Multiple-order Non-negative Matrix Factorization for Speech Enhancement

Xabier Jaureguiberry[1*], Emmanuel Vincent[2], Gaël Richard[1]

[1] Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, 37-39, rue Dareau 75014 Paris, France
[2] Inria, 54600 Villers-lès-Nancy, France

Abstract

Amongst the speech enhancement techniques, statistical models based on Non-negative Matrix Factorization (NMF) have received great attention. In a single channel configuration, NMF is used to describe the spectral content of both the speech and noise sources. As the number of components can have a crucial influence on separation quality, we here propose to investigate model order selection based on the variational Bayesian approximation to the marginal likelihood of models of different orders. To go further, we propose to use model averaging to combine several single-order NMFs and we show that a straightforward application of model averaging principles is inefficient as it turned out to be equivalent to model selection. We thus introduce a parameter to control the entropy of the model order distribution which makes the averaging effective. We also show that our probabilistic model nicely extends to a multiple-order NMF model where several NMFs are jointly estimated and averaged. Experiments are conducted on real data from the CHiME challenge and give an interesting insight on the entropic parameter and model order priors. Separation results are also promising as model averaging outperforms single-order model selection. Finally, our multiple-order NMF shows an interesting gain in computation time.

Index Terms: Variational Bayes, Non-negative Matrix Factorization, Model Averaging, Speech Enhancement

1. Introduction

Speech enhancement has received great attention since it is at stake in numerous industrial applications. The literature proposes a variety of methods which fall into two categories [1] : multichannel vs. single-channel. In a single-channel configuration, which is the focus of this paper, spatial diversity cannot be exploited. Single-channel techniques can be classified into three main categories [2]: spectral subtractive algorithms [3], subspace algorithms [4] and statistical-model-based algorithms [5]. Amongst the latter, various models aim at describing the spectral content of both the speech and background noise sources in order to be able to distinguish between them. To build such models, one can resort for instance to Gaussian mixture models [6], exemplar-based methods [7] or codebook-driven techniques [8, 9]. Non-negative Matrix Factorization (NMF) is one of the most popular class of source models [10, 11, 12, 13] and it has achieved great performance in the latest CHiME contest [13, 14].

The order of an NMF model, also called the number of components, is known to have a noticeable influence on separation quality [15]. However, there is a few literature about how to determine the best number of components [9] and choosing it is often driven by an experimental assessment. The introduction of a statistical formulation of NMF [16] has made the application of model selection principles possible. To do so, the literature advocates the use of a full Bayesian framework [17, 18] in order to compute the marginal likelihood of a model, also named the *evidence*. For a given task, once the marginal likelihood has been estimated for several models of different orders, the model with the highest marginal likelihood is reputed to be the most likely model to explain the observation. In practice, a full Bayesian treatment of NMF is intractable and approximate inference is required instead. In particular, Variational Bayesian (VB) inference is becoming popular since it is less computationally demanding than sampling methods [17, 19].

VB proposes to approximate the marginal likelihood by a lower bound called the *free energy*. This free energy can then be used in place of the true marginal likelihood for model order selection. Two types of approaches have been exploited so far: parametric methods which consist in computing several NMFs and selecting the one which has the highest free energy [19], and nonparametric methods which consider a single NMF with a potentially infinite number of components and which iteratively deactivates the irrelevant components [20].

To go further, we propose here to apply model averaging principles to NMF models [18]. Indeed, the study in [21] has shown that it is worth combining several NMFs of different orders instead of selecting a unique one. The free energy given by VB inference can be used to compute the posterior probability of each number of components and to weight each NMF [17]. We also derive from our averaging of single-order NMFs a novel multiple-order model which jointly estimates and averages several NMFs of different orders. However, our contribution underlines that a straightforward application of model averaging based on the free energy is inefficient as it turns out to select a single NMF. To avoid this behaviour, we propose to use a parameter which controls the entropy of the distribution of the number of components. As this entropic parameter and the order priors need to be chosen beforehand, we propose to learn them on a training database. Our experiments conducted on real data from the CHiME challenge [22] show promising results as both the multiple-order NMF and the averaging of single-order NMFs outperform single-order model selection, thanks to the introduction of the entropic parameter. Moreover, our multiple-order NMF turns out to be less computationally expensive than the averaging of single-order models for equivalent performance.

