix is usually discarded. However, the encoding matrix has a crucial information of the basis use. For example, we can extract the bases from the speech database, and some bases are usually applied at the same time for the specific phoneme. On the other hand, some bases are not utilized at the same time experimentally. Fig. 3.4 shows a relation between the basis and encoding matrices and a role of each matrix. For the piano bases (dictionary), each MIDI note is used to the training phase. (c) and (d) in Fig. 3.4 denote the encoding matrix (activation matrix). The activation matrix shows how each basis is used according to the observation signals and time. To utilize the information of the encoding or activation matrix, the previous work in [15] applied a Gaussian distribution to the encoding vectors. It makes a performance enhancement, but the Gaussian distribution seems to be inadequacy to the encoding vectors. We can enhance the separation performance if we utilize the encoding vectors from the training phase properly.

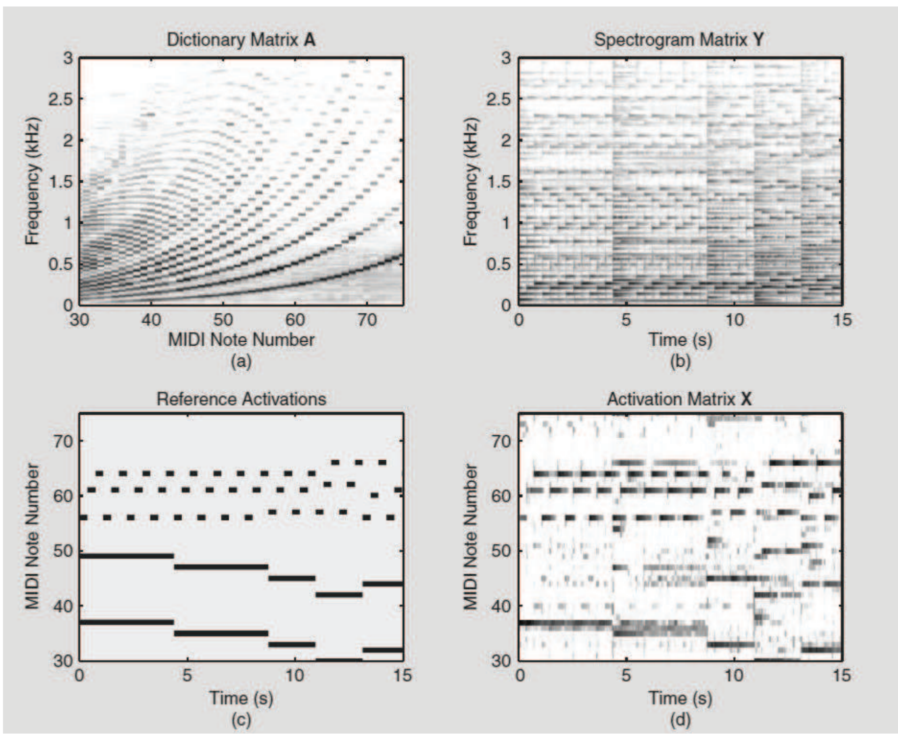


Figure 3.4: A music analysis example where a polyphonic mixture spectrogram (b) is decomposed into a set of note activations (d) using a dictionary (bases) consisting of spectra of piano notes (a) The reference activations are given in (c). [13]

3.2.5 Sparse NMF for the source separation

Over the recent years, the sparse concept has been applied to compositional models [13]. The compositional models include dictionary learning, exemplar-based approaches, and NMF [9], [51], [53]. Among them, the methods based on NMF have shown impressive results in single channel source separation [12]- [14]. In order to promote sparsity of the encoding matrix, the objective function for NMF parameter estimation is modified to have an additional penalty term. The most popular choice for this penalty is the L_1-norm of $\mathbf{H}$ although there are other alternatives such as the $L_{1/2}$ quasi-norm and the geometric mean of the encoding vector [12], [70]. Sparse NMF is applied to different form according to the purpose, e.g., sparse on $\mathbf{H}$ or $\mathbf{W}$ or both of $\mathbf{H}$ and $\mathbf{W}$ [70], [71]. However, more than a few previous works reported that the sparse NMF is not helpful to the source separation and the other applications. In [12], [72] and [73], L_1-norm sparse term degraded the separation performance. In [4], modified L_1-norm sparse term is proposed but it did not obtain a performance improvement. In [37], the performance increased when L_1-norm sparse term was applied to only the noise basis. The work in [47] applied sparse NMF to the document classification, but it could not make a performance improvement. The above works did not analyze the reason of the performance degradation. On the other hand, the approach in [60]- [62] is to directly reformulate the objective function including a normalized version of basis, and L_1-norm sparse term showed a performance improvement. The sparse concept can be helpful for the NMF-based audio source separation, and for this, the precise analysis and proper experiments are need.

Chapter 4

Online bases update

4.1 Introduction

As mentioned in chapter I, the statistical model-based and template-based approaches are major techniques of single channel speech enhancement [24]- [36]. The statistical model-based approach generally assumed that speech and noise have separate parametric distributions for which the parameters are estimated from the input signal [24]- [27]. One of the significant advantages of the statistical model-based techniques is that the models do not need to be trained *a priori.* Since, however, the statistical models are constructed based on a stationarity assumption, the performance deteriorates when the background noise is highly non-stationary. On the other hand, the template-based techniques utilize specific types of the *a priori* information of speech or noise [15]- [33]. One of the predominant approaches in this category is non-negative matrix factorization (NMF) or dictionary learning [15]- [30]. These approaches are more robust to non-stationary noise environments since there is no strict assumption made on the nature of the noise in contrast to the statistical

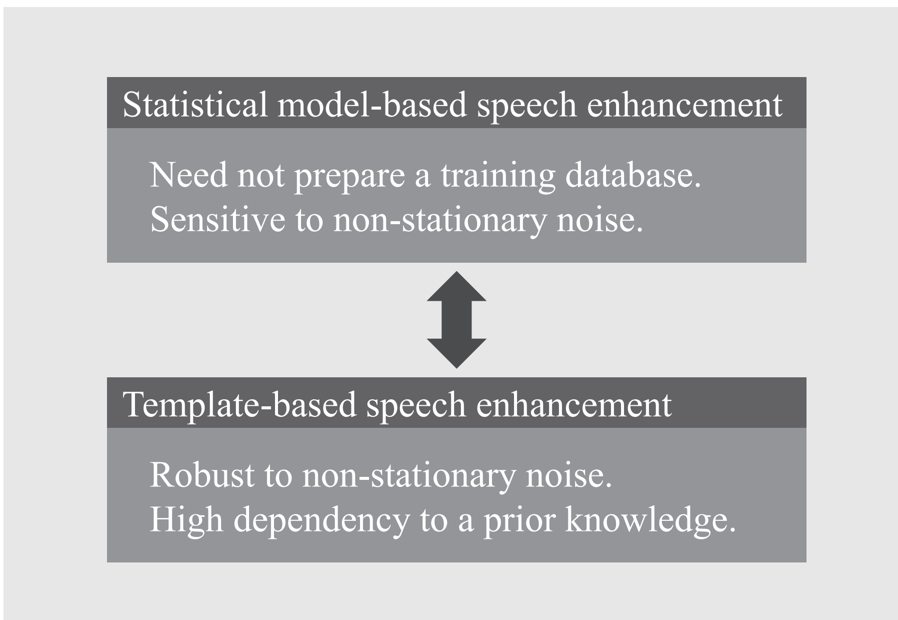


Figure 4.1: A comparison between the statistical model-based and the template-based speech enhancement.

model-based methods. However, if the actual noise is far different from the trained noise model or it fits closely to the trained speech model, the performance degrades seriously [34].

A number of attempts have been made to combine the two aforementioned techniques to achieve better performance. In [31], a template-based method is used to estimate the speech magnitude spectrum, while it is applied to obtain the noise power spectral density (PSD) in [20]. These two methods compute Wiener filter type gain functions using the PSD's obtained from the template-based approaches. In contrast, [35] applies a template-based algorithm to the output of a Wiener filter. This method can take advantage of both the statistical model-based and template-based approaches, but the Wiener filter output may be distorted without any update of the model. We can also find the combination between NMF-based enhancement and VAD in [36] where it is reported that the performance degrades if the trained noise model is different from the actual noise environment.

A majority of template-based approaches employ batch processing which determines the bases based on the entire training dataset [4], [15], [10], [37]. Recently, several on-line methods have been proposed for the bases update. In [29] and [38], the bases are updated based on buffered input frames, while they are updated based on a matrix that summarizes the information of the past and the present inputs in [39] and [40]. In [41], previous basis matrix and the current input are used for bases update. Unfortunately, however, these methods cannot be directly applicable to update multiple sets of bases simultaneously when each set corresponds to separate source.

In this chapter, we propose a cascaded structure that combines a statistical model-based enhancement and a template-based approach with simultaneous update of speech and noise bases. In virtue of the bases update considering the speech presence probability (SPP), the proposed approach can deal with the speech and noise patterns which were not included in the training database, and consequently is less vulnerable on the incomplete *a priori* information. Experimental results showed that the proposed algorithm outperformed not only the statistical model-based and NMF-based methods but also the combination of them.

4.2 NMF-based speech enhancement using spectral gain function

In the NMF analysis of the magnitude spectra, a data set $V\in \mathbb{R}^{M\times N}$ is reconstructed by the product of a basis matrix $W\in \mathbb{R}^{M\times R}$ and an encoding matrix $H\in \mathbb{R}^{R\times N}$ ($V \approx WH$) where M and N denote the numbers of frequency bins and time frames, respectively, and r is the number of basis vectors. For speech enhancement, we assume that W consists of speech basis matrix $W_s\in \mathbb{R}^{M\times R_s}$ and noise basis vectors $W_n\in \mathbb{R}^{M\times R_n}$, i.e., $W = [W_s\ W_n]\in \mathbb{R}^{M\times (R_s+R_n)}$ where R_s and R_n indicate the numbers of corresponding basis vectors while H becomes $H = [H_s^T\ H_n^T]^T\in \mathbb{R}^{(R_s+R_n)\times N}$ with ^{T} denoting matrix transpose [15]- [37]. In the training stage, W_s and W_n are trained separately with clean speech and noise, respectively. If the KLD is chosen as a distance metric, the update rule is given as [9]

$$H_i \leftarrow H_i \otimes \frac{W_i^T \frac{V_i}{W_i H_i}}{W_i^T \mathbf{1}}, \tag{4.1}$$

$$W_i \leftarrow W_i \otimes \frac{\frac{V_i}{W_i H_i} H_i^T}{\mathbf{1} H_i^T} \tag{4.2}$$

where subscript i indicates either speech or noise, $V_i\in \mathbb{R}^{M\times N_i}$ is constructed by stacking the magnitude spectra of the training DB for each source with N_i denoting the total number of frames in the training DB for source i, $\bigotimes$ and $\frac{a}{b}$ denote the element-wise multiplication and division of matrices, and $\mathbf{1}$ is a square matrix of proper size with all its elements equal to one. The updates given by (4.1) and (4.2) are iterated for a sufficient number of times.

At the speech enhancement stage, an optimal spectral gain is determined based on the speech and noise estimates derived from the NMF analysis. Let $Y(t) \in \mathbb{C}^{M\times 1}$, $S(t) \in \mathbb{C}^{M\times 1}$ and $N(t) \in \mathbb{C}^{M\times 1}$ denote the short-time Fourier transform (STFT) coefficients of the noisy speech, clean speech and noise, respectively, for the t-th frame. Then $Y(t) = S(t) + N(t)$ according to the additive noise assumption. In this work, the NMF analysis is performed on the magnitude spectrum domain with the data set $V(t)$ being given by $V(t) = |Y(t)| \in \mathbb{R}^{M\times 1}$ where $|\cdot|$ denotes element-wise magnitude.

In each frame of the input data, an encoding vector $H(t) = [H_s(t)^T \; H_n(t)^T]^T \in \mathbb{R}^{(R_s+R_n)\times 1}$ is computed using the iteration (2.16) while fixing the basis matrix W. For this iteration, $H_s(t)$ is initialized by $H_s(t-1)$, the encoding vector estimated in the previous frame, and $H_n(t)$ is randomly initialized. For the stopping rule, we apply the normalized stopping criterion [38] with a maximum number of iteration.

Once $H(t)$ is calculated, the speech and noise magnitude spectra estimates, $|\hat{S}(t)|$ and $|\hat{N}(t)|$, are obtained as

$$|\hat{S}(t)| = W_s H_s(t), \qquad |\hat{N}(t)| = W_n H_n(t). \tag{4.3}$$

In our approach to speech enhancement, these speech and noise magnitude spectra estimates are applied to derive the spectral gain function $G(m, t)$ with m indicating the frequency index. The spectral gain function $G(m, t)$ is formulated based on a specific statistical model and an optimality criterion, and in this work we employ the minimum mean square error-log spectral amplitude (MMSE-LSA) [24] technique where the gain is given by

$$G(m,t) = \frac{\xi(m,t)}{1+\xi(m,t)} \exp\left(\frac{1}{2}\int_{\nu(m,t)}^{\infty} \frac{e^{-t}}{t} dt\right) \tag{4.4}$$
$$\nu(m,t) = \frac{\gamma(m,t)\xi(m,t)}{1+\xi(m,t)}$$

in which $\xi(m,t)$ is the *a priori* signal-to-noise ratio (SNR) and $\gamma(m,t)$ is the *a posteriori* SNR for the m-th frequency bin at frame t. As in most of the conventional statistical model-based speech enhancement techniques, the *a priori* and *a posteriori* SNR's are estimated through temporal smoothing of the power spectra given as follows:

$$P_s(m,t) = \tau_s P_s(m,t-1) + (1-\tau_s)[(|\hat{S}(t)|)_m]^2, \tag{4.5}$$
$$P_n(m,t) = \tau_n P_n(m,t-1) + (1-\tau_n)[(|\hat{N}(t)|)_m]^2,$$
$$\xi(m,t) = \frac{P_s(m,t)}{P_n(m,t)}, \quad \gamma(m,t) = \frac{[(V(t))_m]^2}{P_n(m,t)}$$

where $P_s(m,t)$ and $P_n(m,t)$ respectively denote the smoothed speech and noise PSDs for the m-th frequency bin at frame t with τ_s and τ_n being the smoothing factors, and $(\cdot)_m$ indicates the m-th element. Finally, the enhanced speech spectrum at the t-th frame is obtained according to $(\hat{S}^{Final}(t))_m = G(m,t)(Y(t))_m$ for $m = 1, 2, \cdots, M$.

