EFFICIENT NEIGHBORHOOD-BASED TOPIC MODELING FOR COLLABORATIVE AUDIO ENHANCEMENT ON MASSIVE CROWDSOURCED RECORDINGS

Minje Kim[*]

University of Illinois at Urbana-Champaign
Department of Computer Science
minje@illinois.edu

Paris Smaragdis

University of Illinois at Urbana-Champaign,
Adobe Research
paris@illinois.edu

ABSTRACT

Collaborative Audio Enhancement (CAE) aims at separating a dominant source from crowdsourced recordings of a scene. This paper proposes a CAE setup as a big ad-hoc microphone array problem, assuming hundreds of sensors scattered over a large scene, e.g. a concert hall or a street riot. An important characteristic in such cases is the fact that not all sensors capture useful information, mainly because of the existence of strong local noise interferences and recording artifacts. This renders traditional array processing techniques inadequate for tasks such as source enhancement. One way to recover the most common source while suppressing recording-specific interference, is to share latent components across simultaneous models on multiple magnitude spectrograms. The proposed method improves on the quality and the computational requirements of such a model by using a two-stage nearest-neighborhood search at every EM update. Its optional first-round search uses Hamming distance between hashed spectrograms to quickly find a redundant candidate set, and then a subsequent step narrows the set down to a subset using more appropriate cross entropy. Experimental results show that the proposed neighborhood schemes converge to the better quality solutions faster than the comprehensive model using all data.

Index Terms— Probabilistic Topic Models, Probabilistic Latent Component Sharing, Ad-hoc Microphone Array, Collaborative Audio Enhancement, Social Data

1. INTRODUCTION

Collaborative Audio Enhancement (CAE) tries to extract the most significant source out of a set of noisy observations, *crowdsourced recordings*, potentially collected from socially shared data. In the CAE scenario, we assume that each observed recording can be uniquely contaminated, e.g. by an artifact from aggressive audio coding, an interfering sound captured only by that sensor, a bad frequency response of the microphone, clipping, etc [1]. Such recordings can also be thought of as signals from an ad-hoc microphone array in the sense that the channels are not synchronized and the sensor locations and characteristics are unknown [2, 3].

One challenge is to synchronize such a large set of signals. An effective approach is to assume a calibration signal in all the recordings as a guide to align with [2]. Another more realistic approach extracts noise-robust landmarks from audio spectrograms to identify videos [4], or to synchronize audio-visual signals [5], where robust matching is done efficiently using integer operations. In this paper we assume that all signals are already aligned using one of the aforementioned methods, and instead we focus on the enhancement part.

Another challenge is to recover the geometric information. A closed form solution for the sensor location estimation problem using signals contaminated with Gaussian noise was proposed in [2]. A beamforming technique for an ad-hoc microphone array was also proposed in [3] in the presence of localized noise sources, while the clustering-based calibration of sensors does not scale up well to the much bigger and heterogeneous array setup that we assume in this paper. More recently, an ad-hoc array calibration method was proposed [6], which enhances a noisy distance matrix using a low-rank approximation. The matrix completion is effectively done by Nonnegative Matrix Factorization (NMF) [7, 8] particularly if some elements in the distance matrix are missing. However, it was evaluated when at least more than half of the locations are known, and the source separation performance was not examined.

Probabilistic Latent Component Sharing (PLCS) allows some topics, or equivalently latent components in the context of the other latent variable models, to share same parameters during simultaneous latent variable modeling on the synchronized magnitude spectrograms [1], as a probabilistic version of Nonnegative Matrix Partial Co-Factorization [9]. Since CAE assumes that all recordings have captured some elements of the most common source, PLCS estimates the dominant source as a mixture of shared components, while the unshared components at each recording-specific topic model are used to explain the unique interferences that are to be discarded. Unfortunately, PLCS does not scale satisfactorily so as to allow analysis of a large number of input recordings: the complexity linearly increases as the number of recordings gets bigger. Furthermore, more social data is not always helpful, because poor recordings at locations far away from the source of interest may include a great deal of interferences and not contribute to the reconstruction of the desired source significantly.