In the following, Section 2 will be dedicated to the presentation of the single-order NMF model and its averaging. In Section 3, we will introduce our novel multiple-order NMF model. The entropic parameter will be presented in Section 4 before being experimentally evaluated in Section 5. Section 6 will give a concise conclusion.

2. Single-order NMF

Our single-order NMF model is a single-channel simplified formulation of the model exposed in [23]. The degraded speech signal is supposed to be a linear mixture of a clean speech sig-

[*] This work was partly supported under the research programme EDiSon3D (ANR-13-CORD-0008-01) funded by ANR, the French State agency for research.

nal and a background noise signal. As such, the mixing equation can be written in the short-time Fourier transform (STFT) domain as

$$x_{fn} = \mathbf{A}\mathbf{s}_{fn} + \epsilon_{fn} = s_{1,fn} + s_{2,fn} + \epsilon_{fn} \tag{1}$$

in which f denotes the frequency bin and n the time frame, $s_{1,fn}$ is the target speech signal, $s_{2,fn}$ is the background noise and ϵ_{fn} represents sensor noise. By denoting the source vector as $\mathbf{s}_{fn} = [s_{1,fn}\, s_{2,fn}]^T$, the mixing equation can also be written in a matrix form thanks to the mixing matrix $\mathbf{A} = [1\, 1]$. Such a formulation permits future extension of our model to more sources and channels.

2.1. Probabilistic model

Speech and background noise sources are both treated in the same way and indexed by $j = \{1, 2\}$. We assume that the source $s_{j,fn}$ follows a circularly-symmetric complex normal distribution $s_{j,fn} \sim \mathcal{N}(0, v_{j,fn})$ whose variance is the result of an NMF so that

$$v_{j,fn} = \sum_{k=1}^{K_j} w_{j,fk} h_{j,kn}. \tag{2}$$

K_j is the order of the NMF, *a.k.a.* the number of components. In this single-order NMF formulation, it is an hyperparameter to be chosen. The variance of source j can also be written in a matrix form so that $\mathbf{V}_j = \mathbf{W}_j\mathbf{H}_j$. $\mathbf{W}_j$ and $\mathbf{H}_j$ are commonly called the dictionary and the activation matrix. To achieve full Bayesian inference, the parameters of both the dictionary and activation matrix are seen as random variables as well. Contrary to [23], we assume as in [20] that the NMF parameters follow a Gamma distribution

$$w_{j,fk} \sim \Gamma(a, a) \quad , \quad h_{j,kn} \sim \Gamma(b, b) \tag{3}$$

where a and b are hyperparameters to be chosen. Finally, sensor noise is supposed to follow a Gaussian distribution of variance σ^2. Denoting $\mathbf{X} = \{x_{fn}\}_{f=1..F}^{n=1..N}$ and $\mathbf{S} = \{\mathbf{s}_{fn}\}_{f=1..F}^{n=1..N}$ for the sake of readability, the log-likelihood can be formulated as

$$\log p(\mathbf{X}|\mathbf{S}) = \sum_{n=1}^{N}\sum_{f=1}^{F} \log \mathcal{N}\left(x_{fn}|\mathbf{A}\mathbf{s}_{fn}, \sigma^2\right). \tag{4}$$

2.2. Variational inference

In the following, we denote the set of all model parameters as $\mathbf{Z} = \{\mathbf{S}, \mathbf{W}, \mathbf{H}\}$ with $\mathbf{W} = \{\mathbf{W}_j\}_{j=1,2}$ and $\mathbf{H} = \{\mathbf{H}_j\}_{j=1,2}$. VB inference aims at approximating the posterior distribution of the model parameters $p(\mathbf{Z}|\mathbf{X})$ with a factorized variational distribution $q(\mathbf{Z})$ which is here defined as $q(\mathbf{Z}) = q(\mathbf{S})q(\mathbf{W})q(\mathbf{H})$ so that

$$q(\mathbf{Z}) = \prod_{fn} q(\mathbf{s}_{fn}) \prod_{j,fk} q(w_{j,fk}) \prod_{j,kn} q(h_{j,kn}). \tag{5}$$

