BLIND SPEECH SEPARATION BASED ON COMPLEX SPHERICAL K-MODE CLUSTERING

Lukas Drude, Christoph Boeddeker, Reinhold Haeb-Umbach

University of Paderborn, Department of Communications Engineering, Paderborn, Germany

ABSTRACT

We present an algorithm for clustering complex-valued unit length vectors on the unit hypersphere, which we call complex spherical k-mode clustering, as it can be viewed as a generalization of the spherical k-means algorithm to normalized complex-valued vectors. We show how the proposed algorithm can be derived from the Expectation Maximization algorithm for complex Watson mixture models and prove its applicability in a blind speech separation (BSS) task with real-world room impulse response measurements. It turns out that the proposed spherical k-mode algorithm is on par with other state-of-the-art BSS algorithms in terms of signal-to-inference ratio gains although being far easier to implement and using fewer calculations.

Index Terms— Directional statistics, speech separation, clustering, complex hypersphere, sparseness

1. INTRODUCTION

We consider the recovery of individual speech signals from reverberant mixtures corrupted by noise. Approaches to blind speech separation (BSS) include Independent Component Analysis (ICA) based [1] and sparseness based methods, where an early approach is time frequency masking [2]. While the first is mostly restricted to scenarios where the number of sensors is at least as large as the number of sources, the second can also be employed in the underdetermined case with fewer sensors than sources. Other approaches rely on dictionary learning, such as nonnegative matrix factorization [3, 4].

Due to the sparseness of speech in the short time Fourier transform (STFT) domain [2] one can assume under fairly general conditions that each time frequency (tf) slot is occupied by a single speech source only, or by noise. BSS then amounts to determining which source is dominant in each tf slot and estimating the source parameters from those tf slots assigned to this source. Since the former depends on the latter and vice versa, iterative algorithms are employed. They can be derived from the well-known Expectation Maximization (EM) algorithm, where the resulting algorithm depends on which features are computed from the microphone signals and by which statistical model they are described [5, 6].

The most prominent models used in this context are Gaussian mixture models (GMMs) where a wide range of features have been compared in the context of k-means clustering [7].

A fairly recent model is a complex Watson mixture model (cWMM). It is a statistical model for the vector of microphone signals in the STFT domain, after normalization to unit length. The key motivation for this normalization is that it decouples the transmission path related features, namely phase and level differences between the microphone signals, from the source related features, namely absolute phase and signal energy [5, 8, 9]. Due to the fact that the complex Watson distribution [10] is sensitive to level and phase differences but invariant with respect to the absolute phase, it ideally accounts for the transmission path related features only.

In this contribution we will derive a clustering algorithm, which is an approximation to the EM algorithm applied to observations drawn from the cWMM. We call this algorithm complex spherical k-mode clustering and show that it relates to the EM for cWMMs in much the same manner as the famous k-means algorithm [11] to a GMM and the spherical k-means to a mixture of von Mises-Fisher distributions [12].

In the following, we will first describe the signal model and review the EM algorithm for cWMMs. From this we will then derive a clustering algorithm and relate it to Lloyd's k-means algorithm [11]. Finally, we will analyze the proposed algorithm in terms of signal to interference ratio (SIR) gains and show, that it is on par with alternative algorithms, while requiring significantly less computational effort.

2. SIGNAL AND STATISTICAL MODEL

Let us assume a convolutive mixture of K independent source signals $S_{ft1}, \dots S_{ftK}$ captured by D sensors in the STFT domain:

$$\mathbf{Y}_{ft} = \sum_{k=1}^{K} \mathbf{H}_{fk} S_{ftk} + \mathbf{N}_{ft}, \tag{1}$$

where $\mathbf{Y}_{ft}$, $\mathbf{H}_{fk}$ and $\mathbf{N}_{ft}$ are the D-dimensional vector of the microphone signals, the acoustic transfer function vector from the k-th source to the D microphones, and the vector of noise signals, respectively. Here, t and f denote the time frame and frequency bin, respectively.

The unit-length normalized observations are then obtained as follows:

$$\widetilde{\mathbf{Y}}_{ft} = \frac{\mathbf{Y}_{ft}}{A_{ft}}, \qquad \text{where} \qquad A_{ft} = \sqrt{\mathbf{Y}_{ft}^{\mathrm{H}} \mathbf{Y}_{ft}}. \tag{2}$$

Since speech signals are sparse in the STFT domain, we may assume that a time frequency slot is occupied either by a single source and noise or by noise only.

Since all frequencies are treated equally, we will drop the frequency index in the following.

The normalized observation vectors $\mathcal{Y} = \{\widetilde{\mathbf{Y}}_t | \forall t\}$ form clusters on the D-dimensional complex unit hypersphere for each frequency bin f independently. Their distribution is modeled by a cWMM, where $\boldsymbol{\theta} = \{\pi_k, \kappa_k, \mathbf{W}_k | \forall k\}$ is the parameter set, comprising the mixture weight π_k, concentration κ_k, and mode direction $\mathbf{W}_k$ for each class k. The class labels $\mathcal{C} = \{c_t | \forall t\}$ are assumed to be categorically distributed with probabilities $\pi_k, k = 1, \ldots, K$.

The hierarchical generative model consists of first sampling an indicator variable from the categorical distribution followed by sampling from the corresponding complex Watson distribution:

$$p(\widetilde{\mathbf{Y}}_t) = \sum_{k=1}^{K} \pi_k p(\widetilde{\mathbf{Y}}_t | c_t = k), \tag{3}$$

$$p(\widetilde{\mathbf{Y}}_t | c_t = k) = \frac{1}{c_{\mathrm{W}}(\kappa_k)} \mathrm{e}^{\kappa_k |\widetilde{\mathbf{Y}}_t^{\mathrm{H}} \mathbf{W}_k|^2}. \tag{4}$$

Here, c_{W} is a normalization constant [10].

