Bitwise Source Separation on Hashed Spectra: An Efficient Posterior Estimation Scheme Using Partial Rank Order Metrics

Lijiang Guo
Dept. of Intelligent Systems Engineering
Indiana University
Bloomington, IN 47401
lijguo@indiana.edu

Minje Kim[*]
Dept. of Intelligent Systems Engineering
Indiana University
Bloomington, IN 47401
minje@indiana.edu

Abstract

This paper proposes an efficient bitwise solution to the single-channel source separation task. Most dictionary-based source separation algorithms rely on iterative update rules during the run time, which becomes computationally costly especially when we employ an overcomplete dictionary and sparse encoding that tend to give better separation results. To avoid such cost we propose a bitwise scheme on hashed spectra that leads to an efficient posterior probability calculation. For each source, the algorithm uses a partial rank order metric to extract robust features that form a binarized dictionary of hashed spectra. Then, for a mixture spectrum, its hash code is compared with each source's hashed dictionary in one pass. This simple voting-based dictionary search allows a fast and iteration-free estimation of ratio masking at each bin of a signal spectrogram. We verify that the proposed BitWise Source Separation (BWSS) algorithm produces sensible source separation results for the single-channel speech denoising task, with 6-8 dB mean SDR. To our knowledge, this is the first dictionary based algorithm for this task that is completely iteration-free in both training and testing.

1 Introduction

The single-channel source separation problem has been widely studied as a latent variable model. The most common practice is to learn a source-specific dictionary from each source during training so that the source spectra can be reconstructed by a linear combination of the dictionary items. Nonnegative Matrix Factorization (NMF) (Lee & Seung, 1999, 2001; Raj & Smaragdis, 2005) and Probabilistic Latent Semantic Indexing (PLSI) (Hofmann, 1999a,b; Smaragdis et al., 2007) are popular choices for the modeling job. Meanwhile, a large overcomplete dictionary is another preferable option to preserve the manifold structure of the source spectra. It can be either learned by a manifold preserving quantization technique (Kim & Smaragdis, 2013) or simply using the entire source spectra directly (Smaragdis et al., 2009; Kim et al., 2015). As these approaches are based on an iterative algorithm to estimate the activation, a practical source separation system needs to be careful about the necessary resources. In this paper we propose a fully BitWise Source Separation (BWSS) scheme, where the dictionary search is done entirely among the hash codes.

[*]This work was supported by Intel Corporation.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2 The Proposed Bitwise Source Separation

2.1 Voting-based Likelihood Estimation: A Fast Dictionary Search in the Hash Code Space

We propose a nonparametric algorithm for estimating the posterior probability of a signal being one of two sources. To this end, we first calculate the likelihood of observing a time-frequency bin given one of the sources, but based on a simple vote-counting method by finding matches between hashed spectra. This algorithm works on two preprocessed dictionaries of clean speech and noise. For a new mixture spectra, the algorithm scans the two dictionaries to generate a mixture distribution of speech and noise, which can be then used to calculate the posterior probability of one of the sources given the time-frequency bin as in (1).

For example, suppose there are two sources which contribute to the observed mixture. We denote the magnitude spectrogram of the mixture signal by $\mathbf{X}$, a $F \times T$ matrix where F and T being the number of frequencies and frames, resepctively. We first use a partial rank order metric as described in Yagnik et al. (2011) to generate L integer embeddings of each column vector $\mathbf{X}_{:,t}$, call each $\mathcal{X}_{\ell,i}, \ell \in \{1, \cdots, L\}$. Let $\mathcal{X}$ denote the $L \times T$ embedding matrix of $\mathbf{X}$. For the dictionaries we use the same procedure to generate their embedding matrices $\mathcal{S}$ and $\mathcal{N}$, respectively with dimension $L \times T_S$ and $L \times T_N$.