In VB inference, minimizing the Kullback-Leibler divergence between $q(\mathbf{Z})$ and $p(\mathbf{Z}|\mathbf{X})$ is equivalent to approximating the log marginal likelihood $\log p(\mathbf{X})$ by the so-called *free energy* $\mathcal{L}[q]$ defined as

$$\mathcal{L}[q] = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})}\, d\mathbf{Z} = \mathbb{E}_{\mathbf{Z}}\left[\log \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})}\right] \tag{6}$$

with $p(\mathbf{X}, \mathbf{Z}) = p(\mathbf{X}|\mathbf{S})\, p(\mathbf{S}|\mathbf{W}, \mathbf{H})\, p(\mathbf{W})\, p(\mathbf{H})$ being the joint distribution. VB inference then consists in iteratively maximizing the free energy with respect to each factor in (5). In practice, the computation of the free energy is intractable and it needs to be further approximated by a parametric lower bound $\mathcal{B}[q]$. For a detailed explanation, the reader is referred to [23].

Deriving $\mathcal{B}[q]$ with respect to each factor leads to the update rules. The variational distribution of the sources is identified to a bivariate Gaussian distribution [23] $q(\mathbf{s}_{fn}) = \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{s},fn}, \boldsymbol{\Sigma}_{\mathbf{s},fn}\right)$ with parameters

$$\boldsymbol{\mu}_{\mathbf{s},fn} = \boldsymbol{\Sigma}_{\mathbf{s},fn}\mathbf{A}^T \frac{x_{fn}}{\sigma^2}, \quad \boldsymbol{\Sigma}_{\mathbf{s},fn} = \left(\mathbf{C}_{fn}^{-1} + \frac{1}{\sigma^2}\mathbf{J}\right)^{-1} \tag{7}$$

where $\mathbf{J}$ is a matrix of ones of size 2×2, $\mathbf{C}_{fn} = \mathrm{diag}(C_{j,fn})_{j=1,2}$ and

$$C_{j,fn} = \sum_{k=1}^{K_j} \mathbb{E}_{\mathbf{Z}\backslash s_j}\left[\frac{1}{w_{j,fk}h_{j,kn}}\right]^{-1} \tag{8}$$

The notation $\mathbb{E}_{\mathbf{Z}\backslash \mathbf{Z}_i}[.]$ denotes the expectation over all model parameters $\mathbf{Z}$ except $\mathbf{Z}_i$.

The variational distributions of the NMF parameters are identified to generalized inverse Gaussian (GIG) distributions [20] which are controlled by three parameters τ, ρ and γ. The updates of these parameters for the activation matrix $\mathbf{H}_j$ of source j are given by:

$$\boldsymbol{\tau}_{\mathbf{H}_j} = \mathbb{E}\left[\frac{1}{\mathbf{H}_j}\right]^{.2} \circ \left[\left(\mathbb{E}\left[\frac{1}{\mathbf{W}_j}\right]^{.-1}\right)^T \left(\mathbb{E}\left[|\mathbf{S}_j|^{.2}\right] \circ \mathbf{C}_j^{.-2}\right)\right]$$
$$\boldsymbol{\rho}_{\mathbf{H}_j} = b + \mathbb{E}\left[\mathbf{W}_j\right]^T \mathbb{E}\left[\mathbf{V}_j\right]^{.-1} \quad , \quad \gamma_{\mathbf{H}_j} = b \tag{9}$$

where the notation $\circ$ denotes the Hadamard product, $\mathbf{M}^{.x}$ and $\mathbf{M}^T$ respectively denote element-wise exponentiation and transposition of matrix $\mathbf{M}$. $\mathbf{C}_j$ is the matrix composed of the coefficients $C_{j,fn}$ defined in (8). Note that the same update rules can be found for $\mathbf{W}_j$ by replacing and reordering the terms accordingly.