4.3 Speech enhancement combining statistical model-based and NMF-based methods with the on-line bases update

The proposed speech enhancement system has a cascaded structure in which the first stage is a statistical model-based enhancement (SE) while the second stage

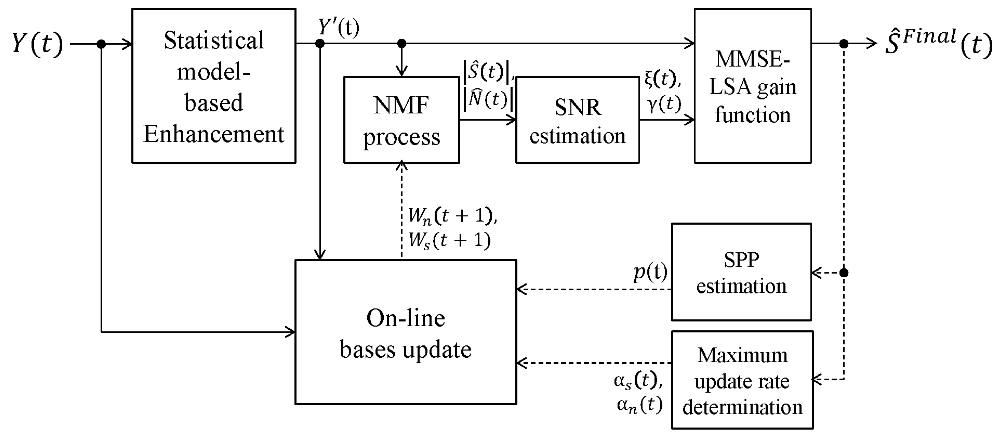


Figure 4.2: Block diagram of the proposed speech enhancement method.

consists of NMF-based noise reduction with the on-line update of both speech and noise bases. The overall block diagram of the proposed technique is illustrated in Figure 1., in which all the blocks run on-line. The first stage performs SE, which produces a pre-enhanced signal $Y'(t)$ with better SNR for the second stage. The second stage implements an NMF-based enhancement module introduced in Section III for which the KL divergence and MMSE-LSA estimator [24] are adopted for the distance metric of NMF analysis and spectral gain function, respectively, and $V(t)$, the input to the NMF analysis, is given by the magnitude spectra of the SE output, i.e., $V(t) = |Y'(t)|$. Both the speech and noise bases are updated with the help of SPP to cope with speech and noises unseen during training. The SPP is computed from the final enhanced speech in the same way as in SE, and the bases updated at the t-th frame is used for the NMF analysis at frame *t+1*. Cascading additional SE or NMF-based enhancement stages did not give any performance improvement over the current structure, maybe because they do not utilize any new characteristics of the signals.

4.3.1 On-line update of speech and noise bases

At each frame, once we obtain the clean speech estimate, we update the speech and noise bases for the NMF analysis in the following frame. This on-line bases update makes it possible to deal with the speech and noise variations that cannot be covered by the training DB and is considered a promising way to cope with the non-stationary nature of the signal.

Even though the first stage output $Y'(t)$ is fed as an input to the second stage, speech bases update using solely $Y'(t)$ may mislead the whole enhancement procedure. This is because $Y'(t)$ possesses distorted speech components, which is unavoidable in most of the statistical model-based speech enhancement techniques. For this reason, we update the speech and noise bases so that they can represent not only $|Y'(t)|$ but also $|Y(t)|$ better. Therefore, the data set $\tilde{V}(t)$ used for the on-line bases update is constructed by concatenating $|Y(t)|$ and $|Y'(t)|$, which showed better performance than using $|Y(t)|$ or $|Y'(t)|$ only, or using $|Y(t)|$ and $|Y'(t)|$ for the updates of W_s and W_n, respectively. $\tilde{V}(t)$ is an $M \times 2$ matrix while the number of basis vectors to be updated is $r_s + r_n$ which is much larger than 2. This turns out to be a severely underdetermined condition and we cannot anticipate a reasonable bases update. For this reason, we add a regularity term to the original objective function as follows:

$$f(\tilde{V}(t), \tilde{W}(t), \tilde{H}(t)) = \\ D_{KL}(\tilde{V}(t), \tilde{W}(t)\tilde{H}(t)) + \lambda \|W(t) - \tilde{W}(t)\|^2 \tag{4.6}$$

where $\tilde{H}(t) \in \mathbb{R}^{(R_s+R_n)\times 2}$ and $\tilde{W}(t) = [\tilde{W}_s(t)\ \tilde{W}_n(t)] \in \mathbb{R}^{M\times(R_s+R_n)}$ are the encoding and basis matrices for $\tilde{V}(t)$, respectively, $W(t) = [W_s(t)\ W_n(t)]$ denotes the basis matrix used to analyzed the t-th frame in the second stage, $D_{KL}(a, b)$ denotes the

KL divergence between a and b, and $\lambda > 0$ is a constant which controls the trade-off between the reconstruction error and the deviation from the previous basis matrix. The iterative update rule is derived in a similar manner to the original NMF update rule [9] with a slight modification due to the regularity term as follows:

$$\tilde{H}(t) \leftarrow \tilde{H}(t) \otimes \frac{\tilde{W}^T(t)\frac{\tilde{V}(t)}{\tilde{W}(t)\tilde{H}(t)}}{\tilde{W}^T(t)\mathbf{1}}, \tag{4.7}$$

$$\tilde{W}(t) \leftarrow \tilde{W}(t) \otimes \frac{\frac{\tilde{V}(t)}{\tilde{W}(t)\tilde{H}(t)}\tilde{H}^T(t) + \lambda W(t)}{\mathbf{1}\tilde{H}^T(t) + \lambda\tilde{W}(t)}$$

where $\tilde{H}(t)$ is randomly initialized and $\tilde{W}(t)$ is initialized by $W(t)$. Although the convergence is not proven like the update rule of the original NMF [56], [58], a continual descent of the objective function was observed in the experiments. The stopping criterion was the same as Section II.

The basis matrix $\tilde{W}(t)$ and the corresponding encoding matrix $\tilde{H}(t)$ obtained after the iterative update can represent the data set $\tilde{V}(t) = [|Y(t)|\ |Y'(t)|]$ well. However this does not mean that $\tilde{W}(t) = [\tilde{W}_s(t)\ \tilde{W}_n(t)]$ is a better estimate for the basis matrix than $W(t)$. Particularly when the speech component is very weak or absent in the current frame, $\tilde{W}_s(t)$ cannot be considered a good estimate for the speech basis matrix.

In order to alleviate this difficulty, we use the SPP $\boldsymbol{p}(t) \in \mathbb{R}^{M\times 1}$, each element of which represents the probability of speech activity in a specific frequency bin, to control the update of speech and noise bases. In the implementation, $\boldsymbol{p}(t)$ is estimated from the final enhanced speech spectra in a similar manner to [27]. In this approach, the speech and noise basis matrices which will be used for the next frame are determined as follows:

$$W_s(t+1) = \boldsymbol{\lambda}_s(t) \otimes \tilde{W}_s(t) + (\mathbf{1}_{M\times r_s} - \boldsymbol{\lambda}_s(t)) \otimes W_s(t), \tag{4.8}$$

$$\boldsymbol{\lambda}_s(t) = \alpha_s(t)\boldsymbol{p}(t)\mathbf{1}_{r_s},$$

$$W_n(t+1) = \boldsymbol{\lambda}_n(t) \otimes \tilde{W}_n(t) + (\mathbf{1}_{M\times r_n} - \boldsymbol{\lambda}_n(t)) \otimes W_n(t), \tag{4.9}$$

$$\boldsymbol{\lambda}_n(t) = \alpha_n(t)(\mathbf{1}_{M\times r_n} - \boldsymbol{p}(t)\mathbf{1}_{r_n})$$

where $0 < \alpha_s(t) < 1$ and $0 < \alpha_n(t) < 1$ are the maximum update rates for W_s and W_n, and $\mathbf{1}_{M\times R_s} \in \mathbb{R}^{M\times R_s}$, $\mathbf{1}_{M\times R_n} \in \mathbb{R}^{M\times R_n}$, $\mathbf{1}_{R_s} \in \mathbb{R}^{1\times R_s}$ and $\mathbf{1}_{R_n} \in \mathbb{R}^{1\times R_n}$ are all-one matrices. This interpolation enables a robust update of the speech and noise basis matrices leading to a stable speech enhancement performance.

4.3.2 Determining maximum update rates

It is clear that $W(t)$ needs to be updated quickly when it does not match the actual speech and noise and vice versa. To achieve this goal, the maximum rates of the on-line bases update, $\alpha_s(t)$ and $\alpha_n(t)$, should be adaptively determined. In the proposed algorithm, the normalized reconstruction error is used to determine $\alpha_s(t)$ and $\alpha_n(t)$. The normalized reconstruction error is defined as

$$e(t) = \frac{\sum_{m=1}^{M}[(V(t))_m - (W(t)H(t))_m]^2}{\sum_{m=1}^{M}(V(t))_m^2} \tag{4.10}$$

where $H(t)$ is obtained from the NMF-based enhancement stage. Since, however, $e(t)$ fluctuates too much from frame to frame, it should be smoothed such that

$$\tilde{e}(t) = \tau_e \tilde{e}(t-1) + (1-\tau_e)e(t) \tag{4.11}$$

where τ_e and $\tilde{e}(t)$ are a smoothing constant and the smoothed reconstruction error, respectively. The maximum rates of speech and noise bases update, $\alpha_s(t)$ and $\alpha_n(t)$, are now determined as a non-decreasing function of $\tilde{e}(t)$ as

$$\begin{aligned} \alpha_s(t) &= \max[sigm(\tilde{e}(t))\alpha_s^{max}, 0.01], \\ \alpha_n(t) &= \max[sigm(\tilde{e}(t))\alpha_n^{max}, 0.01] \\ sigm(x) &= \frac{x}{\sqrt{1+x^2}}, \end{aligned} \tag{4.12}$$

where α_s^{max} and α_n^{max} are the upper limits of the update rates.

4.4 Experiment result

In order to evaluate the performance of the proposed approach, we performed a series of speech enhancement experiments. For the first stage SE module, we employed the algorithm presented in [27], which provides not only the enhanced spectra but also the SPP estimate at each frame.

Speech and noise materials were selected from TIMIT [74] and NOISEX-92 [75] DBs, respectively, and the sampling rate was 16 kHz. A 512 point fast Fourier transform with 75% overlap was used. Each noise basis matrix was trained from 15 s-long noise signal which was not included in the test DB. Speech basis matrix was trained with 78 s-long clean speech spoken by 26 speakers. The test data set comprised utterances spoken by 16 speakers. The number of speech and noise basis vectors was 40 each ($r_n = r_s = 40$). The parameter values related to the on-line bases update and the smoothing were $\lambda = 0.001$, $\alpha_s^{max} = 0.3, \alpha_n^{max} = 0.4$, $\tau_s = 0.5$, $\tau_n = 0.9$ and $\tau_e = 0.98$. The performance was not sensitive to these parameter values.

The performance was measured in terms of the ITU-T Recommendation P.862 Perceptual evaluation of speech quality (PESQ) [43] score. We compared the per-

Table 4.1: PESQ scores for various noises with matched noise basis.

Noise Type	*F-16*	*factory2*	*M109*	*Leopard*	Average
unprocessed	1.9500	1.8235	1.9529	1.7835	1.8775
SE	2.3447	2.2473	2.3101	2.0218	2.2309
NMF	2.3844	2.1268	2.2938	2.3209	2.2815
SE+NMF	2.5104	2.3774	2.6174	2.6313	2.5341
SE+NMF+OU	2.5080	2.4229	2.6255	2.7204	**2.5692**

formance of the following four methods:

- *SE*: Only SE [27] was applied.
- *NMF*: Only NMF based-enhancement was used without the on-line bases update.
- *SE+NMF*: The cascaded form of *SE* and *NMF* without the on-line bases update was used. The noise basis matrix was trained from the noise part of SE output.
- *SE+NMF+OU*: *SE+NMF* with the on-line bases update was used.

The processing time of the proposed algorithm was about 1.32 times of *NMF* when both were implemented by matlab. This shows that the on-line bases update does not require a heavy computation.

The experiments were conducted in three different conditions:

Table 4.2: PESQ scores for various noises for which noise basis was trained with white noise.

Noise Type	*F-16*	*factory2*	*M109*	*Leopard*	Average
SE	2.3447	2.2473	2.3101	2.0218	2.2309
NMF	2.1142	1.8872	2.0115	1.7932	1.9515
SE+NMF	2.4180	2.2567	2.2988	2.0271	2.2502
SE+NMF+OU	2.5497	2.3924	2.6018	2.6122	**2.5390**

1. stationary noise environment with matched noise basis

2. stationary noise environment with mismatched noise basis

3. non-stationary and stationary noises environment with mismatched noise basis

Matched noise basis means that the types of the noise for training and test data are the same. On the contrary, mismatched noise basis was derived from the noise DB different from the actual noise of test DB. In the experiments with mismatched noise basis, we trained the noise basis based on the *white* noise.