For each element $\mathcal{X}_{\ell,t}$, we search $\mathcal{S}_{\ell,:}$ and $\mathcal{N}_{\ell,:}$ to count the number of matches with each dictionary in the ℓ^{th} permutation sample, call it $\mathsf{S}_{\ell,t}$ and $\mathsf{N}_{\ell,t}$. Recall $\mathcal{X}_{\ell,t}$ is the index of the winning element in the ℓ^{th} permutation sample of $\mathbf{X}_{:,t}$. Combining $\mathcal{X}_{\ell,t}$ and θ_ℓ we are able to track back to the corresponding original frequency bin $j = i^{\ell}_{\mathcal{X}_{\ell,t}}$, the true winner of θ_ℓ for $\mathbf{X}_{:,t}$. Thus the total counts of matches with each dictionary that are possibly spread in L slots of $\mathcal{S}_{:,t}$ and $\mathcal{N}_{:,t}$ are defined as follows, respectively:$\bar{\mathsf{S}}_{j,t} = \sum_\ell \mathsf{S}_{\ell,t}, \quad \bar{\mathsf{N}}_{j,t} = \sum_\ell \mathsf{N}_{\ell,t}$.

The total counts $\bar{\mathsf{S}}_{j,t}$ and $\bar{\mathsf{N}}_{j,t}$ approximate the similarity of $\mathbf{X}_{j,t}$ to the two sources, respectively. Therefore, they also approximate the likelihood of observing $\mathbf{X_{j,t}}$ given one of the sources. In ℓ^{th} permutation sample that $\mathbf{X}_{j,t}$ has won, it is greater than the rest $K-1$ frequencies in $\mathbf{X}_{i_1^\ell:i_K^\ell,t}$. Because we encode the rank order of only K partial dimensions, the same relationship can be likely to be found in one of the source dictionaries more than in the other. Therefore, for $\mathcal{S}_{j,t'}$ to win in the same ℓ^{th} permutation sample, $\mathcal{S}_{j,t'}$ must be greater than the rest $K-1$ frequencies in $\mathcal{S}_{i_1^\ell:i_K^\ell,t'}$. As we discussed earlier, this is a binarized cosine similarity.

2.2 Estimation of the Posterior Probability

Once we calculate the likelihoods in the form of the number of partial matches to the two dictionaries as in section 2.1, the rest of the job is to compute the posterior probabilities over the sources given the mixture spectrogram. In the proposed BWSS system, we escape from the EM iterations, but instead propose a bitwise method to estimate the posterior probabilities.

Let $\mathbf{Y}_{j,t}$ denote a Bernoulli random variable where 0 is clean speech and 1 is noise. Thus the likelihood of observing $\mathbf{X}_{j,t}$ is $P(\mathbf{X}_{j,t}) = \sum_{\mathbf{Y}_{j,t}=\{0,1\}} P(\mathbf{Y}_{j,t}) P(\mathbf{X}_{j,t}|\mathbf{Y}_{j,t})$. We define a prior distribution on $\mathbf{Y}_{j,t}$ with a Bernoulli distribution with $p = 0.5$ to give a fair chance to both sources. Another assumption is that each frequency bin is independent of all the other bins in a different time frame, while it is dependent on the other frequency bins in the same time frame due to the rank ordering during hashing.

To adjust for the difference in the number of frames of the two dictionaries, we normalize the count of matches accordingly. Finally, the posterior probability for a given time-frequency bin is:

$$P(\mathbf{Y}_{j,t} = 0|\mathbf{X}_{j,t}, \mathcal{S}, \mathcal{N}) = \frac{\bar{\mathsf{S}}_{j,t}}{\bar{\mathsf{S}}_{j,t} + \bar{\mathsf{N}}_{j,t} \cdot r}, \quad P(\mathbf{Y}_{j,t} = 1|\mathbf{X}_{j,t}, \mathcal{S}, \mathcal{N}) = \frac{\bar{\mathsf{N}}_{j,t} \cdot r}{\bar{\mathsf{S}}_{j,t} + \bar{\mathsf{N}}_{j,t} \cdot r}, \qquad (1)$$