In a speech enhancement context, the variable in which we are interested is the clean speech source $s_{1,fn}$. The STFT coefficients of the estimated sources are given by the mean $\boldsymbol{\mu}_{\mathbf{s},fn}$ of the posterior distribution $q(\mathbf{s}_{fn})$ in (7). For the speech source $s_{1,fn}$, this expectation simplifies to

$$\mu_{s_{1,fn}} = \frac{C_{1,fn}}{C_{1,fn} + C_{2,fn} + \sigma^2}\, x_{fn}. \tag{10}$$

We recognize the classical expression of the Wiener filter where deterministic estimates of the source power spectra have been replaced by the expectations $C_{j,fn}$.

2.3. Model averaging

We now assume that the above single-order NMF framework has been used with M different models. The model m is defined by its order denoted $\mathbf{K}_m = \{K_{1m}, K_{2m}\}$, in which K_{1m} (resp. K_{2m}) is the number of components of the speech source (resp. background noise source). The posterior distribution $q_m(\mathbf{Z})$ has thus been estimated for each model $m = 1..M$.

Bayesian model averaging [18] proposes to average these posterior probabilities with respect to the posterior probability $p(\mathbf{K}_m|\mathbf{X})$ of each model. Thanks to Bayes' rule, this posterior probability can be expressed as the product of the prior probability π_m of model m and its likelihood $p(\mathbf{X}|\mathbf{K}_m)$ so that

$$p(\mathbf{K}_m|\mathbf{X}) \propto \pi_m\, p(\mathbf{X}|\mathbf{K}_m). \tag{11}$$

As we have already highlighted, the computation of the likelihood $p(\mathbf{X}|\mathbf{K}_m)$ is intractable. However, the choice of a VB framework gives us the opportunity to replace it with the free energy expressed in (6) so that $p(\mathbf{X}|\mathbf{K}_m) \approx \exp(\mathcal{L}_m)$.

Applying model averaging to the M posterior distributions of the sources leads to the new estimate

$$q(\mathbf{s}_{fn}) = \frac{1}{\delta}\sum_{m=1}^{M} \pi_m e^{\mathcal{L}_m}\, q_m(\mathbf{s}_{fn}) \tag{12}$$

where $q_m(\mathbf{s}_{fn})$ is the source posterior distribution estimated for model m and $\delta = \sum_{m=1}^{M} \pi_m e^{\mathcal{L}_m}$ aims at normalizing the posterior probability of $\mathbf{K}_m$ so that it sums to 1. The STFT coefficients of the estimated speech source is now given by a linear combination of the expectations $\mu_{s_{1m,fn}}$ computed for each model m as in (10). This formulation is equivalent to the *temporal fusion by linear combination* introduced in [21] but differs from it in that the fusion coefficients now depend on the signal to be processed through the free energy $\mathcal{L}_m$.

3. Multiple-order NMF

Our novel multiple-order NMF model introduced here aims at jointly estimating and averaging several NMFs of different orders. For more details on the model and the derivation of the VB inference, the reader is referred to [24].

3.1. Probabilistic model

Contrary to the single-order NMF model of Section 2, the number of components K_j of source j is now seen as a random variable which follows a categorical distribution

$$K_j \sim \text{Cat}\left(\pi_{j1}, ..., \pi_{jm}, ..., \pi_{jM_j}\right) \tag{13}$$

where m indexes the M_j possible number of components $\left\{K_{j1}, ..., K_{jm}, ..., K_{jM_j}\right\}$, each having an *a priori* probability of π_{jm}. As above, we assume that each number of components has its specific NMF parameters so that in the following, they will be indexed with m such as $w_{jm,fk}$ and $h_{jm,kn}$. The priors on the NMF parameters and the sources remain unchanged in comparison with Section 2. The main difference is that a single posterior distribution of the sources is now estimated.