Table 4.1 shows the PESQ scores obtained with rather stationary noises such as *factory2*, *F-16 cockpit (F-16)*, *M109* and *Leopard* noises when the tested noise type was included in the training DB. The SNR for each noise type was set to provide similar PESQ score for the unprocessed signals, which ranged from -5 to 5 dB. From the result, we can see that *SE+NMF+OU* outperformed other enhancement techniques

Table 4.3: PESQ scores for various noises mixed with non-stationary machinegun noise at 0 dB SNR with mismatched noise basis.

Noise Type (+*machinegun*)	*F-16*	*factory2*	*M109*	*Leopard*	Average
unprocessed	1.6890	1.6277	1.7584	1.6875	1.6906
SE	1.7474	1.7752	1.9317	1.8943	1.8372
NMF	1.7894	1.7752	1.9317	1.7004	1.7428
SE+NMF	1.8165	1.8205	1.9571	1.8962	1.8726
SE+NMF+OU	2.1746	2.3098	2.4540	2.4999	**2.3596**

Table 4.2 shows the PESQ scores obtained in the same noise environments to Table 4.1, but this time W_n was trained with *white* noise DB only. The performance of *NMF* was poor since W_n did not match the actual noise. It is apparent that the on-line bases update gave rise to a significant improvement on NMF performance.

The PESQ scores obtained when the stationary noises that were used in the previous experiments were mixed with the non-stationary *machinegun* noise are presented in Table 4.3. It is evident from the results that the proposed approach could also deal with non-stationary noise well.

Chapter 5

Discriminative NMF

5.1 Introduction

NMF shows a good performance when the audio signal which is formed of single category of source is represented or reconstructed. However, when the audio signal is corrupted with noise or mixed with the other category of source, the performance of reconstruction or separation is severely degraded, mainly due to the optimal solution of source separation and objective function of NMF algorithm is different when two or more of audio sources are mixed. In case of the application for speech enhancement, the amount that the speech basis vectors used for the reconstruction of noise source is the same to the speech distortion, and the amount that the noise basis vectors used for the reconstruction of speech signal means the residual noise.

In order to solve this issue, many papers named discriminative NMF (DNMF) have been published [6]- [8], [76], [77]. Although the detailed methods are different to each other, these works of DNMF have the same aim to make the basis vectors of a target source that reconstruct only the target source and utilize the other source

DBs or DBs mixed with a target source DB. In [6], the basis vectors of target source are made with the constraint that orthogonal to the basis vectors of the other source. However, the above constraint can make the negative effect such as a high reconstruction error because of the audio nature. In [7], the basis vectors of the each source are updated by the reconstruction error of the each source, and the encoding vectors are updated by the whole reconstruction error, alternatively. In [8], the clean source and the same clean source mixed with the other source are used for the training phase.

In this chapter, we propose discriminative NMF that makes the higher reconstruction error to the other source than standard NMF does. Namely, the proposed DNMF gives the constraint to have a high reconstruction error of the other source, not target source. There are many candidates for basis vectors, because of non-convexity of NMF. The proposed DNMF finds a proper basis matrix that makes a high reconstruction error to the other source from the above candidates. In this chapter, the application for the performance evaluation is speech enhancement, and perceptual evaluation speech quality (PESQ) [43] and signal-to-distortion (SDR) [42] are used for the measurement.

5.2 Discriminative NMF utilizing cross reconstruction error

In the source separation, some of the basis vectors from the specific source are used for the reconstruction of the other source because of the nature of the audio signal. Namely, one frame data of the audio signal can have a faint resemblance with the audio signal of other category, and this similarity appears in the between each basis vectors. Because of this similarity, the basis vectors obtained independently to the other source make a misuse to the representation of the other source during the reconstruction of the mixed signal. That is, the regions made from the linear combination of basis vectors of each source are overlapped. Consequently, this misuse and overlapped regions of basis vectors make a degradation on the performance of separation.

Since NMF does not satisfy convexity, the specific source has many candidate bases from NMF algorithm, and the basis matrix can be adjusted and fixed by the initial value and the constraint or penalty function [6]- [8], [10], [70]. Among these candidate bases, some bases are apt to be consumed for the reconstruction of the other source, but some bases are used less than other bases. Surely, the latter bases show better performance than the former bases does in the source separation. In order to obtain desired bases, the latter basis, certain constraint or penalty function is applied to the objective function of NMF. Although the reconstruction error of target source can be increase because of the constraint function, if the overlapped range and the misuse of each basis matrix are reduced, the higher performance can be expected than before.

5.2.1 DNMF using the reconstruction error of the other source

For such a constraint function, we propose DNMF which utilizes the reconstruction error of the background source data. In the case of speech enhancement, the objective function of proposed method is defined by

$$
\begin{aligned}
&f(W_S, H_S, H_N) = \\
&\qquad D(V_S \parallel W_S, H_S) - \lambda D(V_N \parallel W_S, H_N)
\end{aligned}
\tag{5.1}
$$

where V_S, W_S, and H_S are input data, basis, and encoding matrix for the speech signal, and V_N and H_N are input data and encoding matrix for the background noise signal. The λ plays a role of the trade-off between the reconstruction error of the speech and noise signals. If $\lambda = 0$, (5.1) is the same to the standard NMF, on the other hand, when $\lambda > 0$, W_S becomes discriminative basis matrix to the representation of the noise signal.

The update equations of W_S, H_S and H_N are obtained by the same way to Sec. 2. as follows:

$$H_S \leftarrow H_S \otimes \frac{W_S^T \frac{V_S}{W_S H_S}}{W_S^T \mathbf{1}}, \quad H_N \leftarrow H_N \otimes \frac{W_S^T \frac{V_N}{W_S H_N}}{W_S^T \mathbf{1}}, \tag{5.2}$$

$$W_S \leftarrow W_S \otimes \frac{\frac{V_S}{W_S H_S} H_S^T}{H_S^T \mathbf{1} - \lambda(\frac{V_N}{W_S H_N} H_N^T)}. \tag{5.3}$$

In the training phase, the speech basis matrix of DNMF,(5.3), is obtained from the speech and noise samples by (5.2) and (5.3) with the fixed number of iteration and the initialization of non-negative random values.

In the speech enhancement phase, any noise basis matrix can be applied with discriminative speech basis matrix, (5.3), and the speech STFT coefficients are finally estimated as Sec. 2.

5.2.2 DNMF using the interference factors

For the comparison, we propose another DNMF which constraint function consists of interference factor. When speech signal is reconstructed by both speech and noise bases, the reconstruction part from noise bases becomes the speech distortion, meanwhile, the reconstruction part from speech bases when noise signal is reconstructed denotes the residual noise. These parts should be reduced, and for this issue, the objective function is defined as:

$$f(W,H) = D(V \parallel W,H) + \gamma \sum (W_S H_{SN} + W_N H_{NS}), \tag{5.4}$$

$$H = \begin{pmatrix} H_{SS} & H_{SN} \\ H_{NS} & H_{NN} \end{pmatrix}$$

where $W = [W_S, W_N]$ and $V = [V_S, V_N]$ are the cascade of speech and noise part, V_S and V_N indicate clean speech and noise DBs. In (5.4), the second term denotes interference factors, the speech distortion and residual noise. The bigger γ is, the more discriminative bases are obtained, but the performance of representation degrades. The update equations of each basis and encoding matrices as follows:

$$H \leftarrow H \otimes \frac{W^T \frac{V}{WH}}{W^T \mathbf{1} + \gamma C}, \tag{5.5}$$

$$C = \begin{pmatrix} \mathbf{0} & W_S \mathbf{1} \\ W_N \mathbf{1} & \mathbf{0} \end{pmatrix},$$

$$W \leftarrow W \otimes \frac{\frac{V}{WH} H_S^T}{H^T \mathbf{1} + \gamma B}, \tag{5.6}$$

$$B = \begin{pmatrix} H_{SN} \mathbf{1} & H_{NS} \mathbf{1} \end{pmatrix}$$

where $\mathbf{0}$ is a matrix of suitable size with all elements equal to zero. Unlike the Sec. 3.1., the speech and noise basis matrices are obtained at the same time, and both these basis matrices should be applied to the speech enhancement simultaneously, since these are a pair of basis matrices.

5.3 Experiment result

Speech and noise samples were selected from TIMIT and NOISEX-92 DBs, respectively, with a sampling rate of 16 kHz. A 512-point discrete Fourier transform with 75% overlap was used. The basis matrices for each noise type were obtained from a half of sample, about 120-second long noise signal, and the speech DB for training consists of 56-sample files different from test set. The speech test data set consisted of 48 sentences uttered by 24 different speakers. We applied the proposed basis matrix on 4 different types of noise signals including *F-16*, *factory1*, *babble* and *machinegun*. The number of speech and noise bases is the same in the experiment, and the number of basis vectors was either 64 or 128.

The performance of the proposed method was measured using the perceptual evaluation of speech quality (PESQ) [43] and signal-to-distortion ratio (SDR) [42]. In order to evaluate the proposed method, the following methods are performed with the same enhancement system and dataset, i.e., only the basis matrix is different, and we selected and compared the algorithm [6], which makes basis matrices orthogonal each other.

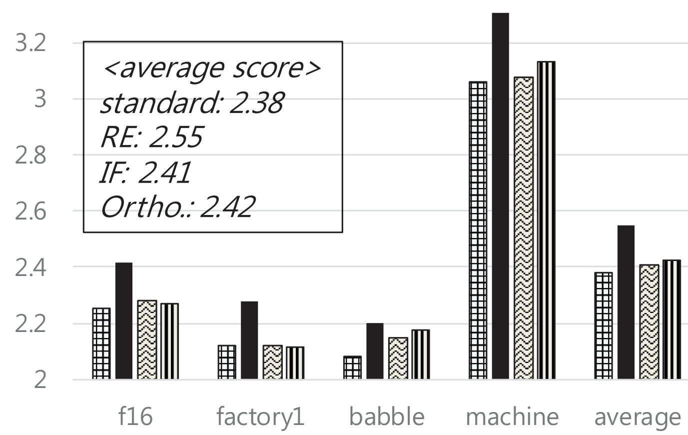
•*standard*: without any constraint function [9]

•*RE*: the proposed method using the reconstruction error of the other source (Sec. 3.1.)

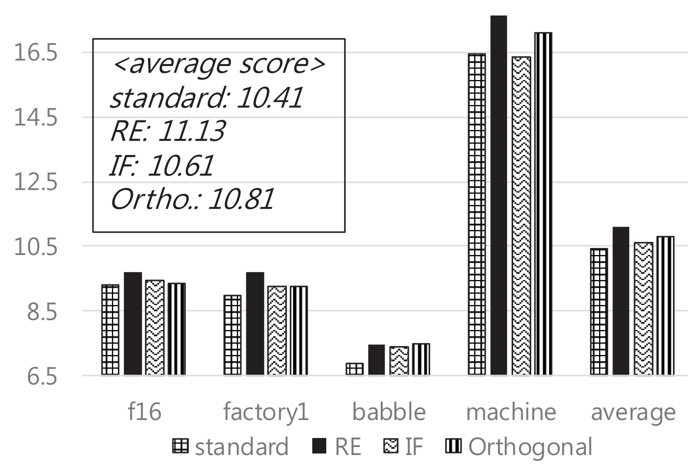
•*IF*: the proposed method using the interference factors (Sec. 3.2.)

•*Ortho.*: using the basis matrix of [6]

Fig.5.1 shows SDR and the PESQ score obtained from the noisy signal (input SNR 5dB) where the number of speech basis vectors is 64. One of the proposed method, *RE*, outperformed the others. This observation indicates that using the

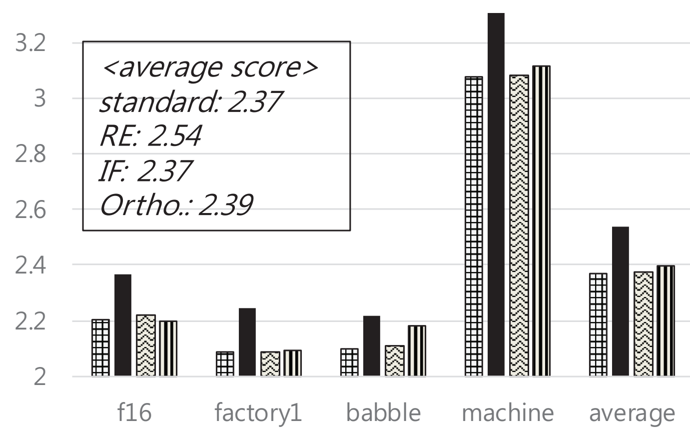


(a) PESQ score

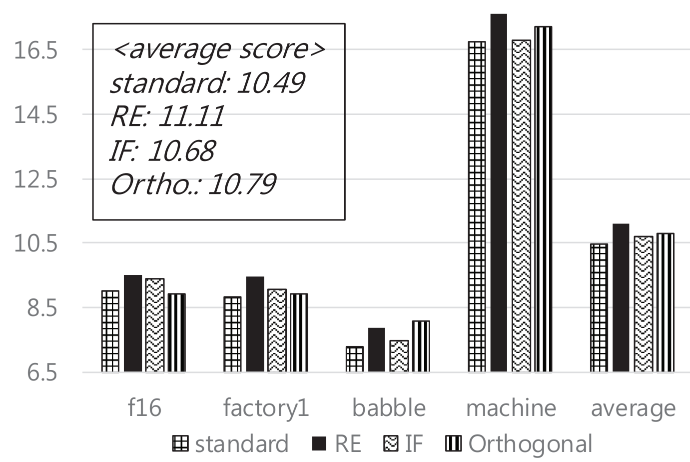


(b) SDR

Figure 5.1: Experiment result 1 ($r_s = 64$): input SNRs of test dataset is 5 dB.