We propose a streamlined PLCS algorithm that adaptively uses a subset of the available recordings rather than the whole set. The selection is done at every EM update based on cross entropy between the normalized magnitude spectrograms of the dominant source estimation and the noisy observations. Additionally, we show how to employ hashing to improve the complexity of the requisite search. If Hamming distance between hashed spectrograms is correlated to the original distance, it can reduce the search space by providing with a candidate set, followed by the second search with the full metric only on those candidates. The proposed methods indeed give better performances in simulated CAE tasks by selectively utilizing crowdsourced data made up from up to a thousand individual recordings. Finally, the proposed methods converge to a better solution faster.

[*]This material is based upon work supported by the National Science Foundation under Grant No. 1319708.

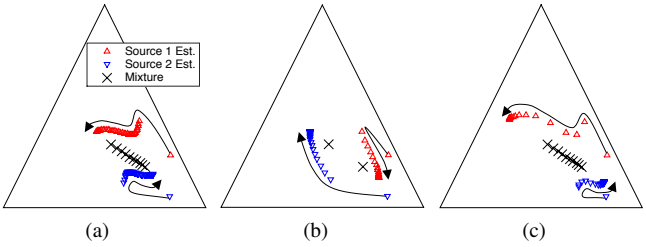


Fig. 1: PLSI topic modeling with various conditions: (a) Ordinary PLSI (b) An oracle PLSI only on the data samples that are closest to the optimal solutions (c) PLSI updates only on the running nearest neighbors to the current estimates of the sources. In the figure, all the source estimates are shown from every iteration, and their order in time is represented with a curved arrow. All parameters start from the same position for comparison.

2. NEIGHBORHOOD-BASED TOPIC MODELING

Probabilistic Latent Semantic Indexing (PLSI) [10, 11] has been widely used in audio research, and it has been known that sparse overcomplete coding is beneficial for the separation performance, since sparse coding on the training spectra leads to a tighter convex hull for the source [12][1]. In this paper we do not specifically focus on the sparseness of encoding, but it is also true for the proposed model that the data samples that are closer to the corners of the convex hull (or the normalized basis vectors of NMF) contribute more to the parameter estimation. For example, Fig. 1 (a) shows that the ordinary PLSI learns a subspace that reconstructs the mixture samples as convex combination of the two converged sources. In other words, the mixture samples should lie on or near the line that connects the learned sources for a quality approximation. In (b) we do the same PLSI modeling, but only on the two mixture points closest to the optimal solutions. This is an oracle run, because usually we cannot know the optimal solution in advance. What we can see is that the two samples are just enough to recover the same subspace with a permutation ambiguity.

An alternative to the oracle scenario is to find the nearest neighbors at every iteration given a set of source estimates. Fig. 1 (c) demonstrates this case. Once again, the source estimates converge to the same subspace as in the ordinary PLSI and the one with an oracle minimal set of inputs. Formally, we can intervene in the EM updates of original PLSI by forming a tentative set of neighbors so that the computation is done only on the set as follows:

$$P(f|z) \leftarrow \frac{P(f|z)\sum_{t\in\mathcal{N}_z} V_{f,t}P_t(z)}{\sum_z P(f|z)P_t(z)}, P(f|z) \leftarrow \frac{P(f|z)}{\sum_f P(f|z)}, \quad (1)$$

$$P_t(z) \leftarrow \frac{P_t(z)\sum_f V_{f,t}P(f|z)}{\sum_z P(f|z)P_t(z)}, P_t(z) \leftarrow \frac{P_t(z)}{\sum_z P_t(z)}, \quad (2)$$

where the M-step updates are equivalently reformulated to include the E-step, so that the posterior probabilities are updated more often. In (1) we do the summation only over the nearest neighbor set $\mathcal{N}_z$, which is a subset of all spectra whose elements are always closer to the z-th basis vector $P(f|z)$ than the non-neighbors as follows:

$$-\sum_f \hat{V}_{f,t\in\mathcal{N}_z} \log P(f|z) < -\sum_f \hat{V}_{f,t'\notin\mathcal{N}_z} \log P(f|z), \quad (3)$$

[1]NMF with KL-divergence as the error function is also known to be equivalent to PLSI with a proper parameter normalization [13]. Therefore, the proposed methods can be directly used in NMF-based systems as well.

where the column vectors of $\hat{V}$ are normalized, i.e. $\sum_f \hat{V}_{f,t} = 1$, for the proper calculation of cross entropy. The set should be small enough to effectively represent the corners of the convex hull, similarly to the locality preservation issues in manifold learning, such as in [14]. We refresh $\mathcal{N}_z$ according to the new $P(f|z)$ after every M-step in (1).