By denoting $\mathbf{K} = \{K_1, K_2\}$, $\mathbf{W} = \{\mathbf{W}_{jm}\}_{j=1,2}^{m=1..M_j}$, $\mathbf{H} = \{\mathbf{H}_{jm}\}_{j=1,2}^{m=1..M_j}$ and $p(\mathbf{K}) = p(K_1)p(K_2)$, the joint distribution becomes

$$p(\mathbf{X}, \mathbf{Z}, \mathbf{K}) = p(\mathbf{X}|\mathbf{S})\, p(\mathbf{S}|\mathbf{W}, \mathbf{H})\, p(\mathbf{W}|\mathbf{K})\, p(\mathbf{H}|\mathbf{K})\, p(\mathbf{K}).$$

3.2. Variational inference

In order to include the variables related to the numbers of components, the variational distribution of (5) is modified as follows

$$q(\mathbf{Z}, \mathbf{K}) = q(\mathbf{S})\, q(\mathbf{W}|\mathbf{K})\, q(\mathbf{H}|\mathbf{K})\, q(\mathbf{K}). \tag{14}$$

By minimizing the corresponding free energy, the posterior probability of the number of components $q(K_j)$ is obtained as

$$\log q(K_j) = \mathbb{E}_{\mathbf{Z}\backslash K_j}\left[\log p(\mathbf{X}, \mathbf{Z}, \mathbf{K})\right]$$
$$- \mathbb{E}_{\mathbf{Z}\backslash K_j}\left[\log q(\mathbf{W}_j|K_j)\right] - \mathbb{E}_{\mathbf{Z}\backslash K_j}\left[\log q(\mathbf{H}_j|K_j)\right] + \text{const.}$$

By developping, reordering and taking the exponential of these terms, the posterior probability of K_j can be formulated as

$$\forall m, \quad q(K_{jm}) \propto \pi_{jm}\, e^{\mathcal{L}_{jm}} \tag{15}$$

in which the term $\mathcal{L}_{jm} = \mathbb{E}_{\mathbf{Z}\backslash K_j}\left[\log \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\right]$ is similar to the free energy expressed in (6) in the single-order NMF case. The inference of the variational distributions over the NMF parameters remains unchanged in comparison to the single-order NMF model and the variational distribution of the sources is still identified as a bivariate Gaussian distribution with the parameters defined in (7). The only difference relies in the term $\mathbf{C}_{fn}^{-1}$ which is now a linear combination of terms related to each NMF order so that, with $\delta = \sum_{m=1}^{M_j} \pi_{jm} e^{\mathcal{L}_{jm}}$,

$$\forall j, \quad C_{j,fn}^{-1} = \frac{1}{\delta} \sum_{m=1}^{M_j} \pi_{jm} e^{\mathcal{L}_{jm}}\, C_{jm,fn}^{-1}. \tag{16}$$

4. Controlling the order posterior entropy

We introduced two NMF frameworks which both combine several NMFs of different orders. In Section 2, the scheme consists in computing several single-order NMFs and combining them after the variational inference thanks to the linear combination in (12). In Section 3, the scheme jointly estimates and average several NMFs according to (16). However, the averaging is similar in both cases as the averaging weights of (12) and (16) are of the form $\pi_m \exp(\mathcal{L}_m)$.

To enforce this comparison, we now assume as a particular case that the number of components of the background noise K_2 is fixed. In the single-order NMF framework, the averaging rule (12) remains unchanged. The model m is now entirely defined by the number of components K_{1m} of the speech source. In the multiple-order case, the number of components K_1 of the speech source is now the only one to be considered as a random variable and the average is only effective on the related NMF parameters and not on the background noise side. Hence, (16) only holds for $j = 1$ and by dropping the index j when unnecessary, it becomes $C_{1,fn}^{-1} = \frac{1}{\delta} \sum_{m=1}^{M} \pi_m e^{\mathcal{L}_m}\, C_{1m,fn}^{-1}$. As a consequence, $\mathcal{L}_m$ and π_m now have the exact same signification in both cases. π_m denotes the prior probability of model m of order K_{1m} whereas $\mathcal{L}_m$ denotes its free energy. Note however that in the multiple-order case, all speech models share the same background noise estimate whereas in the single-order case, one background noise model is estimated per order K_{1m}.