where $r = T_S/T_N$. Recall $\bar{\mathsf{S}}_{j,t}$ and $\bar{\mathsf{N}}_{j,t}$ are counts of matches with clean speech and noise dictionaries for a given frequency bin $\mathbf{X}_{j,t}$. For $\bar{\mathsf{S}}_{j,t}$, it is the number of votes on the clean speech dictionary for $\mathbf{X}_{j,t}$ based on all the permutation samples that $\mathbf{X}_{j,t}$ has been involved in comparison and won; similarly $\bar{\mathsf{N}}_{j,t}$ corresponds to the number of votes that $\mathbf{X}_{j,t}$ received from the noise dictionary. Thus, $P(\mathbf{Y}_{j,t} = 1|\mathbf{X}, \mathcal{S}, \mathcal{N})$ reflects the proportion of votes from clean speech dictionary for a frequency.

3 Experiments

3.1 The Data Set

TIMIT training set contains 136 female and 326 male speakers, while the testing set contains 56 female and 112 male speakers, which are from eight dialect regions in the US. Each TIMIT speaker has 10 short utterances. TSP dataset has over 1400 short utterances from 25 speakers. We downsample the TSP dataset, so that all signals are with a 16kHz sampling rate. We mix each test utterance with 10 kinds of noises as proposed in Duan et al. (2012). These noises are: 1. birds, 2. casino, 3. cicadas, 4. computer keyboard, 5. eating chips, 6. frogs, 7. jungle, 8. machine guns, 9. motorcycles, and 10. ocean. Short-Time-Fourier-Transform (STFT) with a Hann window of 1024 samples and a hop size of 256 transforms the signals. To evaluate the final results, we used Signal-to-Distortion Ratio (SDR) as an overall source separation measurement along with Signal-to-Interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR) Vincent et al. (2006), and Short-Time Objective Intelligibility (STOI) Taal et al. (2010).

3.2 Experiment Design

In our experiments, we first construct the hash code dictionaries, which yields a set of clean speech dictionaries and 10 noise dictionaries. During source separation, an unseen noisy utterance is processed using the corresponding clean speech dictionary and a noise dictionary of the same noise type. Since the noise type is known, we vary between known and unknown speaker identity to perform supervised and semi-supervised separation.

Our algorithm has three parameters K, L, and r, and there is no clear guideline in choosing their values. As in dictionary based source separation, Duan et al. (2012) and Sun & Mysore (2013) empirically choose the parameters for number of NMF or PLCA basis vectors. For BKL-NMF Sun & Mysore (2013), there is an additional regularization parameter λ . We take similar approach to searches for the optimal parameter combination for each testing case. Further details are discussed in section 3.3.

Experiment #1 Speaker-dependent supervised separation with small dictionaries: For each TIMIT speaker we use the first 9 out of 10 utterances to create a speaker-specific speech dictionary. By mixing the 10th one with 10 noise types with 0 dB Signal-to-Noise Ratio (SNR) we get 462×10 noisy test utterances for 462 speakers. Supervised separation was done by assuming the noise type and the identity of the speaker are known.

Experiment #2 Speaker-dependent supervised separation with large dictionaries: We suspect that a larger speech dictionary better represents the speaker. For this we use the TSP dataset with roughly 56 utterances per speaker. For each speaker, we once again hold out one last utterance for testing and build the dictionary from the rest. This gives us $25 \times 10 = 250$ noisy utterances. Supervised separation is done as in Experiment #1.