Note that if this set is optimal from the start and does not change, the results will be similar to Fig. 1 (b). If $\mathcal{N}_z$ always includes all data samples, then the updates in (1) and (2) are equivalent to the ordinary PLSI's by resulting in Fig. 1 (a).

3. NEIGHBORHOOD-BASED PROBABILISTIC LATENT COMPONENT SHARING

3.1. Neighborhood-Based Extension

Neighborhood-based extensions of a probabilistic topic model were proposed in [15, 16] for manifold preserving source separation. However, those models are designed to find out a sparse code from an overcomplete dictionary, where the dictionary is a large collection of clean source spectra. Therefore, the search is on the clean spectra, not on the noisy spectrograms.

In this section we propose a new PLCS model employing a neighbor search on the spectrograms. At every EM iteration, we update the parameters by analysing only a subset of the recordings. The subset can be also refreshed every time according to the distances between the recordings and the new source estimation. In the CAE scenario, the source estimate is not just a component anymore, but a convex combination of the shared components across the multiple simultaneous probabilistic topic models.

First, for the l-th recording, we assume that its magnitude spectrogram is generated from a distribution that can be decomposed into the common and individual topics:

$$V^l_{f,t} \sim \sum_{z\in z_C} P_C(f|z)P_C(t|z)P^l(z) + \sum_{z\in z^l_I} P^l_I(f|z)P^l_I(t|z)P^l(z), \quad (4)$$

where the parameters with C as subscripts, but with no superscript, are common across all models, while the subscript I stands for the recording-specific individual parameters along with a superscript l to distinguish recordings. We use a symmetric PLSI version [11] in which the temporal encoding $P(t|z)$ represents probabilities over time, and consequently requiring an addition weight $P(z)$ that governs the global activation of the component. The global weights are grouped into two, i.e. z_C and z^l_I, which denote the sets of indices for the common and individual components, respectively.

The parameters are updated with the EM algorithm as in [1], but this time we introduce the neighborhood concept as in Section 2. Therefore, the updates use only some selected recordings that belong to the neighbor set $\mathcal{N}_S$ that are nearest to the common source. The series of M-steps are as follows:

For $l \in \mathcal{N}_S$ and $z \in z^l_I$,

$$P^l_I(f|z) \leftarrow \frac{P^l_I(f|z)P^l(z)\sum_t V^l_{f,t}P^l_I(t|z)}{\sum_{z'\in z_C\cup z^l_I} P^l(f|z')P^l(t|z')P^l(z')},$$

$$P^l_I(f|z) \leftarrow P^l_I(f|z)/\sum_f P^l_I(f|z), \quad (5)$$

$$P^l_I(t|z) \leftarrow \frac{P^l_I(t|z)P^l(z)\sum_f V^l_{f,t}P^l_I(f|z)}{\sum_{z'\in z_C\cup z^l_I} P^l(f|z')P^l(t|z')P^l(z')},$$

$$P_I^l(t|z) \leftarrow P_I^l(t|z) / \sum_t P_I^l(t|z), \tag{6}$$

For $l \in \mathcal{N}_S$ and $z \in z_C$,

$$P_C(f|z) \leftarrow \sum_{l \in \mathcal{N}_S} \frac{P_C(f|z) P^l(z) \sum_t V_{f,t}^l P_C(t|z)}{\sum_{z' \in z_C \cup z_I^l} P^l(f|z') P^l(t|z') P^l(z')} + \beta \alpha_{f,z},$$

$$P_C(f|z) \leftarrow P_C(f|z) / \sum_f P_C(f|z), \tag{7}$$

$$P_C(t|z) \leftarrow \sum_{l \in \mathcal{N}_S} \frac{P_C(t|z) P^l(z) \sum_f V_{f,t}^l P_C(f|z)}{\sum_{z' \in z_C \cup z_I^l} P^l(f|z') P^l(t|z') P^l(z')},$$

$$P_C(t|z) \leftarrow P_C(t|z) / \sum_t P_C(t|z), \tag{8}$$

For $l \in \mathcal{N}_S$ and $z \in z_C \cup z_I^l$

$$P^l(z) \leftarrow \frac{P^l(z) \sum_{f,t} V_{f,t}^l P^l(f|z) P^l(t|z)}{\sum_{z' \in z_C \cup z_I^l} P^l(f|z') P^l(t|z') P^l(z')}. \tag{9}$$