As stated in [21, 24], it is worth combining NMFs of different orders instead of selecting a unique order as it can improve the separation performance. However, our models, which apply model averaging in a straightforward way, are not achieving that goal. Indeed, preliminary tests have shown that the free energies $\mathcal{L}_m$ being very large valued, the posterior probabilities $q(K_{1m})$ are all equal to zero except the one which depends on the highest free energy and which is thus equal to 1. As such, model averaging turns out to be model selection. To avoid this behaviour, we propose to scale the free energies $\mathcal{L}_m$ by a factor $\beta \geq 1$ to be determined. Hence, the posterior probability of order K_{1m} becomes

$$q(K_{1m}) \propto \pi_m\, e^{\mathcal{L}_m/\beta}. \tag{17}$$

This is equivalent to controlling the entropy of the distribution $q(K_{1m})$ in a way similar to [25]. Small values of β will favor peaky distributions with one $q(K_{1m})$ close to 1 as in model selection, whereas higher values of β will result in a more uniform distribution. The determination of the values of β and π_m will be addressed in Section 5.

5. Experiments

In this section, we propose to evaluate and compare both the multiple-order NMF model and the model averaging of single-order NMFs. We also explore oracle and learning methods to determine the prior probabilities of the number of components π_m as well as the entropic parameter β. To do so, we rely on the PASCAL CHiME corpus [22] which features recordings of real domestic noise and speech utterances from diverse speakers.

5.1. Dataset

The CHiME corpus is divided in three datasets using 34 distinct speakers. Firstly, the training set is composed of 500 utterances of each speaker in reverberated conditions. This allows us to learn an NMF dictionary $\mathbf{W}_{1m}$ to describe each speaker by applying a single-order NMF on the concatenation of these utterances with a chosen number of components K_{1m}. All these learned dictionaries hence describe the same source but at different levels of details. The other datasets, namely the develop-

		-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Average	Average time (ms)
Baseline	VB selection	4.56	6.84	3.26	9.91	7.34	10.60	7.08	54.4
	Oracle selection	5.28	7.87	5.52	10.34	9.22	12.03	8.38	54.4
Oracle π_m **and** β	so-NMF averaging	5.80	8.16	5.55	10.34	9.79	12.34	8.66	54.4
	mo-NMF	5.18	8.11	4.65	11.14	9.70	11.07	8.31	13.3
Learned π_m **and** β	so-NMF averaging	5.47	7.81	4.52	10.54	9.44	11.26	8.17	54.4
	mo-NMF	5.25	7.80	4.37	10.99	9.63	11.30	8.22	13.4

Table 1:Average SDR (dB) and computation time for VB and oracle selections, simple-order NMF averaging and multiple-order NMF

ment and test sets, are both composed of 600 reverberated utterances. Each utterance, pronounced by one of the 34 speakers, has been mixed within a noise background so that the mixture of clean speech and noise achieved a given signal-to-noise ratio (SNR) amongst: -6, -3, 0, 3, 6 and 9 decibels (dB). The noise background signals come from real recordings of a domestic living room and the utterances have been placed at controlled time stamps according to the desired SNR. This allows us for each utterance to select 10 seconds of noise without speech, after and/or before the utterance, in order to learn an NMF dictionary $\mathbf{W}_2$ in a way similar to the learning of the NMF speaker models. We chose a single-order NMF with a fixed number of components $K_2 = 16$.

In the following, we have used the development set for learning purpose, whereas for evaluation, we have randomly selected in the test set a total of 24 utterances, *i.e.*, 4 by SNR. As the original data were two channel mixtures, we will work here on the mean of both channels to restrain our study to the single channel case. Both the single-order NMFs and the multiple-order NMF have been used with numbers of components $K_{1m} = 2^m$ with $m = 1..7$ for the speaker source. The dictionaries of noise and speech NMFs have been fixed to their learned values and the activation matrices have been initialized with the mean activation values estimated in the corresponding learning step. The hyperparameter of the Gamma prior related to the activation matrices has been fixed to $b = 0.2$ and sensor noise is assumed to be of variance $\sigma^2 = 10^{-6}$. Finally, the separation quality has been evaluated by the signal-to-distortion ratio (SDR) of the target speech, expressed in decibels [26].