Table 1: Experiment Results (dB)

	SDR	**SIR**	**SAR**	**STOI**
BWSS Experiment #1 supervised (TIMIT)				
Male	6.6433	8.7644	9.1727	0.0077
Female	6.7548	8.8735	9.3551	0.0063
Male and Female	6.6761	8.7965	9.2264	0.0073
BWSS Experiment #2 supervised (TSP)				
Male and Female	6.9898	9.3538	9.6739	0.0128
BWSS Experiment #3 semi-supervised (TIMIT)				
Male	7.4213	9.7257	9.2759	-0.0104
Female	7.5271	10.343	9.4262	-0.0050
Male and Female	**7.4742**	10.035	9.351	-0.0077
KL-NMF (TIMIT) Sun & Mysore (2013)				
Male (supervised)	10.23	-	-	-
Male (semi-supervised)	**7.22**	-	-	-
BKL-NMF+USM (TIMIT) Sun & Mysore (2013)				
Male (supervised)	10.41	-	-	-
Male (semi-supervised)	6.23	-	-	-
Online PLCA (NOIZEUS) Duan et al. (2012)				
Male and Female	**6.180**	11.710	8.450	-

Experiment #3 Pooled-speaker semi-supervised separation: We apply BWSS in a semi-supervised setting where speaker identity is unknown during separation, while the gender is known. From each dialect of TIMIT training set we select 4

males and 4 females. This gives
us $4 \times 8 \times 10 = 320$ clean utterances for each gender, which are pooled into one clean speech dictionary for each gender. For testing, we select 1 male and 1 female from each dialect of TIMIT testing set and mix their utterances with 10 noises to create $2 \times 8 \times 10 \times 10 = 1600$ noisy utterances. During separation, we use the clean speech dictionary of same gender and the noise dictionary of same noise type, which is still a semi-supervised separation with unknown speaker identity.

3.3 Separation Results

- ***Variations in parameters***: There are three model parameters in the BWSS algorithm, L, K, and r. L is the number of permutation samples to be drawn from a time frame $\mathbf{X}_{:,t}$. As L goes to ∞, the sample posterior probability will converge to the true mixing distribution. In our experiment we found the algorithm approximates stable posterior probability quickly as we increase L. For $L = 2F$ the result is already very close to $L = 8F$. K is the size of each permutation sample. More random samples means the distribution of $\mathbf{X}_{:,t}$ is more exploited, and the better approximation to the posterior probability of each frequency bin $\mathbf{X}_{j,t}$. However, it is not always guaranteed, because a too large K value can break down the locality of the comparison process. The relative sizes of clean speech dictionary $\mathcal{S}$ and noise dictionary $\mathcal{N}$ are compensated by the parameter r. This is because of the possibility that a larger dictionary with more repeating training samples can exaggerate the number of matches for that source. However, note that because of the other chance that the dictionary is indeed with many unique items, the choice of r is not always related to good separation.
- ***Size of clean speech dictionary***: In a fully supervised setting, both the speaker identity and noise type are known, and the proposed algorithm achieves 6.68 dB mean SDR on TIMIT dataset (Exp. #1) and 6.99 dB mean SDR on TSP dataset (Exp. #2), as shown in Table 1. Although in Exp. #2, each clean speech dictionary has roughly 5 times more speakers than Exp. #1, we see the performance gap is not very large. Therefore, we conclude that the BWSS algorithm works reasonably well on small, but quality speech dictionaries.
- ***Speaker identity and pooled dictionary***: We notice that knowing speaker identity does not provide significant improvement in separation performance compared to the semi-supervised setting with a very large dictionary. For Exp. #3 the test speaker is unseen, but noise type is known in advance, where the proposed algorithm achieves a mean SDR of 7.47 dB which is 0.80 dB higher than in supervised setting (Exp. #1). Note that this gender-specific dictionary of 32 speakers is larger than USM's 20 speaker model Sun & Mysore (2013), and would be computationally demanding to handle if it were not for the proposed bitwise mechanism.
- ***Comparison with other dictionary based methods***: In Table 1, we include results of two NMF-based methods reported in Sun & Mysore (2013) and one PLSI-based method reported in Duan et al. (2012) in addition to BWSS results. All these experiments use the same 10 noise types. In Sun & Mysore (2013) 20 male speakers from TIMIT were used as training set, learning 10 basis vectors from each speaker. In Duan et al. (2012) 3 male and 3 female speakers from NOIZEUS formed the training set. The proposed algorithm achieves competitive results using hashed spectra, so that it can employ large training data in a memory-saving manner. Also, its iteration free separation is a plus for the run-time efficiency.