(a) PESQ score



(b) SDR

Figure 5.2: Experiment result 2 ($r_s = 128$): the experimental condition is the same to Fig.5.1 without r_s (input SNRs of test dataset is 5 dB.).

reconstruction error of the other source is effective. The emphasis of the result is the PESQ score of *RE*. The order of performance improvement differs by type of the background noise, for example, *Ortho.* shows better performance than the other methods in the case of *babble* noise type.

Fig.5.2 shows the performance when the number of speech basis vectors is 128. As can be seen by this figure, slightly worse performance is obtained for all the methods in this case. Again, *RE* yielded the best performance, outperforming the *standard* by around 0.17 PESQ score and 0.62 SDR in average.

Chapter 6

Incremental approach for bases estimate

6.1 Introduction

Though NMF shows an impressive performance in several fields, one of its weakness is that the final result is so sensitive to the initial values of the bases [13]. Because the specified objective function of NMF is not convex, the optimized solution obtained from iterative update of the basis matrix can be stuck to a local minimum, which implies that the overall performance may significantly depend on the initial parameter values. For this reason, several previous works attempt to provide systematic ways to initialize the basis and encoding matrices such as the centroids of k-means clustering and singular value decomposition (SVD)-based method [63]-[69]. Though some of these methods show a lower reconstruction error and a faster convergence speed than the random value initialization, they do not carefully consider the performance in source separation. Moreover, the SVD-based methods can not support over-complete bases in which the number of bases is larger than the dimension of the input vector.

The conventional vector quantization task can be interpreted as a special case of the matrix factorization where each basis vector corresponds to a codeword and only a single basis is activated at each time [79]. This analogy implies that the data clustering techniques can provide some useful cues for the initialization of the NMF bases. Unfortunately, however, conventional codebook training approaches such as the k-means clustering can only guarantee suboptimal solutions similar to the case of NMF bases estimation and the final centroids are sensitive to the initialization of the code vectors. In order to alleviate this difficulty, several modified k-means algorithms have been developed [48], [80], [81]. The core idea of these algorithms is to increase the number of code vectors gradually while optimizing a certain criterion so that the final result can be less dependent on the initial parameter values.

In this chapter, we propose a novel approach to estimate the basis and encoding matrices for the NMF analysis. Exploiting the analogy between NMF analysis and data clustering, a systematic method for estimating the NMF basis matrix is proposed by combining the standard NMF basis training procedure and an efficient codebook learning algorithm. The proposed method borrows an idea from the modified k-means algorithm [48]. One of the prominent features of this algorithm is that it estimates the parameters incrementally, i.e. increases the number of bases at each iteration. In order to evaluate the performance of the proposed technique, we carried out an experiment on target source separation. In the experimental result, we can see that the proposed method outperformed the other bases initialization methods.

6.2 Incremental approach based on modified k-means clustering and Linde-Buzo-Gray algorithm

6.2.1 Based on modified k-means clustering

In this section, we propose a novel approach to estimate NMF bases, which is based on an analogy between the NMF basis training and the codebook design in vector quantization. If the encoding vector of the NMF analysis is allowed to have only one non-zero component, then each NMF basis can be viewed as a codeword vector and the reconstruction error can be treated as the distance between the input vector and its nearest codeword. Our approach to NMF bases estimation is motivated by the global k-means (GKM) clustering technique [48], which has demonstrated smaller clustering error than several other variants of the k-means clustering approach. In general, for data clustering, we need to find R codewords and a rule to map any M-dimensional input vector into one of the R codewords for the sake of minimizing the sum of the squared Euclidean distances between each input vector and the corresponding code vector. GKM starts with one cluster ($R = 1$) for which the optimal codeword is set at the centroid of the whole data set. At each iteration of the GKM algorithm, a new codeword is added to refine the clusters and the conventional k-means algorithm is run until convergence.

In the proposed approach, the NMF bases are trained incrementally like the GKM technique while the iterative update rules in (2.16) and (2.17) are adopted instead of the k-means clustering operation. Fig. 6.1 illustrates how a new basis is determined in the proposed approach when $M = 3$. The simplex shown in Fig. 6.1 represents the space of all possible basis vectors since each column of $\mathbf{W}$ is constrained to have unit L_1-norm. A set of bases $\mathbf{W}$ forms a convex hull on his

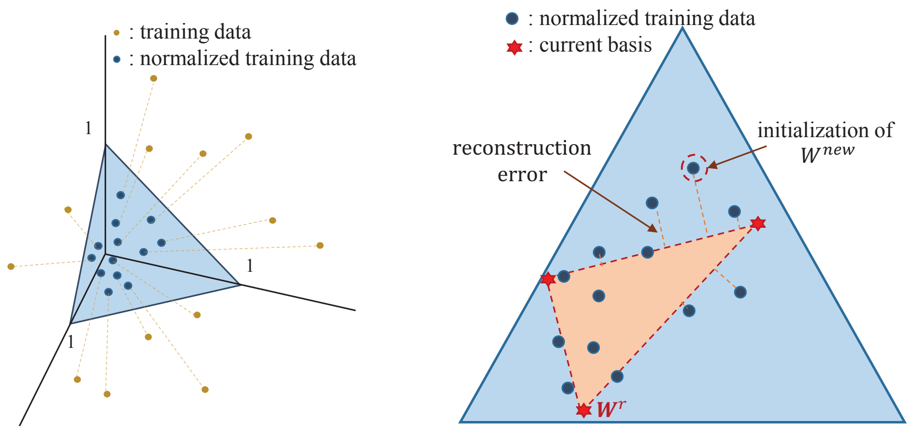


Figure 6.1: The data point for the new basis when $M = 3$ (left: data and pullback onto the simplex, right: data on the simplex).

simplex. If a vector, when projected onto the simplex i.e., normalized to have unit norm, is far from this convex hull, it means that **W** produces a large reconstruction error for this vector in NMF analysis. In the proposed algorithm, the training data which shows the maximum reconstruction error is appended to the current basis matrix **W** as a new basis vector. This initializes the basis matrix where the number of columns is incremented by one and then the conventional update algorithm in (2.16) and (2.17) is iterated until convergence.

The pseudo code of the proposed incremental approach to NMF basis estimation is given in Fig. 6.2. The input of the algorithm is the training data matrix $\mathbf{V} \in \mathbb{R}_+{}^{M \times N}$ and the number of bases R, while the output is the trained basis matrix $\mathbf{W}$. Let $\mathbf{W}^r$ and $\mathbf{H}^r$ respectively denote the basis matrix and the corresponding encoding matrix when the number of bases is r. We begin with $r = 1$ and increments it at each iteration until $r = R$. The proposed algorithm proceeds as follows:

1) Initialization $r = 1$. Compute the centroid $c(\mathbf{V})$ of $\mathbf{V}$, the entire training set. The centroid is computed as the mean of the whole training data vectors. Then, take the L_1-based normalization of $c(\mathbf{V})$, i.e.

$$\mathbf{W}^1 = \frac{c(\mathbf{V})}{\parallel c(\mathbf{V}) \parallel_1 \mathbf{1}} \tag{6.1}$$

where $\| \cdot \|_1$ and $\mathbf{1}$ represent L_1-norm and a vector with all elements equal to one, respectively.

2) Compute $\mathbf{H}^1$ which minimizes the reconstruction error. It is easy to find $\mathbf{H}^1 = \mathbf{W}^{1^T}\mathbf{V}$.

3) Let V_{n^*} be the training vector showing the maximum reconstruction error. V_{n^*} is decided as

$$V_{n^*} = \underset{V_n}{\operatorname{argmax}} \|V_n - \mathbf{W}^r\mathbf{H}^r(n)\|_2 \tag{6.2}$$

where $\mathbf{H}^r(n)$ denotes the n-th column of $\mathbf{H}^r$, and $\| \cdot \|_2$ represents L_2-norm.

4) A new basis vector $W^{new} \in {\mathbb{R}_+}^{M\times 1}$ and the corresponding encoding vector $H^{new} \in {\mathbb{R}_+}^{1\times N}$ are determined as

$$W^{new} = \frac{V_{n^*}}{\parallel V_{n^*} \parallel_1 \mathbf{1}}, \qquad H^{new} = \parallel V_{n^*} \parallel_1 \mathbf{e}_{n^*}^T \tag{6.3}$$

where $\mathbf{e}_{n^*}$ denotes the standard vector having all its elements zero except for the n^*-th element which is set to 1.

5) Increment the number of bases by one for which

$$\mathbf{W}^{r+1} = [\mathbf{W}^r, W^{new}] \in {\mathbb{R}_+}^{M\times(r+1)}, \tag{6.4}$$

$$\mathbf{H}^{r+1} = [\bar{\mathbf{H}}^{rT}, H^{newT}]^T \in {\mathbb{R}_+}^{(r+1)\times N}$$

where $\bar{\mathbf{H}}^r$ is the same as $\mathbf{H}^r$ except the n^*-th column is replaced by a zero vector, which makes the perfect reconstruction of V_{n^*}.

6) Iteratively update $\mathbf{W}^{r+1}$ and $\mathbf{H}^{r+1}$ according to (2.16) and (2.17) for a fixed number of times to obtain $M \times (r+1)$ basis matrix.

7) The steps from 3) to 6) are repeated until we get R bases.

6.2.2 LBG based incremental approach

Because the general objective function of NMF is biconvex in $\mathbf{W}$ and $\mathbf{H}$, different algorithms and their initializations lead to different solutions. It means that NMF algorithm does not satisfy the uniqueness [13]. To get more accurate solutions with complex models, carefully designed initializations or regularizations may be needed. Since, however, it is hard to exploit a general prior knowledge of the parts of a source data, nonnegative random values have been widely applied for NMF initialization. Although this method has shown a somewhat proper performance experimentally [9], [66], due to the non-covexity of the objective function and iterative nature of the algorithm, it cannot be considered to provide an optimal initial point for successful NMF analysis.

In this section, we propose novel approach to estimate the basis for NMF analysis which is based on the clustering approach. The proposed method is motivated by accepting a general premise that the best basis is the centroid of the whole training DB when the number of bases is set to one. The core idea of the proposed approach is to estimate the bases incrementally. The incremental approach where a new centroid is searched at each step can be a good strategy for basis estimation.

LBG algorithm is the most cited and widely used algorithm on designing the vector quantization codebook [67], and it is similar to the k-means algorithm in data clustering. At each iteration of LBG, each vector is split into two new vectors. LBG algorithm can be employed in the making of the basis. In this case, its number of bases doubles in every procedure.

The pseudo code for the proposed incremental approach to the NMF basis estimation is given in Fig. 6.3. The input of the algorithm is the training data matrix $\mathbf{V} \in \mathbb{R}^{m \times n}$ and the integer k. The output is the matrix composed of 2^k basis vectors. $\mathbf{W}^i$ and $\mathbf{H}^i$ in Fig. 1 denote the basis matrix with 2^i bases and the corresponding encoding matrix, respectively. In order to decide the single basis case, $W^0 \in \mathbb{R}^{m \times 2^0}$ is obtained by the centroid of $\mathbf{V}$ with unit-norm normalization. $\| \cdot \|_1$, $\mathbf{1}$, and *centroid*$(\mathbf{V})$ represent unit-norm, a vector of a proper size with all elements equal to one, and the centroid of the matrix $\mathbf{V}$, respectively. For the encoding of W^0, H^0 which minimizes the Euclidean distance is given in a closed-form as $H^0 = \mathbf{V}^T W^0$. For the bases $\mathbf{W}^1 \in \mathbb{R}^{m \times 2^1}$, $[W^{0+}, W^{0-}]$ is applied for the initialization of *NMF process* where $W^{0+} \in \mathbb{R}^{m \times 2^0}$ and $W^{0-} \in \mathbb{R}^{m \times 2^0}$ are obtained by addition and subtraction of a very small value ϵ to W^0, respectively. *NMF process* indicates the alternative update phase, (1) and (2), for a fixed number of iterations. The initial value of $\mathbf{H}^1$ is given as $[(\mathbf{H}^0/2)^T, (\mathbf{H}^0/2)^T]^T \in \mathbb{R}^{2^1 \times n}$. The procedures from 3 to 8 are repeated until we get 2^k bases.

6.3 Experiment result

6.3.1 Modified k-means clustering based approach

To evaluate the performance of the proposed algorithm, experiments on audio source separation were performed in a variety of noisy conditions with the target source being speech signal. Speech and interfering signals were selected from TIMIT [74] and NOISEX-92 [75] DBs, respectively. A 512-point discrete Fourier transform with 75% overlap was used to form the spectrogram with hamming window. The sampling rate was 16kHz. The magnitude spectra were used as data vectors for the NMF analysis. The basis matrix for each noise type was obtained from about 120-second long noise signal, and the training speech DB was 130-second long spoken by 56 different speakers. The test speech data set consisted of 32 sentences from 32 different speakers. We tested 4 different types of noises including *F-16*, *factory1*, *babble*, and *white* noises. There was no overlap between the training and test data in both the speech and noise types. The numbers of the bases for each source varied among 64, 128, 256, and 512, and the number of iterations for the separation was 30.

The performance of the proposed approach was evaluated in terms of PESQ [43] and SDR [42]. To demonstrate the performance improvement achieved by the proposed method, four versions of the NMF-based source separation algorithm described in Section II for which only the training methods of the basis matrix differed were compared:

- **Rand**: NMF basis estimation with random initialization [9]
- **SVD**: NMF basis estimation which utilizes singular value decomposition [69]
- **Cent**: NMF basis estimation which utilizes the centroids of the k-means clustering [68]
- **Prop**: proposed incremental approach to the NMF basis estimation

To facilitate the robustness of the proposed algorithm to the outliers of the training DBs, the average of the data vectors corresponding to the top 10% of the largest reconstruction error was used for W^{new} instead of a single data vector with the highest reconstruction error. In **Cent**, the number of the clusters was kept the same to the number of the NMF bases. The iterative update rule used for the NMF analysis was the PGD with Euclidean distance [56] for training and the multiplicative update rule with KL [9] for separation, which showed high performance and low computation time in the experiment when **Rand** was performed. The number of iterations during the training phase was decided to maximize the separation performance for each algorithm as in [21], which turned out to be 100 for **Rand**, **SVD**, and **Cent** and integers in the range $[1, 3]$ for each stage of **Prop**.