Note that the E-step is absorbed in the listed M-steps as well. For the parameters notated without subscripts C or I their associations should be obvious from the context, or it does not matter whether they are shared or not. For the shared basis vectors, which make up the dominant source, we can also use prior knowledge if the nature of the target source is known, e.g. clean bases from female speech, from a studio recording of the same song played in the scene, etc. We use a conjugate prior $\alpha_{f,z}$ for this, whose contribution is controlled by β as in [1].

Sharing is achieved by doing another average over l for those common parameters in (7) and (8). Therefore, when it comes to hundreds of recordings, this summation would be computationally burdensome had it not been for using the neighboring subset $\mathcal{N}_S$. Focusing only on $\mathcal{N}_S$ also helps speed up learning the individual parameters in (5) and (6), because we can largely skip all the recording-specific modeling when l does not belong to $\mathcal{N}_S$. Once again, $\mathcal{N}_S$ is a set based on cross entropy relationships, but between the observed noisy spectrogram and the estimated source spectrogram as follows:

$$-\sum_{f,t} \hat{V}_{f,t}^{l \in \mathcal{N}_S} \log \hat{S}_{f,t} < -\sum_{f,t} \hat{V}_{f,t}^{l' \notin \mathcal{N}_S} \log \hat{S}_{f,t}, \tag{10}$$

where the spectrograms $\hat{S}$ and $\hat{V}^l$ are normalized along both axes, e.g. $\sum_{f,t} \hat{V}_{f,t}^l = 1$. The source is estimated from the average of the common components over the participating recordings:

$$S_{f,t} \leftarrow \frac{1}{|\mathcal{N}_S|} \sum_{l \in \mathcal{N}_S} V_{f,t}^{(l)} \frac{\sum_{z \in z_C} P_C(f|z) P_C(t|z) P^{(l)}(z)}{\sum_{z \in z_C \cup z_I^{(l)}} P_C(f|z) P_C(t|z) P^{(l)}(z)}. \tag{11}$$

We refresh $\mathcal{N}_S$ at every iteration, and therefore, S should be updated before that, too. Note that (11) is suboptimal, because the amount of the contribution from each recording to the final result is not known.

3.2. The Proposed Two-Stage Method

We define F, T, Z, and L to denote the number of rows, columns, components, and recordings, respectively. Then, the computational complexity of an EM update in (5)-(8) is in the order of $\mathcal{O}\big(FTZ|\mathcal{N}_S|\big)$. In original PLCS $\mathcal{N}_S = L$, but in the proposed systems we set $|\mathcal{N}_S| < L$, so that the use of $\mathcal{N}_S$ is beneficial. The complexity for the construction of the neighbor set in (10) and the source in (11) are $\mathcal{O}(FTL)$ and $\mathcal{O}\big(FT|\mathcal{N}_S|\big)$, respectively.

In this section we propose a two-stage neighborhood search method that further decreases the complexity of (10) from $\mathcal{O}(FTL)$ down to $\mathcal{O}\big(FT|\mathcal{N}_H|\big)$, by introducing a set of candidate neighbors $\mathcal{N}_H$, where $|\mathcal{N}_S| < |\mathcal{N}_H| < L$ and $\mathcal{N}_S \subset \mathcal{N}_H$. Construction of $\mathcal{N}_H$ is cheaper, because we use binary operations.

To this end, we propose to hash all the recordings into a binary representation that can be more efficient in some arithmetic operations. Any hash function can be used if it meets the conditions for the family of locality sensitive hashing: (a) originally closer data points are more probable to collide into the same hash code (b) Hamming distance between the codes approximates the original distance metric [17]. We are particularly interested in Winner-Take-All (WTA) hashing, which also holds those properties [18, 19].