5.2. Learning the priors and the entropic parameter

In order to average the NMFs, we need to determine the prior probabilities π_m of each number of components as well as the entropic parameter β. To do so, we propose two approaches. In particular, we propose to learn both π_m and β thanks to the development set. Indeed, as the original reverberated signals are available, we can evaluate the performance of the single-order NMF scheme on each mixture l of the development dataset and for each number of components $K_{1m} = 2^m$. We can thus find the set of π_m and β which maximizes the mean SDR of the speech signals over the L examples of the development set. This is equivalent to solving the non-linear optimization problem

$$\underset{(\pi_1,\ldots,\pi_7,\beta)}{\operatorname{argmin}} \sum_{l,fn} \left|\left|\mu^{(l)}_{s_{1,fn}} - \tilde{s}^{(l)}_{1,fn}\right|\right|^2 \qquad (18)$$

in which $\tilde{s}^{(l)}_{1,fn}$ is the original reverberated speech signal of example l, $\mu^{(l)}_{s_{1,fn}} = \sum_m \pi_m \exp(\mathcal{L}_m/\beta)\mu^{(l)}_{s_{1m,fn}}$ and $\mu^{(l)}_{s_{1m,fn}}$ is defined for example l as in (10) for the number of components K_{1m}. Note that such a learning is impossible in the multiple-order NMF case as the averaging rule (16) is computed at each iteration of the inference algorithm. However, these learned values of π_m and β can be used in both the averaging of single-order NMFs and the multiple-order NMF model. For computational convenience, we have restrained the size of the development set for learning. We thus propose to learn the parameters on $L = 36$ randomly picked examples, *i.e.*, 6 by SNR.

In order to evaluate the performance of the learned π_m and β, we can also solve the optimization problem (18) for the example of the test set being processed. However, it is worth noting that such a result is unreachable in practice and is thus denoted as the *oracle* result. As in the learning case, oracle π_m and β will be used in both the single-order and multiple-order schemes.

5.3. Results and comments

Table 1 shows the results of our study grouped in three categories. The SDRs are averaged for each SNR over the 4 selected examples as well as over the whole selected test set. The baseline is given by the VB selection result, which consists in selecting for each example the single-order model that has the highest free energy $\mathcal{L}_m$, and the oracle selection result which, on the basis of the original reverberated speech signal, consists in selecting for each example the single-order model which gives the best SDR. Note that VB selection is equivalent to the averaging of single-order NMFs without using the entropic parameter we proposed in Section 4. The second group of results is based on the oracle π_m and β values whereas the third group is based on the learned π_m and β values, as explained in Section 5.2. For each example of the test set, these oracle and learned parameters have been used for both the single-order NMF (so-NMF) averaging and the multiple-order NMF (mo-NMF).

These results first show that regardless of the SNR, the VB selection always fails to select the best model in term of SDR. In average, the VB selection underperforms oracle selection by 1.3 dB. Oracle results show that the averaging of single-order NMFs can always outperform both VB and oracle selection. The multiple-order NMF using these same oracle parameters implies a loss of only 0.35 dB in comparison with the single-order averaging. However, this loss is somehow counterbalanced by an interesting gain in computation time as mo-NMF is 4 times faster than the averaging of so-NMFs.

Recalling that oracle methods are not reachable in practice, the results based on learned π_m and β show an interesting performance. Indeed, both single-order and multiple-order methods outperform VB selection which is the only other method to be practicable. Moreover, both methods nearly reach the SDR of the oracle selection. Our proposed method to learn π_m and β is thus efficient as it allows to improve separation performance by almost 1.1 dB. We can also notice that in realistic conditions, the multiple-order NMF gives similar SDRs to the averaging of single-order NMFs but it is far less computationally demanding.