Fig. 6.4 shows the source separation performance obtained while varying the number of the bases ($R = 64, 128, 256, 512$). When $R = 512$, the result of **SVD** is omitted because **SVD** cannot be applied to the over-complete basis case. **Rand** shows the optimal performance when $R = 256$, but its performance much degraded when $R = 512$. **Cent** showed a similar performance to **Rand**. **Prop** outperformed the other methods at any R. This results imply that the basis vectors obtained from the incremental approach are more discriminative than those of the other approaches.

The source separation performance for different noise types is given in Fig. 6.5. The values of **Rand**, **Cent**, and **Prop** were the average of the cases of $R = 64, 128, 256, 512$. Since **SVD** cannot make the over-complete basis matrix, the average of the cases of $R = 64, 128, 256$ was used for the performance comparison. **Prop** outperformed the other methods at any type of noise source, and the performance improvement over **Rand** were about 1.64 dB and 0.14 in terms of SDR and PESQ scores, respectively. In particular, **Prop** showed a high performance improvement when *white* noise was mixed, while **SVD** demonstrated a performance degradation.

6.3.2 LBG based approach

To evaluate the performance of the proposed algorithm, audio source separation was performed in a variety of noisy conditions, and the target sources were speech and violin signals. The whole data for the training and test was not overlapped. A 512-point discrete Fourier transform with 75% overlap was used to form the spectrogram with a sampling rate of 16 kHz ($m = 257$).

The performance of the proposed methods were evaluated in terms of PESQ [43] and SDR [42]. To demonstrate the performance improvement achieved by the proposed methods, four source separation systems for which only the basis matrices were trained in different ways were compared:

- *Rand.*: the initialization of nonnegative random values [9]
- *SVD*: the initialization in [69] which utilizes the result of singular value decomposition
- *Cent.*: the initialization in [63] which utilizes the centroids from the k-means clustering (The number of the cluster is the same to the number of the basis vectors.)
- *LBG*: the proposed method using the LBG-based basis estimation for NMF ($\epsilon = 1^{-14}$)

The whole bases of above systems were trained by PGD [56], on the other hand, the source separation was performed by multiplicative update rule with KLD [9]. This is because such a system experimentally shows a high performance and a low computational time of the separation when random nonnegative values are used for

NMF initialization. Each number of iteration during the training phase was decided based on the separation performance [21]. (*Rand.*=*SVD*=*Cent.*=100, *LBG*=10.)

The target source: speech signal

Speech and noise samples were selected from TIMIT [74] and NOISEX-92 [75] DBs, respectively. The basis matrix for each noise types was obtained from about 120-second long noise signal, and the speech DB for the training was 130-second long spoken by 56 different speakers. The speech test data set consisted of 32 sentences from 32 different speakers. We tested 4 different types of noises including *F-16*, *factory1*, *babble*, and *white* noises. The numbers of the bases were 64, 128, 256, and 512, and the number of iteration for the separation phase was 30.

Fig. 6.6 shows the PESQ scores and SDRs when the input signal-to-noise ratio (SNR) was 0 dB. For all cases of r, the proposed algorithm outperformed other methods in terms of both the PESQ score and the SDR. In particular, *LBG* produced the best separation performance, and the optimal r for the performance was different from the case of *Rand.*. In the over-complete case, $r = 512 > m$, the performance of *Rand.* was the lowest, but the proposed methods maintained the performance or outperformed the case of $r \leqq 256$. One of the previous method, *SVD*, showed the minimum reconstruction error during our training phase, but it cannot support the over-complete case and its separation performance was lower than *Rand.* This result denotes that the performance of source separation is not proportional to the reconstruction error during the training phase, and the proposed method which utilizes the incremental approach may extract the proper character of the source to the bases.

Fig. 6.7 denotes the source separation performance as the interfering sources, *F*-

Table 6.1: The information of the data for the bases estimation and source separation (resampled to 16kHz/s)

The phase	Title	Artist	Instrument
Training	Partita No.1 (BWV 1002) - Double	Ida Haendel	Violin
	Blind Film	Yiruma	Piano
Separation	Sonata No.2 (BWV 1003) - Allegro	Ida Haendel	Violin
	Waltz In C Minor (Only For Piano)	Yiruma	Piano

16, *factory1*, *babble*, and *white* signals. In the same manner as Fig. 2, the proposed method *LBG* outperformed other methods in the face of all interfering sources. The previous works, *SVD* and *Cent.*, show a similar performance as *Rand.*, but the performance degraded when *white* and *babble* signals are mixed, respectively. *LBG* made a performance improvement at every interfering sources, and it showed a prominent improvement in term of SDR.

The target source: violin signal

For violin and piano data, we used four songs to the training and separation phases, and Table. 6.1 shows the information of the data. The interfering sources were 4 different types of sources including *piano*, *factory1*, *babble*, and *machinegun* sources, and the test data set of violin consisted of 10 clips which are 5 seconds long each.

The number of bases r for each source was set to 128, which provided a good trade-off between the reconstruction error and the computational complexity. The experimental results when the input SNR is 0 are illustrated in Fig. 6.8. The proposed algorithms outperformed other methods at all interfering types. In particular, LBG shows a high performance improvement when *piano* or *factory1* source is mixed. The performance improvements of LBG were 1.43 and 2.05 in term of the SDR over *Rand.* and SVD, respectively.

Figure 6.2: Pseudo code for the proposed incremental approach to the NMF basis estimation.

Input: Matrix $\mathbf{V} = (V_1, \cdots, V_N) \in \mathbb{R}_+{}^{M\times N}$**, integer** R

Output: Matrix $\mathbf{W} \in \mathbb{R}_+{}^{M\times R}$

---

1. $\mathbf{W}^1 = \frac{c(\mathbf{V})}{\|c(\mathbf{V})\|_1 \mathbf{1}}$

2. $\mathbf{H}^1 = \mathbf{W}^{1^T}\mathbf{V}$

for $r = 1 : R - 1$

3. Initialize $\mathbf{W}^{r+1}$ and $\mathbf{H}^{r+1}$

   1) Find the vector V_{n^*} with the maximum reconstruction error in $\mathbf{V}$.

   2) $W^{new} = \frac{V_{n^*}}{\|V_{n^*}\|_1 \mathbf{1}}$

      $H^{new} = \| V_n \|_1 \mathbf{e}_{n^*}^T$

   3) $\mathbf{W}^{r+1} = [\mathbf{W}^r, W^{new}] \in \mathbb{R}_+{}^{M\times(r+1)}$,

      $\mathbf{H}^{r+1} = [\bar{\mathbf{H}}^{rT}, H^{newT}]^T \in \mathbb{R}_+{}^{(r+1)\times N}$

4. Iteratively update $\mathbf{W}^{r+1}$ and $\mathbf{H}^{r+1}$ with (2.16) and (2.17)

end

5. $\mathbf{W} = \mathbf{W}^R$

Figure 6.3: Pseudo code for the proposed incremental approach to the NMF basis estimation

---

Input: Matrix $\mathbf{V} = (V_1, V_2, \cdots, V_n) \in \mathbb{R}^{m \times n}$,

integer k

Output: Matrix $\mathbf{W} \in \mathbb{R}^{m \times 2^k}$

---

1. $W^0 = \frac{centroid(\mathbf{V})}{\|centroid(\mathbf{V})\|_1 \mathbf{1}}$

2. $H^0 = \mathbf{V}^T W^0$

for $i = 0 : k - 1$

3. $\mathbf{W}^{i+} = \mathbf{W}^i + \epsilon$, $\mathbf{W}^{i-} = \mathbf{W}^i - \epsilon$

5. $\mathbf{W}^{temp} = [\mathbf{W}^{i+}, \mathbf{W}^{i-}] \in \mathbb{R}^{m \times 2^{i+1}}$

6. $\mathbf{H}^{temp} = [(\mathbf{H}^i/2)^T, (\mathbf{H}^i/2)^T]^T \in \mathbb{R}^{2^{i+1} \times n}$

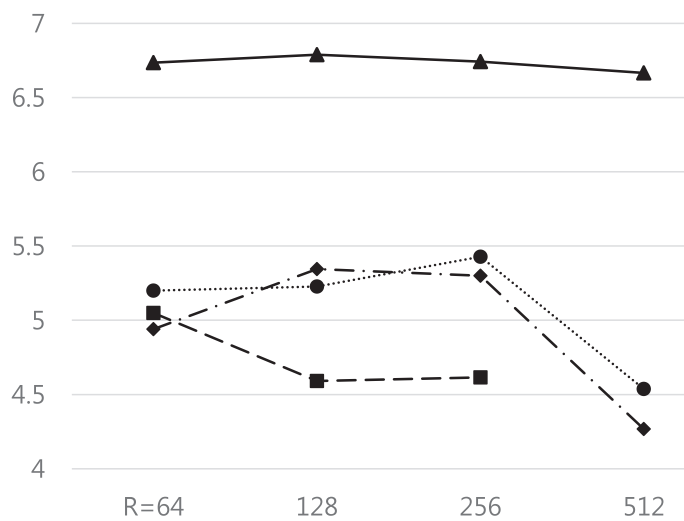
7. Do *NMF process* by $\mathbf{W}^{temp}$ and $\mathbf{H}^{temp}$

$\longrightarrow \mathbf{W}^{i+1}$, $\mathbf{H}^{i+1}$

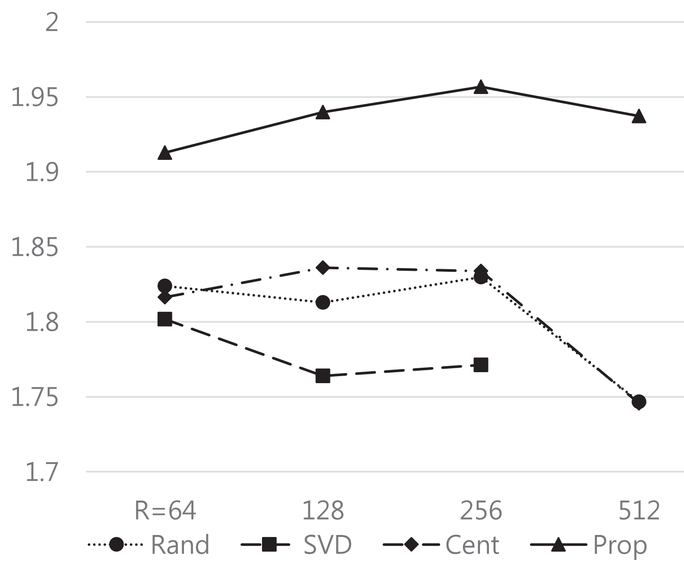
8. $\mathbf{W}^{i+1} = \frac{\mathbf{W}^{i+1}}{\|\mathbf{W}^{i+1}\|_1 \mathbf{1}}$

end

9. $\mathbf{W} = \mathbf{W}^k$

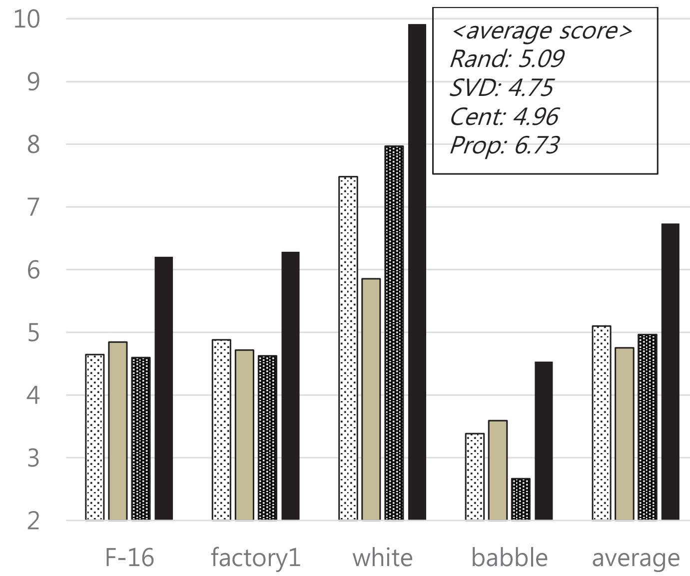


(a) SDRs

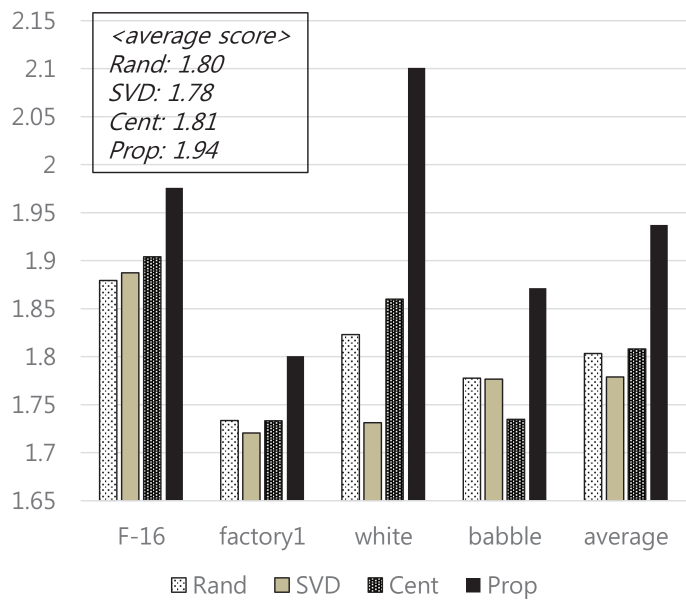


(b) PESQ scores

Figure 6.4: The source separation performance with various basis training methods according to the number of basis vectors. (input SNR = 0 dB)