In the recent application of WTA hashing for speeding up the sparse encoding process of dictionary-based source separation, a few candidates of the overcomplete dictionary items are selected based on the WTA Hamming distance in the first place, and then the selection is refined using cross entropy [20]. There are pros and cons of the two searches. Cross entropy is more accurate, yet costly due to the floating-point operations and logarithm. On the other hand, Hamming distance is cheap to calculate thanks to the binary representation of the hash codes, while inaccurate. Therefore, the two-stage search on the dictionary consists of (a) the first round that constructs a bigger candidate set, e.g. $3K$ items, using Hamming distance (b) the second round only on the $3K$ candidates rather than the whole dictionary based on cross entropy, where $K = |\mathcal{N}_S|$.

A WTA hash code is generated by randomly choosing M elements out of the input vector, and then writing down the index of the maximum among them by flipping the corresponding element of an all-zero binary vector of length M as the indicator. By repeating this experiment Q times, the total bits used for an input vector are QM bits[2]. Hamming distance among the two binary hash codes $x, y \in \mathbb{B}^{QM \times 1}$ can be calculated by a bitwise AND operation followed by bit counting, i.e. $\sum_i x_i \wedge y_i$, and then inverting the result. Since this procedure approximates a rank order metric, a relative ordering of the input elements, it can still be a discriminative feature even with its binary representations. On the other hand, the mismatch between Hamming distance and cross entropy hinders WTA hash codes from replacing the original distance measure.

We start from hashing all the recordings in advance by using the WTA hash function ϕ: $\tilde{v}^l \leftarrow \phi(\text{vec}(V^l))$, where vec() vectorizes a matrix. Now after every source update (11), we update the source' hash code as well: $\tilde{s} \leftarrow \phi(\text{vec}(S))$. In the first-round search, using this new hash code along with the already prepared ones for the recordings, we calculate another candidate set $\mathcal{N}_H$, which meets an inequality as follows:

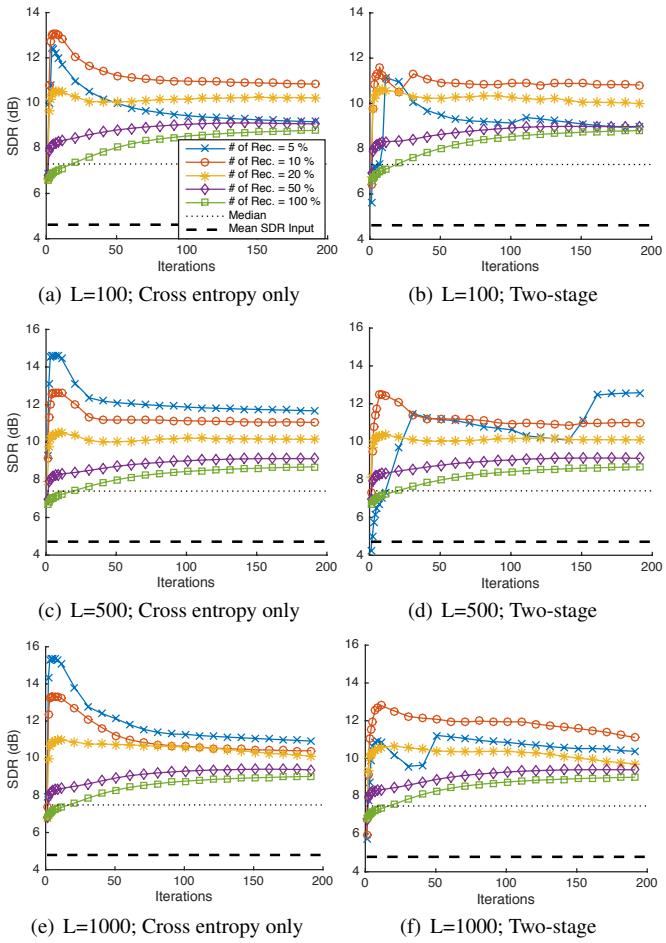
$$\sum_n^{QM} \big(\tilde{v}_n^{l \in \mathcal{N}_H} \wedge \tilde{s}_n\big) > \sum_n^{QM} \big(\tilde{v}_n^{l' \notin \mathcal{N}_H} \wedge \tilde{s}_n\big), \tag{12}$$

Now that we have a candidate set, the second search is limited only on $\mathcal{N}_H$ rather than all the L recordings. In other words, we still do the search based on (10), but now the inequality is guaranteed only in the candidate set $l, l' \in \mathcal{N}_H$, assuming $\mathcal{N}_S \subset \mathcal{N}_H$.