6. Conclusion

We introduced two different NMF frameworks which aim at combining several NMFs of different orders: a single-order model scheme which applies Bayesian model averaging and a novel multiple-order model which jointly estimates and averages several NMFs. The parameter we introduced to control the entropy of the order posterior has been shown to be essential in order to make the averaging effective in both models. Finally, experimental results on real data showed that both models outperform traditional VB selection, the multiple-order NMF being furthermore less computationally demanding. Future works will focus on the study of a time-varying model averaging.

7. References

[1] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*, Springer, 2005.

[2] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[4] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[5] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

[6] J. Hao, H. Attias, S. Nagarajan, T.-W. Lee, and T. J. Sejnowski, "Speech enhancement, gain, and noise spectrum adaptation using approximate Bayesian estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 24–37, 2009.

[7] J. F. Gemmeke, A. Hurmalainen, and T. Virtanen, "HMM-regularization for NMF-based noise robust ASR," *Proc. of CHiME-2013*, pp. 47–52, 2013.

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

[9] P. Mowlaee, R. Saeidi, Mads G. Christensen, Z.-H. Tan, T. Kinnunen, P. Franti, and S. H. Jensen, "A joint approach for single-channel speaker identification and speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2586–2601, 2012.

[10] B. Raj, R. Singh, and T. Virtanen, "Phoneme-dependent NMF for speech enhancement in monaural mixtures," in *Interspeech*, 2011, pp. 1217–1220.

[11] N. Mohammadiha, J. Taghia, and A. Leijon, "Single channel speech enhancement using Bayesian NMF with recursive temporal updates of prior distributions," in *Proc. of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2012, pp. 4561–4564.

[12] N. Moritz, M. R. Schädler, K. Adiloǧlu, B. T. Meyer, T. Jürgens, T. Gerkmann, B. Kollmeier, S. Doclo, and S. Goetze, "Noise robust distant automatic speech recognition utilizing NMF based source separation and auditory feature extraction," *Proc. of CHiME-2013*, pp. 1–6, 2013.

[13] J. T. Geiger, F. Weninger, A. Hurmalainen, J. F. Gemmeke, M. Wöllmer, B. Schuller, G. Rigoll, and T. Virtanen, "The TUM + TUT + KUL approach to the 2nd CHiME challenge: Multi-stream ASR exploiting BLSTM networks and sparse NMF," *Proc. of CHiME-2013*, pp. 25–30, 2013.

[14] E. Vincent, J. Barker, Sh. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: An overview of challenge systems and outcomes," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 162–167.

[15] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," in *Proc. of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2007, vol. 1, pp. I–65–68.

[16] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2008, pp. 1825–1828.

[17] H. Attias, "A variational Bayesian framework for graphical models," *Advances in neural information processing systems*, vol. 12, no. 1-2, pp. 209–215, 2000.

[18] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: a tutorial," *Statistical science*, pp. 382–401, 1999.

[19] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, 2009, Article ID 785152.

[20] M. Hoffman, D. M. Blei, and P. R. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *Proc. of International Conference on Machine Learning (ICML)*, 2010, pp. 439–446.

[21] X. Jaureguiberry, G. Richard, P. Leveau, R. Hennequin, and E. Vincent, "Introducing a simple fusion framework for audio source separation," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2013, pp. 1–6.

[22] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: datasets, tasks and baselines," in *Proc. of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2013, pp. 126–130.

[23] K. Adiloǧlu and E. Vincent, "Variational Bayesian inference for source separation and robust feature extraction," Tech. Rep. RT-0428, Inria, 2012.

[24] X. Jaureguiberry, E. Vincent, and G. Richard, "Variational Bayesian model averaging for audio source separation," in *Proc. of IEEE Statistical Signal Processing Workshop (SSP)*, 2014, pp. 33–36.

[25] K. Katahira, K. Watanabe, and M. Okada, "Deterministic annealing variant of variational Bayes method," *Journal of Physics: Conference Series*, vol. 95, no. 1, 2008.

[26] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.