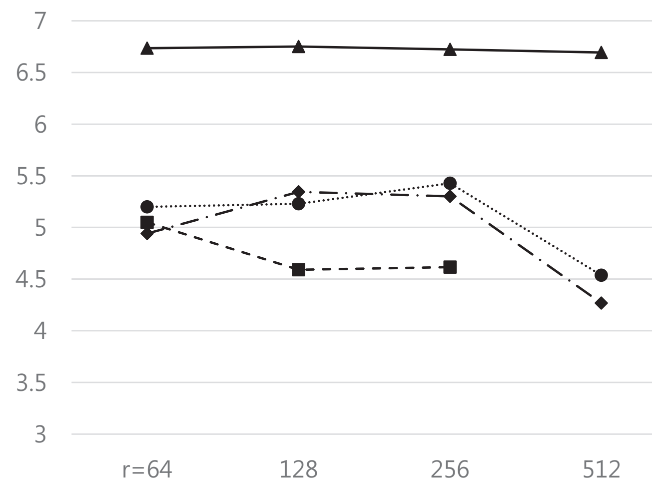


(a) SDRs

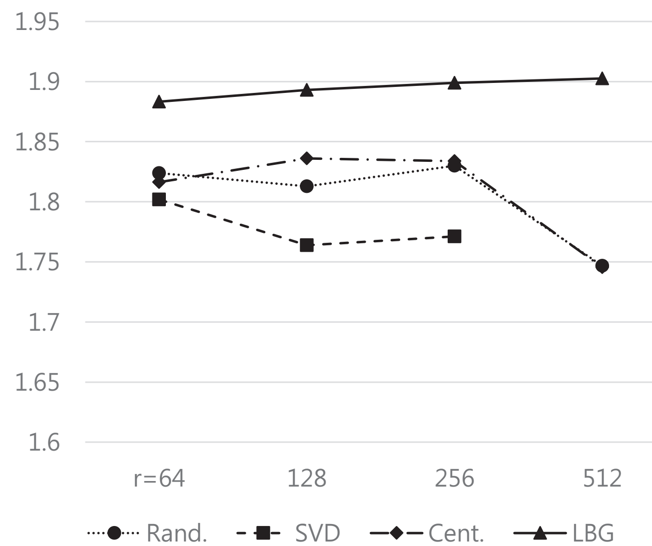


(b) PESQ scores

Figure 6.5: The source separation performance with various basis training methods according to the interfering source (input SNR = 0 dB)

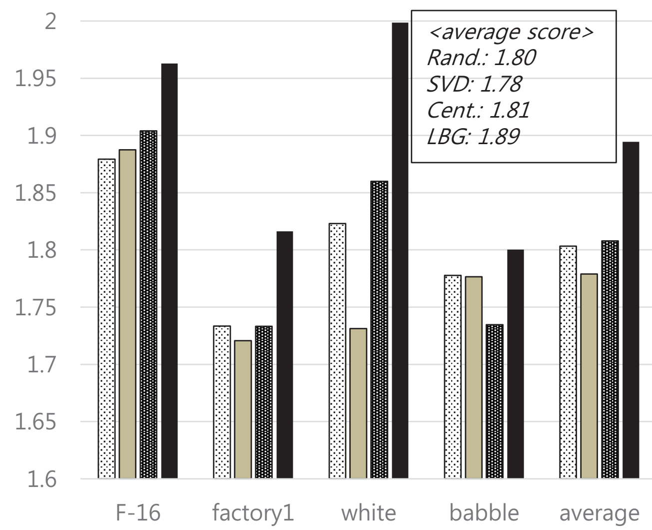


(a) SDRs

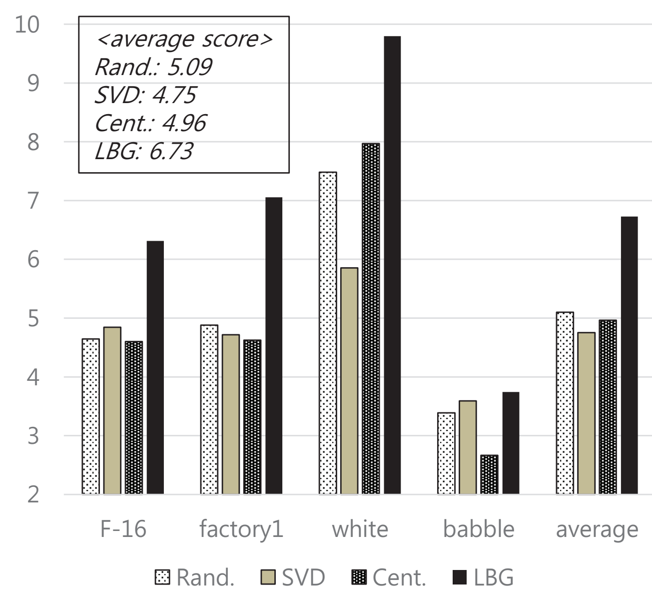


(b) PESQ scores

Figure 6.6: The source separation performances based on the numbers of basis (target source = speech, input SNR = 0 dB)



(a) PESQ scores



(b) SDRs

Figure 6.7: The source separation performances based on the interfering sources (target source = speech, input SNR = 0 dB)

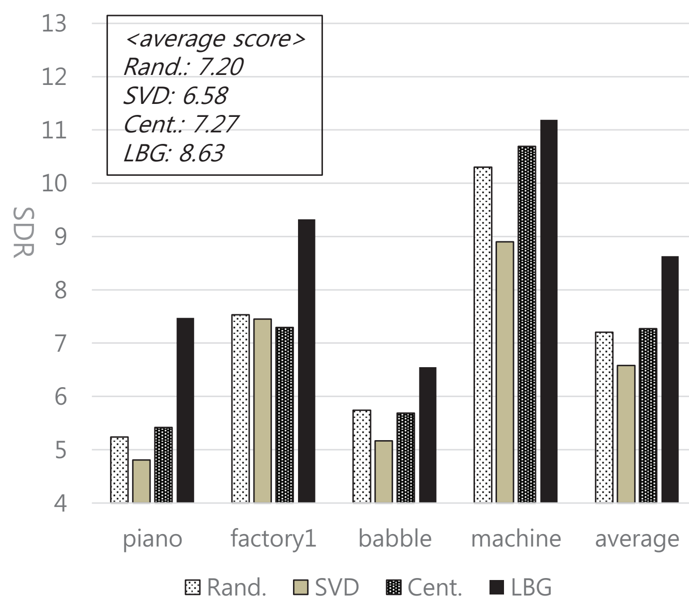


Figure 6.8: The source separation performances as the interfering sources (target source = violin, input SNR = 0 dB, r=128)

Chapter 7

Prior model of encoding vectors

7.1 Introduction

Most of the NMF-based source separation approaches compute the basis matrix $\mathbf{W}$ from a set of given training data and then use it for specific source separation with possible update of the basis matrix [14]- [21]. The encoding matrix obtained for the training data, $\mathbf{H}^{train}$, is usually ignored after training although it bears useful information on how often each basis was utilized. In [10], the distribution of the logarithm of the encoding vector is modeled as a multivariate Gaussian distribution, and the log-likelihood of the current estimate for the encoding vector is incorporated in the objective function. However, our analysis on $\mathbf{H}^{train}$ revealed that each row of this matrix was highly sparse, which implies that the lognormal distribution may not be the best parametric form of prior knowledge. As alternative models for the distribution of the encoding vector, an independent exponential or gamma distribution was proposed in our previous research [49], which resulted in better source separation performance. But the performance of [10] and [49] degrades significantly when the signal and interference levels of the test data deviate much from those of the training data. In [58], [86]- [88], several prior distributions for the encoding vector were proposed in the Bayesian framework, for which the parameters were not estimated from the training data.

In this chapter, we extend our previous study in [49] to address the problem of possible mismatch between the training and test data levels by introducing a new penalty function of parameter training. Assuming that the statistical characteristics of the encoding vector for a specific source are stationary except for the level of the signal, we model the distribution of the components of the encoding vector as a multivariate exponential PDF (MVE) with a single time-varying scaling factor for each source. The parameters of the MVE are initially estimated from $\mathbf{H}^{train}$ and then continuously adjusted with suitable scaling factors to match the current input level. The scaling factor is estimated according to the maximum likelihood criterion in conjunction with temporal smoothing. Experimental results on audio source separation in which the target signal was speech showed that the proposed method could enhance the separation performance in term of the signal-to-distortion ratio (SDR) [42] even in the presence of the signal level mismatch.

7.2 Prior model of encoding vectors based on multivariate exponential distribution

Although most of the previous works use only the trained basis matrix during the source separation phase, the encoding matrix obtained for the training data is considered to possess important information as to how each basis is utilized to reconstruct the clean source signals. In the training procedure, $\mathbf{W}_S \in \mathbb{R}_+^{M \times R_s}$ and

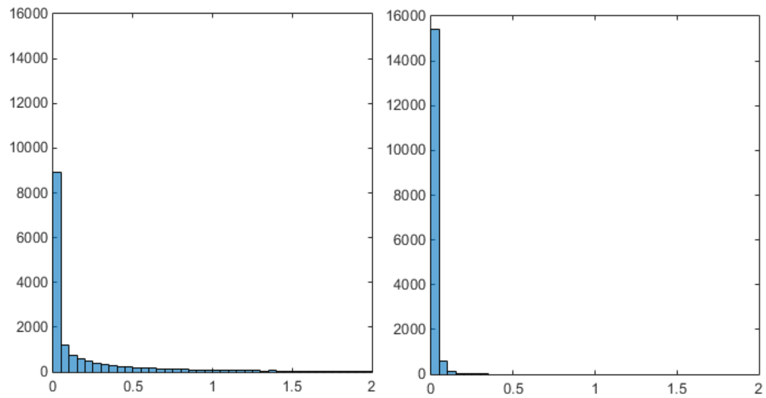


Figure 7.1: The histograms of two rows of $\mathbf{H}_S^{train}$ corresponding to the most frequently and rarely used basis vectors.

$\mathbf{H}_S^{train} \in \mathbb{R}_+^{R_s \times N_s}$ are obtained through the NMF analysis of the clean target signal data $\mathbf{V}_S^{train} \in \mathbb{R}_+^{M \times N_s}$ while $\mathbf{W}_N \in \mathbb{R}_+^{M \times R_n}$ and $\mathbf{H}_N^{train} \in \mathbb{R}_+^{R_n \times N_n}$ are computed from the interference signal data $\mathbf{V}_N^{train} \in \mathbb{R}_+^{M \times N_n}$. In [10], the distribution of the elements of an encoding vector is modeled as R-dimensional multivariate lognormal PDF of which the parameters are estimated from $\mathbf{H}_S^{train}$ and $\mathbf{H}_N^{train}$ assuming that H_S and H_N are independent. Based on this statistical model, the log-likelihood of the current estimate of H for a test data V is also maximized while minimizing the reconstruction error in the source separation phase. Although the utilization of the prior knowledge on H brought about some performance improvement, the choice of the statistical model in [10] does not fit well to the actual distribution of $\mathbf{H}_S^{train}$ and $\mathbf{H}_N^{train}$. Fig. 1 shows the histograms corresponding to the two rows of $\mathbf{H}_S^{train}$ having the largest and smallest L_1 norms, where $\mathbf{H}_S^{train}$ was obtained through the standard NMF analysis of the clean speech training database. These rows roughly correspond to the most frequently and rarely used basis vectors. The shape of the histograms reveals that $\mathbf{H}_S^{train}$ can be better approximated by the gamma or exponential distribution rather than the lognormal distribution as used in [10]. The analysis on the distributions of the elements of $\mathbf{H}_N^{train}$ for various interfering signals also showed similar tendency. Based on this analysis, the distribution of each component of the encoding vector is modeled by an independent exponential or gamma distribution in [49].

The biggest issue on the approaches in [10] and [49] which utilizes the distribution of the $\mathbf{H}^{train}$ may be that the distributions of the training and test data can be different, especially in signal level. This discrepancy would lead to a degradation of the performance in the presence of the signal or interference level mismatch, as shown in Section IV. On the other hand, if the parameters of the exponential distributions adapt to the test data independently, all the prior informations obtained from the training data would become useless. We can also consider to model the distribution of the normalized encoding vectors as in [87] and [88], which, however, usually results in too diverse distribution patterns depending on the signal level to be modeled with a single parametric distribution.

In order to compensate the level mismatch between the training and test data while maintaining the prior information collected from the training data, we model the distribution of an R-dimensional encoding vector as an R-parameter multivariate exponential distribution PDF (MVE) [89], [90] where we add a single time-varying scaling factor. The cumulative distribution function of the R-parameter MVE for a random vector $X = (X_1, ..., X_R)$ is given as

$$P(X_1 > x_1, ..., X_R > x_R) = exp[-\sum_{r=1}^{R} \eta_r x_r], \tag{7.1}$$

for nonnegative x_r's where η_r is a nonnegative rate parameter corresponding to the r-th component [90]. The parameters of the R-parameter MVE are initially estimated from $\mathbf{H}^{train}$ by the MLE, as given by

$$\eta_r = \frac{N}{\sum_{p=1}^{N} [\mathbf{H}^{train}]_{r,p}}, r = 1, 2, ..., R, \tag{7.2}$$

where $[\cdot]_{r,p}$ denotes the (r, p)-th component of a matrix.