4. EXPERIMENTS

For experimental validation, we simulate an audio scene where 30 sources are randomly scattered in a square space of 50 meters by 50

[2]For example, $3 = 0100$ when $M = 4$, while in the usual binary representation with $\lceil \log_2(M) \rceil$ bits per integer, $3 = 11$. This consumes more bits, but the Hamming distance calculation is more convenient.



(a) L=100; Cross entropy only (b) L=100; Two-stage

(c) L=500; Cross entropy only (d) L=500; Two-stage

(e) L=1000; Cross entropy only (f) L=1000; Two-stage

Fig. 2: Average SDR performances of the systems with different numbers of input signals and nearest neighboring measures.

meters. Among the sources, one female speech is randomly chosen and placed at the center with 20dB louder volume than the others. 30 sources consist of 10 male and 10 female TIMIT speech signals, and 10 non-stationary noise signals used in [21]. All sources are either cut to 10 seconds long if longer than that or repeated otherwise, and sampled at 16KHz. $L = \{100, 500, 1000\}$ are the number of sensors that are randomly placed. Therefore, a recording is a mixture of all the sources with different sound pressure levels depending on the distance between the source and the sensor. This is a challenging setup that models real-world situations because the sensors far from the dominant source can exhibit significantly louder interference. For now we assume that the signals are already aligned, mixing is instantaneous, and no additional artifacts are present.

A short-time Fourier transform is done with 1024 pt Hann windowing with 75% overlap. We set up 50 components for sharing while 10 others are assigned per recording to capture interferences. The a prior bases α for the common components were trained from 20 different female speakers using an ordinary PLSI. We choose a big number, e.g. 5000, to initialize β, and decay it exponentially during the updates. We change the number of neighbors to be one of $|\mathcal{N}_S| = L \times \{0.05, 0.1, 0.2, 0.5, 1\}$. The number of WTA candidates are set to be $|\mathcal{N}_H| = \min(3|\mathcal{N}_S|, L)$. WTA parameters Q and M are set to be $\lceil \log_2 FT \rceil = 2^{19}$ and 2, respectively.

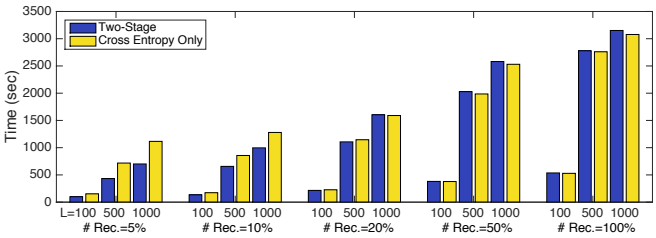


Fig. 3: Run-time analysis of the PLCS system and the proposed neighborhood-based methods.

Each sub-figure in Fig. 2 is an average of five repeated experiments using five different choices of the dominant female source, and randomized geometric configurations of the other sources and sensors accordingly. Signal-to-Distortion Ratio (SDR) was measured up to 200 EM iterations as an overall separation quality score. First thing we notice in (a), (c), and (e) is that the neighbor set in terms of cross entropy successfully enhances the performance using only a small set of nearest neighbors: all the choices using this neighborhood concept converge to a better solution earlier than the full PLCS that uses 100% of the recordings (green squares). When there are enough number of recordings (500 or 1000), only 5% of them are needed to obtain the best results (blue crosses). All the systems including the full PLCS perform better than the average of the input SDRs, or the SDR of a the median spectrogram of all recordings, i.e. $V_{f,t}^{\text{median}} = \text{median}([V_{f,t}^1, V_{f,t}^2, \cdots V_{f,t}^L])$.