Since the experimental setup is different, a fair comparison is not possible to those existing methods, but we can still gauge the performance of BWSS. For example, a completely supervised model where both the speaker identity and noise type is known (KL-NMF supervised), the usual KL-NMF performs very well (10.23 versus 6.99 dB in Exp. #2). For the semi-supervised case using KL-NMF, a direct comparison is not possible because it assumes unknown noise, while in Exp. #3 we assumed anonymity of speakers (7.22 versus 7.47 dB). USM catches up the performance by introducing a larger dictionary and the block sparsity as regularizer, whose supervised case loosely corresponds to Exp. #3 (10.41 versus 7.47 dB).

Another comparison would be with Duan et al. (2012), where an online PLCA algorithm was proposed, but tested with a different speech dataset. For a rough comparison, in all three BWSS experiments, we obtained better SDR and SAR than online PLCA, but marginally worse SIR.

From this comparison, we see that BWSS does not outperform the existing dictionary-based algorithm. Also note that the STOI improvement of BWSS results are not very impressive. However, BWSS performs reasonably well given its low operational cost thanks to its bitwise operations. For example, BWSS can be a viable solution in an extreme environment with little resource. Or, it can be used to better initialize the full NMF/PLCA-based models.

References

Duan, Zhiyao, Mysore, Gautham J, and Smaragdis, Paris. Online plca for real-time semi-supervised source separation. In *International Conference on Latent Variable Analysis and Signal Separation*, pp. 34–41. Springer, 2012.

Hofmann, Thomas. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57. ACM, 1999a.

Hofmann, Thomas. Probablistic latent semantic analysis. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*, 1999b.

Kim, Minje and Smaragdis, Paris. Manifold preserving hierarchical topic models for quantization and approximation. In *International Conference on Machine Learning*, pp. 1373–1381, 2013.

Kim, Minje, Smaragdis, Paris, and Mysore, Gautham J. Efficient manifold preserving audio source separation using locality sensitive hashing. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 479–483. IEEE, 2015.

Lee, Daniel D and Seung, H Sebastian. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Lee, Daniel D and Seung, H Sebastian. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pp. 556–562, 2001.

Raj, Bhiksha and Smaragdis, Paris. Latent variable decomposition of spectrograms for single channel speaker separation. In *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, pp. 17–20. IEEE, 2005.

Smaragdis, Paris, Raj, Bhiksha, and Shashanka, Madhusudana. Supervised and semi-supervised separation of sounds from single-channel mixtures. *Independent Component Analysis and Signal Separation*, pp. 414–421, 2007.

Smaragdis, Paris, Shashanka, Madhusudana, and Raj, Bhiksha. A sparse non-parametric approach for single channel separation of known sounds. In *Advances in neural information processing systems*, pp. 1705–1713, 2009.

Sun, Dennis L and Mysore, Gautham J. Universal speech models for speaker independent single channel source separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 141–145. IEEE, 2013.

Taal, Cees H, Hendriks, Richard C, Heusdens, Richard, and Jensen, Jesper. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4214–4217. IEEE, 2010.

Vincent, Emmanuel, Gribonval, Rémi, and Févotte, Cédric. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4): 1462–1469, 2006.

Yagnik, Jay, Strelow, Dennis, Ross, David A, and Lin, Ruei-sung. The power of comparative reasoning. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 2431–2438. IEEE, 2011.