During the separation phase, the parameters $\boldsymbol{\eta} = [\eta_1, \eta_2, ..., \eta_R]^T$ are modified with a time varying scaling factor $\alpha_i(t) \in \mathbb{R}_+$ for each source to compensate the level mismatch. We will assume that there are two signals in the mixture for simplicity. With the modified parameters $\boldsymbol{\eta}(t) = [\alpha_S(t)\boldsymbol{\eta}_S{}^T, \alpha_N(t)\boldsymbol{\eta}_N{}^T]^T \in \mathbb{R}_+^{(R_s+R_n)\times 1}$, the objective function of the NMF in the separation phase is combined with the log-likelihood of the current estimate for $H(t) = [H_S{}^T(t), H_N{}^T(t)]^T$ which gives as

$$f(H(t)) = D(V(t) \mid \mathbf{W}H(t)) + \gamma_p[(\boldsymbol{\eta}(t) \otimes H(t)) \cdot \mathbf{1}] \tag{7.3}$$

where $\mathbf{1}$ indicates a vector of suitable size with all elements equal to one, γ_p denotes a parameter controlling the trade-off between the reconstruction error and the log-likelihood, and $\cdot$ represents inner product. The MuR in (7.3) with KLD in the separation phase now becomes

$$[H(t)]_r \leftarrow [H(t)]_r \frac{\sum_{k=1}^{M} \frac{\mathbf{W}_{k,r}[V(t)]_k}{\sum_{f=1}^{R} \mathbf{W}_{k,f}[H(t)]_f}}{\sum_{k=1}^{M} \mathbf{W}_{k,r} + \gamma_p[\boldsymbol{\eta}(t)]_r}. \tag{7.4}$$

The time-varying scaling factor for each source, $\alpha_i(t)$, is determined through the MLE with temporal smoothing while utilizing a crude estimate of $H(t)$, $\check{H}(t) = [\check{H}_S^T(t), \check{H}_N^T(t)]^T$, obtained by a single iteration of (2.8) in which $\check{H}(t)$ is initialized by $\mathbf{W}^T V(t)$ [68]. The instantaneous estimate of $\alpha_i(t)$, $\check{\alpha}_i(t)$, is obtained through the MLE which is given as

$$\check{\alpha}_i(t) = \frac{R_i}{[\boldsymbol{\eta}_i \otimes \check{H}_i(t)] \cdot \mathbf{1}} \tag{7.5}$$

where the subscript i denotes either target or interfering signal. For a more robust estimation, it is useful to apply temporal smoothing such that

$$\alpha_i(t) = (1-\beta)\alpha_i(t-1) + \beta\check{\alpha}_i(t) \tag{7.6}$$

where β indicates a forgetting factor.

7.3 Experiment result

To evaluate the performance of the proposed algorithm, we applied it to audio source separation in which the target source was speech. The proposed objective function based on the prior distribution of H modeled by the MVE with adaptive scaling factor was compared with those based on the lognormal distribution [10] and the exponential distribution without parameter adaptation [49]. In addition, we also compared the proposed algorithm with that based on independent exponential distributions with individual parameter adaptation.

Speech samples were selected from TIMIT DB [74] while the interference signals used for the experiments were the *F-16*, *factory1*, and *babble* noises extracted from the NOISEX-92 DB [75]. Each signal was sampled at 16 kHz, and the Hamming window and a 512-point discrete Fourier transform with 75% overlap were applied to form a spectrogram. The training DB for speech consisted of 102-second long speech data spoken by 40 different speakers, while the noise data for training were 117-second long in total for each type of noise. To test the proposed and conventional methods, 32 sentences spoken by 32 different speakers which were not included in the training DB were mixed with the aforementioned three types of noise data at 0 dB SNR. There was no overlap between the training and test data in both the speech and interfering types. MuR was applied with the distance measure of KLD

Table 7.1: The signal-to-distortion ratios for the same test signals of 0dB SNR with various training data levels.

The levels of the training signals		The penalty terms used in the separation phase					
		None (Standard)	**L1**	**Lognorm**	**Exp**	**Exp+IndAd**	**Prop**
matched		4.1985	5.2314	**4.4233**	**5.7746**	**6.1359**	**6.7656**
mismatched	S+10dB,N+10dB			3.3182	5.1503	6.1584	6.7514
	S-10dB,N-10dB			4.8476	5.7834	6.1833	6.7609
	S+0dB,N+10dB			3.5075	5.8872	6.1315	6.7465
	S+0dB,N-10dB			4.4920	3.5713	6.0952	6.7404
	S+10dB,N+0dB			4.3324	3.9949	6.0447	6.7562
	S-10dB,N+0dB			4.3205	3.6359	6.2242	6.7783
	S-10dB,N+10dB			3.5774	2.8089	6.2484	6.7755
	S+10dB,N-10dB			4.3725	1.9302	5.9443	6.7100
	mismatched average			**4.0960**	**4.0952**	**6.1287**	**6.7524**

(The None and L1 values, **4.1985** and **5.2314**, span all rows.)

in the NMF analysis, and the numbers of iterations for the training and test phases were set to 100 and 10, respectively. The number of bases R for each source was set to 64, which provided a good trade-off between the reconstruction error and the computational complexity.

The penalty terms used for the separation phase were:

- **None**: without any penalty term (standard NMF)
- **L1**: L_1 norm-based sparsity penalty term [91], [92]
- **Lognorm**: the negative log-likelihood assuming that H follows the multivariate lognormal PDF [10]
- **Exp**: the negative log-likelihood of H where the distribution for H is assumed to be an independent exponential PDF [49]
- **Exp+IndAd**: *Exp.* with individual parameter adaptation, i.e., with separate scaling factor for each basis
- **Prop**: the negative log-likelihood of H where the distribution for H is assumed to be MVE with one scaling factor for each source

Since the penalty terms for **Exp** and **Prop** have a similar form with the L_1 norm-based constraint, **L1** was also applied for performance comparison [91], [92]. The parameters of the lognormal and exponential PDFs were obtained from the 1st and 2nd order moments of $\mathbf{H}_S^{train}$ and $\mathbf{H}_N^{train}$ or logarithm of them. The γ parameters which control the trade-off between the reconstruction error and the penalty term were chosen to maximize the source separation performance for the validation set of 4 sentences with matched level, which resulted in the ranges $\gamma_{L1}\in [1, 3]$, $\gamma_{lognorm}\in [0.01, 10]$, $\gamma_{Exp}\in [0.16, 0.18]$, $\gamma_{Exp+IndAd}\in [0.16, 0.18]$ and $\gamma_p\in [0.16, 0.18]$. The scaling factors for the **Exp+IndAd** were computed in a similar way to (7.5) and (7.6) with a single element of $\check{H}(t)$.

The SDRs for the same test signals with various training data levels averaged over all noise types are shown in Table I. For the 'matched' case in which the levels of the signal and interference in the test data were the same as those in the training data, the algorithms based on sparse prior models showed superior performances. The fact that the adaptation of the parameters brought about the performance improvement even in the matched case may be due to its ability to track the temporal evolution of the signal and interference level. The performance gap between **Prop** and **Exp+IndAd** implies that keeping relative magnitudes among the elements of $\boldsymbol{\eta}(t)$ is effective.

For the mismatched case in which the levels of the signal and interference in the training data differ from those in test data, the performances of the conventional prior model-based systems deteriorated significantly as expected. In Table I, '$S + 10dB, N - 10dB$' for instance indicates that the average magnitudes of speech and noise signals during the training phase were 10 dB higher and lower than those of the test data, respectively. In contrast, two systems with parameter adaptation turned out to be very robust to the magnitude level mismatch. As in the matched case, **Prop** outperformed **Exp+IndAd**, which implies that maintaining relative magnitudes of the parameters for each source is important to guarantee performance in mismatched condition.

Chapter 8

Conclusions

The audio source separation is a extracting or reconstruction of the target source signals from the mixture signals which consist of the target and interfering sources. The compositional model-based approaches which have a part-based feature have been actively utilized for the audio source separation and these approaches show a better performance than those of the statistical model, PCA, and ICA. In the compositional models, NMF is representative and widely used for the audio source separation. Although NMF-based audio source separation shows a proper performance, it has various issues for the performance. In this dissertation, we propose diverse approach for NMF-based audio source separation and analysis sparse NMF for the audio source separation.

The first approach is a speech enhancement technique combining statistical models and non-negative matrix factorization (NMF) with on-line update of speech and noise bases. The statistical model-based enhancement methods have been known to be less effective to non-stationary noises while the template-based enhancement techniques can deal with them quite well. However, the template-based enhancement techniques usually rely on *a priori* information. To overcome the shortcomings of both approaches, we propose a novel speech enhancement method that combines the statistical model-based enhancement scheme with the NMF-based gain function. For a better performance in time-varying noise environments, both the speech and noise bases of NMF are adapted simultaneously with the help of the estimated speech presence probability. Because the proposed online update properly reflects each source information of the mixture signals, the separation performance of the semi-supervised case is similar or better than those of the supervised case.

In the second approach, we propose discriminative NMF using the reconstruction error of the other source, which means that the basis vectors of the target source should not be used to represent of the background source. A desired bases of target source should reconstruct the target source well, while do not represent the other source. For this issue, we add a proper constraint to the objective function of NMF. This proper constraint is utilized cross-reconstruction error of the interfering signals, and this approach outperformed the other DNMF algorithms.

In the third approach, the incremental approaches for NMF bases estimation are proposed. Since, however, the objective function of NMF is non-convex, the performance of the source separation can degrade when the iterative update of the basis matrix in the training procedure is stuck to a poor local minimum. In most of the previous studies, the whole basis matrix for a specific source is iteratively updated to minimize a certain objective function with random initialization although a few approaches have been proposed for the systematic initialization of the basis matrix such as the singular value decomposition and k-means clustering. Based on an analogy between clustering and NMF analysis, we incrementally estimate the NMF bases similar to the modified k-means and Linde-Buzo-Gray algorithms popular in the data clustering area.

In the last approach, the distribution of the encoding vector is modeled as a multivariate exponential PDF (MVE) with a single scaling factor for each source. The parameters of the MVE are initially estimated from the encoding matrix of the training data, and adjusted in the test phase with a single scaling factor for each source which is updated by the maximum likelihood estimation to deal with potential level mismatch between the training and test data. A new objective function of NMF analysis based on the prior model is proposed. The proposed method shows a better performance than before regardless of the mismatch between the training and test DBs.

Bibliography

[1] M. Zibulevsky and B. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural computation*, vol. 13, no. 4, pp. 863-882, 2001.

[2] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 191-199, 2006.

[3] O. Nieto and T. Jehan, "Convex non-negative matrix factorization for automatic music structure identification," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 236-240, 2013.

[4] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1066-1074, 2007.

[5] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 550-563, 2010.

[6] E. M. Grais and H. Erdogan, "Discriminative nonnegative dictionary learning using crosscoherence penalties for single channel source separation," *INTERSPEECH*, pp. 808-812, 2013.

[7] F. Weninger, J. L. Roux, J.R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," *Proc. of ISCA Interspeech*, 2014.

[8] Z. Wang and F. Sha, "Discriminative non-negative matrix factorization for single-channel speech separation," *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE*, pp. 3749-3753, 2014.

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788-791, 1999.

[10] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," *INTERSPEECH*, pp. 411-414, 2008.

[11] Blind source separation advances in theory, algorithms and applications, Springer, 2014.

[12] C. Joder, F. Weninger, D. Virette, and B. Schuller, "A comparative study on sparsity penalties for NMF-based speech separation: beyond LP-norms," *Acoustics, Speech and Signal Process.*, pp. 858-862, 2013.

[13] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 125-143, Mar. 2015.

[14] P. Smaragdis, C. Fevotte, G. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 66-75, 2014.

[15] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," *in IEEE Int. Conf. on Acoustics, Speech and Signal Process.*, 2008.

[16] A. Cichocki, R. Zdunek, A.-H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation.* New York: Wiley, 2009.

[17] K. Kwon, J. W. Shin, and N. S. Kim, "Target Source Separation Based on Discriminative Nonnegative Matrix Factorization Incorporating Cross-Reconstruction Error," *IEICE Trans. on Information and Systems*, vol. E98-D. no. 11, pp.-, Nov. 2015.

[18] K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based source separation with prior models," *IEEE Signal Process. Lett.*, vol. , no. , pp. -, Apr. 2017.

[19] K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based speech enhancement using bases update," *IEEE Signal Process. Lett.*, vol. 22, no. 4, pp. 450-454, Apr. 2015.

[20] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization," *In Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on,* pp. 45-48, 2011.

[21] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. on Audio, Speech, and Language process.*, vol. 15, no. 1, pp.1-15, Jan. 2007.

[22] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden markov modeling of audio with application to source separation," in *Int. Conf. Latent Variable Analysis and Signal Separation*, pp. 140-148, 2010.

[23] G. Zhou, A. Cichocki, Q. Zhao, and S. Xie, "Nonnegative matrix and tensor factorizations," *IEEE Signal Process. Mag.*, vol.31, no.3, pp. 54-65, May 2014.

[24] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-33, pp. 443-445, Apr. 1985.

[25] N. S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Processing Lett.*, vol. 7, no. 5, May 2000.

[26] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403-2418, Nov. 2001.

[27] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Commun.*, vol. 48, pp. 220-231, 2006.

[28] D. Ellis and R. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," *in IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, pp. 957-960, 2006.

[29] N. Mohammadiha, P. Smaragdis and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140-2151, 2013.

[30] C. D. Sigg, T. Dikk and J. Buhmann, "Speech enhancement using generative dictionary learning," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 6, Aug. 2012.

[31] J. Ming, R. Srinivasan and D. Crookes, "A corpus-based approach to speech enhancement from nonstationary noise," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 822-836, May 2011.

[32] P. Smaragdis, M. V. Shashanka and B. Raj, "A sparse non-parametric approach for Single Channel Separation of Known Sounds." *In NIPS*, pp. 1705-1713, 2009.