Two-stage methods in (b), (d), and (f) also show on average better performance than the full PLCS model or the median spectrogram of the recordings. However, it is noticeable that its convergence is not stable when only 5% are used. We believe that this fluctuation can be mitigated when we increase the size of $\mathcal{N}_H$, but it comes with the cost of increased run-times. 10% and 20% cases are more stable than the 5% case and comparable to their cross entropy counterparts.

Fig. 3 compares the average run-times of all the cases. If the neighbor sets are reasonably small (5% or 10% in the first two bar groups), we see a speed-up by using the two-stage method. Furthermore, the gap becomes larger as L increases. However, if $|\mathcal{N}_S|$ is too big ($\geq 20\%$) the two-stage method starts to add more overhead than the desired speed-up. Overall, both neighborhood methods with a reasonably smaller neighborhood set, $\mathcal{N}_S \leq 0.2L$, reduce the run-times down to from around 16% to 50% of the the full PLCS model. It is a significant saving given the fact that our MATLAB-based implementation is not well-suited for a fair comparison in this case, penalizing the efficiency of bitwise operations.

5. CONCLUSION AND FUTURE WORK

We proposed a neighborhood-based extension for the PLCS model to handle a large number of recordings that can be found in abundance through social data. We proposed two different neighbor search schemes, one using the comprehensive cross entropy between a tentative source spectrogram and the noisy recordings, and the other first reduces the set down to some candidates based on WTA hash codes followed by the comprehensive search only on those candidates. Experimental results clearly supported the merit of the proposed methods in terms of separation performances, convergence behaviors, and run-times. Although separation was successful, the robustness of the neighborhood-based extension to the other types of recording artifacts, such as band-pass filtering, clipping, reverberation, etc, is not shown here, and we leave it for future work.

6. REFERENCES

[1] M. Kim and P. Smaragdis, "Collaborative audio enhancement using probabilistic latent component sharing," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013.

[2] V. Raykar, I. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 70–83, Jan 2005.

[3] I. Himawan, I. McCowan, and S. Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays. audio," *ieeeaslp*, vol. 19, no. 4, pp. 661–676, 2011.

[4] C. Cotton and D. P. Ellis, "Audio fingerprinting to identify multiple videos of an event," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2010.

[5] N. J. Bryan, P. Smaragdis, and G. J. Mysore, "Clustering and synchronizing multicamera video via landmark cross-correlation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012.

[6] A. Asaei, N. Mohammadiha, M. J. Taghizadeh, S. Doclo, and H. Bourlard, "On application of non-negative matrix factorization for ad hoc microphone array calibration from incomplete noisy distances," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

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

[8] ——, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 13. MIT Press, 2001.

[9] M. Kim, J. Yoo, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for spectral and temporal drum source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1192–1204, 2011.

[10] T. Hofmann, "Probablistic latent semantic analysis," in *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*, 1999.

[11] ——, "Probablistic latent semantic indexing," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1999.

[12] M. Shashanka, "Latent variable framework for modeling and separating single channel acoustic sources," Ph.D. dissertation, Boston University, Aug. 2007.

[13] C. Ding, T. Li, and W. Peng, "On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing," *Computational Statistics and Data Analysis*, vol. 52, pp. 3913–3927, 2008.

[14] S. T. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[15] P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, BC, Canada, 2009.

[16] M. Kim and P. Smaragdis, "Manifold preserving hierarchical topic models for quantization and approximation," in *Proceedings of the International Conference on Machine Learning (ICML)*, Atlanta, Georgia, 2013.

[17] P. Indyk and R. Motwani, "Approximate nearest neighbor – towards removing the curse of dimensionality," in *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, 1998, pp. 604–613.

[18] J. Yagnik, D. Strelow, D. A. Ross, and R. Lin, "The power of comparative reasoning," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011, pp. 2431–2438.

[19] T. Dean, M. A. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik, "Fast, accurate detection of 100,000 object classes on a single machine," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[20] M. Kim, P. Smaragdis, and G. J. Mysore, "Efficient manifold preserving audio source separation using locality sensitive hashing," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2015.

[21] Z. Duan, G. J. Mysore, and P. Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in non-stationary noise environments," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Portland, OR, 2012.