[33] R. Talmon, I. Cohen, S. Gannot and R. R. Coifman, "Supervised graph-based processing for sequential transient interference suppression," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2528-2538, 2012.

[34] S. Srinivasan, J. Samuelsson and W.B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Transaction on Audio, Speech, And Language Processing*, vol. 14, no. 1, pp. 163-176, Jan. 2006.

[35] S. M. Kim, J. H. Park, H. K. Kim, S. J. Lee and Y. K. Lee, "Non-negative matrix factorization based noise reduction for noise robust automatic speech recognition," *Lecture Notes in Computer Science*, vol. 7191, pp. 338-346, 2012.

[36] G. Cabras, S. Canazza, P. L. Montessoro and R. Rinaldo, "Restoration of audio documents with low SNR: a NMF parameter estimation and perceptually

motivated Bayesian suppression rule," *in Proc. Sound and Music Computing Conference*, pp. 314-321, 2010.

[37] M. N. Schmidt, J. Larsen, and F. T. Hsiao, "Wind noise reduction using non-negative sparse coding," *Machine Learning for Signal Processing, 2007 IEEE Workshop on*. IEEE, pp.431-436, 2007.

[38] N. Guan, D. Tao, Z. Luo and B. Yuan, "Online nonnegative matrix factorization with robust stochastic approximation," *IEEE Trans. Neural networks and Learning systems*, vol. 23, no. 7, Jul. 2012.

[39] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *The Journal of Machine Learning Research*, vol. 11, pp. 19-60, 2010.

[40] D. Wang, R. Vipperla and N. Evans, "Online pattern learning for non-negative convolutive sparse coding," *INTERSPEECH*, pp. 65-68, 2011.

[41] S. Rebhan, W. Sharif and J. Eggert, "Incremental learning in the non-negative matrix factorization," *Advances in Neuro-Information Processing*, Springer Berlin Heidelberg, pp. 960-969, 2009.

[42] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp. 1462-1469, 2006.

[43] "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.

[44] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229-238, 2008.

[45] H. Liu, Z. Wu, D. Cai, and T. S. Huang, "Constrained nonnegative matrix factorization for image representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1299-1311, 2012.

[46] F. Shahnaz, M. M. Berry, V.P. Pauca, and R. J. Plemmons, "Document clustering using nonnegative matrix factorization," *Information Process. & Management*, vol. 42, no. 2, pp. 373-386, 2006.

[47] H. A. Song, B. K. Kim, T. L. Xuan, and S. Y. Lee, "Hierarchical feature extraction by multi-layer non-negative matrix factorization network for classification task," *Neurocomputing*, 165, pp. 63-74, 2015.

[48] A. Likas, N. Vlassis, J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognition*, vol. 36, no. 2, pp. 451-461, 2003.

[49] K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based source separation utilizing prior knowledge on encoding vector," *in IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[50] J. Mairal, F. Bach, and J, Ponce, "Sparse modeling for image and vision processing," *Foundations and Trends in Computer Graphics and Vision*, vol. 8, no. 2-3, pp. 85-283. 2014.

[51] J. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 19, no. 7, pp. 2067-2080, 2011.

[52] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, no. Article ID 785152, 2009.

[53] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," *Proceedings of the 26th annual international conference on machine learning. ACM*, pp. 689-696, 2009.

[54] M. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete latent variable decomposition of counts data," *in Proc. Neural Information Processing Systems, Canada*, pp. 1313-1320, 2007.

[55] C. Ding, T. Li, and W. Ping, "On the equivalence between nonnegative matrix factorization and probabilistic latent semantic indexing," *Computat. Stat. Data Anal.*, vol. 52, no. 8, pp. 3913-3927, 2008.

[56] C. J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural computation*, vol. 19, no. 10, pp. 2756-2779, 2007.

[57] N. Guan, D. Tao, Z. Luo, and B. Yuan, "NeNMF: an optimal gradient method for nonnegative matrix factorization," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2882-2898, 2012.

[58] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793-830, 2009.

[59] F. Weninger, J. L. Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," *INTERSPEECH*, pp. 865-869, 2014.

[60] J. L. Roux, F. Weninger, and J. R. Hershey, "Sparse NMF-half-baked or well done?," Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023, 2015.

[61] J. Eggert and E. Korner, "Sparse coding and NMF," *Proc. of Neural Networks*, vol. 4, pp. 2529-2533, 2004.

[62] F. Weninger, J, Feliu, and B. Schuller, "Supervised and semi-supervised suppression of background music in monaural speech recordings," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 61-65, 2012.

[63] S. Wild, "Seeding non-negative matrix factorizations with spherical k-means clustering," *Master's thesis*, University of Colorado, 2003.

[64] S. Wild, J. Curry, and A. Dougherty, "Improving non-negative matrix factorizations through structured initialization," *Pattern Recognition*, vol. 37, no. 11, pp. 2217-2232, 2004.

[65] Z. Zheng, J. Yang, and Y. Zhu, "Initialization enhancer for non-negative matrix factorization," *Engineering Applications of Artificial Intelligence*, vol. 20, no. 1, pp. 101-110, 2007.

[66] R. Albright, J. Cox, D. Duling, A. N. Langville, and C. D. Meyer, "Algorithms, initializations, and convergence for the nonnegative matrix factorization," *presented at the 12th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Aug. 2006.

[67] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design,"*IEEE Transactions on Communications*, vol. com-28, no. 1, pp. 84-95, Jan. 1980.

[68] L. Gong and A. K. Nandi, "An enhanced initialization method for non-negative matrix factorization." *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*, IEEE, 2013.

[69] C. Boutsidis and E. Gallopoulos, "SVD based initialization: A head start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, no. 4, pp. 1350-1362, 2008.

[70] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457-1469, 2004.

[71] A. Pascual-Montano, J. Carazo, K. Kochi, D. Lehmann, and R. D. Pascual-Marqui, "Nonsmooth nonnegative matrix factorization (NSNMF)," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 403-415, Mar. 2006.

[72] Y. Qian, S. Jia, J. Zhou, and A. Robles-Kelly, "Hyperspectral unmixing via sparsity-constrained nonnegative matrix factorization," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 49, no. 11, pp. 4282-4297, 2011.

[73] P. O'Grady and A. P. Barak, "Discovering convolutive speech phones using sparseness and non-negativity," *Int. Conf. on Independent Component Analysis and Signal Separation*, Springer Berlin Heidelberg, 2007.

[74] L. Larnel, R. Kassel, and S. Seneff, "Speech database development: design and analysis of the acoustic-phonetic corpus," *Proc. DARPA Speech Recognition Workshop*, pp. 26-32, Mar. 1987.

[75] A. Varga, H.J.M Steenneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," 1992. Documentation included in the NOISEX-92 CD-ROMs.

[76] N. Guan, D. Tao, Z. Lwo, and B. Yuan, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent," *Image Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 2030-2048, 2011.

[77] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Discriminative Non-negative Matrix Factorization for Multiple Pitch Estimation," *ISMIR*, pp. 205-210, 2012.

[78] P. O'grady and B. A. Pearlmutter, "Convolutive non-negative matrix factorisation with a sparseness constraint," *Machine Learning for Signal Process. 2006. Proc. of the 2006 16th IEEE Signal Process. Society Workshop on*, pp. 427-432, 2006.

[79] A. Singh and G. Gordon, "A unified view of matrix factorization models," *Mach. Learn. Knowledge Discovery Databases*, vol. 5212, pp. 358-373, 2008.

[80] M.E. Celebi, H.A. Kingravi, and P.A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," *Expert Syst. Appl.*, vol. 40, no. 1, pp. 200-210, 2013.

[81] P. Bradley and U. Fayyad, "Refining initial points for K-means clustering," *in Proc. 15th Int. Conf. Machine Learning*, pp. 91-99, 1998.

[82] A. Belouchrani, K. A. Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique based on second-order statistics," *IEEE Trans. Signal Process.*, vol. 45, pp. 434-44, Feb. 1997.

[83] M. E. Davies and C. J. James, "Source separation using single channel ICA," *Signal Process.*, vol. 87, no. 8, pp. 1819-1832, 2007.

[84] A. Hyvarinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Net*, vol. 13, no. 4-5, pp. 411-430, Jun. 2000.

[85] A. Cichocki, R. Zdunek, A.H. Phan, and S. Amari, "Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation," Wiley-Blackwell, 2009.

[86] O. Dikmen, and C. Févotte, "Maximum marginal likelihood estimation for non-negative dictionary learning in the Gamma-Poisson model," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5163-5175, 2012.

[87] T. Virtanen and A. T. Cemgil, "Mixtures of gamma priors for non-negative matrix factorization based speech separation," *International Conference on Independent Component Analysis and Signal Separation*, Springer Berlin Heidelberg, pp. 646-653, 2009.

[88] E. M. Grais and H. Erdogan, "Regularized nonnegative matrix factorization using Gaussian mixture priors for supervised single channel source separation," *Computer Speech & Language*, vol. 27, no. 3, pp. 746-762, 2013.

[89] A. W. Marshall and I. Olkin, "A multivariate exponential distribution," *Journal of the American Statistical Association*, vol. 62, no. 317, pp. 30-44, 1967.

[90] F. Proschan and P. Sullo, "Estimating the parameters of a multivariate exponential distribution," *Journal of the American Statistical Association*, vol. 71, no.354, pp. 465-472, 1976.

[91] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization." *Spoken Language Processing, ISCA International Conference on (INTERSPEECH)*, 2006.

[92] P. O. Hoyer, "Nonnegative Sparse Coding," *Proc. IEEE Workshop Neural Networks for Signal Process.*, pp. 557-565, 2002.

국문초록

많은 종류의 데이터들은 부분들로 이루어진 조합들로 표현 될 수 있다. 자연적으로 발생한 거의 모든 신호와 데이터는 비음수 값들로 나타낼 수 있고 이는 조합 모델로 설명이 가능하다. 이러한 조합 모델에서는 오로지 더하기 형태의 조합만 가능하다. 즉 어떠한 부분을 빼는 것은 고려되지 않는다. 조합 모델들(compositional models)에는 사전 학습, 전형 기반 기법, 그리고 비음수 행렬 인수 분해(nonnegative matrix factorization, NMF) 등이 있다. 조합 모델들은 영상 및 이미지 신호 처리, 문서 정보 처리, 음향 신호 처리, 음악 정보 처리 등 다양한 분야에 사용되고 있다. 본 논문에서는 조합 모델들 중 NMF를 이용하여 음향 신호 분리를 수행한다.

목표 음원 분리는 목표 음원과 그 외 방해 음원이 섞인 입력 신호로부터 목표 음원 만을 복원 또는 추출 하는 것을 말한다. 목표 음원 분리는 블라인드 소스 분리(blind source separation, BSS)로 설명할 수 있다. BSS는 최소한의 사전 정보로 목표 음원을 복원한다. 하지만 최근에는 많은 양의 사전 정보가 음원 분리에 사용되고 있다. 단일 채널 소스 분리를 위해 기존에 독립 원소 분석, 스파스(sparse) 분해, 주 원소 분석, 단수 값 분해, 계산적 청각 장면 분석, NMF 등의 알고리즘들이 적용되었다.

NMF는 비음수 데이터 행렬 V를 비음수 기저와 인코딩 행렬들의 곱으로 근사시킨다. 기저와 인코딩이 비음수이기 때문에 NMF 는 부분 기반의 데이터 표현을 보인다. NMF 기반 방법들은 단일 채널 소스 분리에서 좋은 결과를 보이고 있다. NMF 의 목적 함수는 일반적으로 유클리디언 거리(Euclidean distance), 컬백 라이블러 발산(Kullback-Liebler divergence), 그리고 이타쿠라 사이토 발산(Itakura-Saito divergence) 등으로 나타낸다. 이 목적함수로부터 기저와 인코딩

행렬을 구하기 위해 증가 업데이트 방식, 투영된 기울기 하강(projected gradient descent), NeNMF 등의 다양한 최적화 기법이 적용되었다. 그러나 NMF 기반 음원 분리에는 여러가지 문제점들이 존재한다. 문제점들에는 기저들의 유일성 불만족, 사전 정보에 높은 의존성, 목적 기저와 방해 기저의 겹치는 영역, 인코딩 정보의 부족한 활용, 스파스 NMF 에 대한 분석 부족 등이 있다. 본 논문은 이러한 문제점을 해결하는 방법들을 제시하였다.

4장에서는 통계 모델 기반 음성 향상과 NMF 기반 음성 향상 기법을 결합하여 기저들을 실시간으로 업데이트 하는 방법을 제안했다. 목표 음원의 종류만 알면 배경 잡음에 대한 사전 정보 없이 높은 음성 향상을 보임을 확인하였다. 5장에서는 목표 기저들로 만들어지는 영역이 방해 기저들의 것들과 최소한으로 겹치게 하는 새로운 배타적(discriminative) NMF 를 제안했다. 기존 배타적 NMF 에 비해 매우 높은 성능 향상을 보였다. 6장에서는 기저들을 증가 방식에 따라 개수를 늘려가는 기법들을 제안하였다. 음성 외에 바이올린, 피아노 음원에서도 높은 성능을 보임을 확인했다. 7장에서는 인코딩 벡터들의 정보를 이용하여 특정 통계 모델을 만들고 이를 음원 분리 과정에 활용하는 기법을 제안했다. 기저 뿐만 아니라 인코딩 정보 또한 사전 모델을 이용하여 조절 해줌으로써 높은 성능 향상을 가져 올 수 있었다.

주요어 : 조합 모델, 음향 신호 분리(음원 분리), 음성 향상, 비음수 행렬
인수분해(NMF), 기저 업데이트, 배타적 NMF, 증가 방식, 인코딩 벡터
학 번 : 2011-20788