The parameter set $\boldsymbol{\theta}$ can be estimated via the EM algorithm [5], which is performed for each frequency bin separately. During the E-step, the source posteriors are computed

$$\gamma_{tk} := P(c_t = k | \widetilde{\mathbf{Y}}_t) = \frac{P(c_t = k) p(\widetilde{\mathbf{Y}}_t | c_t = k)}{\sum_{k=1}^{K} P(c_t = k) p(\widetilde{\mathbf{Y}}_t | c_t = k)} \tag{5}$$

while the M-step delivers the parameter updates. For the mixture weights we obtain

$$\pi_k = N_k / T, \qquad \text{where} \qquad N_k = \sum_{t=1}^{T} \gamma_{tk}, \tag{6}$$

where T is the total number of frames.

The estimate for the mode direction $\mathbf{W}_k$ is the principal component, i.e., the eigenvector with the largest eigenvalue, of the weighted correlation matrix $\boldsymbol{\Phi}_k$:

$$\mathbf{W}_k = \mathcal{P}\left\{\boldsymbol{\Phi}_k\right\}, \quad \text{where} \quad \boldsymbol{\Phi}_k = \frac{1}{N_k} \sum_{t=1}^{T} \gamma_{tk} \widetilde{\mathbf{Y}}_t \widetilde{\mathbf{Y}}_t^{\mathrm{H}}. \tag{7}$$

The concentration parameter κ_k is obtained by solving the implicit equation:

$$\frac{{}_1F_1(2, D+1, \kappa_k)}{D \; {}_1F_1(1, D, \kappa_k)} = \mathbf{W}_k^{\mathrm{H}} \boldsymbol{\Phi}_k \mathbf{W}_k, \tag{8}$$

where the right hand side is simply the eigenvalue belonging to $\mathbf{W}_k$ and ${}_1F_1(\cdot, \cdot, \cdot)$ is the confluent hypergeometric function.

3. SPHERICAL K-MODE ALGORITHM

Now we introduce similar approximations as in the derivation of the k-means algorithm from the EM for GMMs.

The main simplification is a quantization of the posteriors γ_{tk}. We introduce the hard class assignments

$$\hat{c}_{tk} = \begin{cases} 1, & k = \operatorname*{argmax}_{\tilde{k}} \gamma_{t\tilde{k}}, \\ 0, & \text{else.} \end{cases} \tag{9}$$

This is motivated by the fact, that the assumption in the first place was a sparse signal model, where, despite the mixture at the input, each tf slot is dominated by a single source.

We further assume equal mixture weights, $\pi_k = \pi$ for all k, and shared concentration parameters $\kappa_k = \kappa$ for all k. Using this in Eq. (5) with the definition of the complex Watson distribution in Eq. (4) we arrive at:

$$\begin{aligned} k &= \operatorname*{argmax}_{\tilde{k}} \pi_{\tilde{k}} \frac{1}{c_{\mathrm{W}}(\kappa_{\tilde{k}})} \mathrm{e}^{\kappa_{\tilde{k}} |\widetilde{\mathbf{Y}}_t^{\mathrm{H}} \mathbf{W}_{\tilde{k}}|^2} \\ &= \operatorname*{argmax}_{\tilde{k}} \mathrm{e}^{|\widetilde{\mathbf{Y}}_t^{\mathrm{H}} \mathbf{W}_{\tilde{k}}|^2} \\ &= \operatorname*{argmax}_{\tilde{k}} |\widetilde{\mathbf{Y}}_t^{\mathrm{H}} \mathbf{W}_{\tilde{k}}|^2. \end{aligned} \tag{10}$$

It finally turns out, that the class affiliations can be determined by simply maximizing the squared cosine distance between the normalized observations $\widetilde{\mathbf{Y}}_t$ and the class dependent mode directions $\mathbf{W}_k$.

Thus, the complexity of the E-step is greatly reduced, because the evaluation of the complete probability density function is replaced by the computation of the square of an inner product of two vectors. The complexity of the M-step is also reduced, since it is no longer necessary to estimate the concentration parameters and mixture weights.

The mode direction still remains to be the principal component of the covariance matrix of normalized observations using the newly derived class affiliations:

$$\mathbf{W}_k = \mathcal{P}\left\{ \left(\sum_{t=1}^{T} \hat{c}_{tk} \right)^{-1} \sum_{t=1}^{T} \hat{c}_{tk} \widetilde{\mathbf{Y}}_t \widetilde{\mathbf{Y}}_t^{\mathrm{H}} \right\}. \tag{11}$$

The relation between the proposed clustering algorithm and the EM for the cWMM is the same as the relation between Lloyd's k-means algorithm [11] and the EM for the GMM, or between the spherical k-means algorithm and the EM for a mixture of von Mises-Fisher distributions [12]. In all these cases, soft assignments are replaced by hard class affiliations, and both class priors and class conditional variances/precisions are neglected.

The proposed clustering algorithm differs from k-means in the distance function used (squared cosine distance vs. Euclidian distance) and in the way the prototypes are computed (principal component vs. sample mean). Similarly, it differs

from the spherical k-means [13] in the distance function used (squared cosine distance vs. cosine distance) and in the way the prototypes are computed (principal component vs. normalized sample mean).

Especially the fact that the phase offset of each vector is caused by the signal source and is unrelated to the transmission path shows, that a simple Euclidean distance will not yield meaningful clusters in the considered BSS problem, neither does a cosine distance without calculating its squared absolute value.

4. EXTENSION TO NOISE-ONLY OBSERVATIONS

Up to now, the clustering algorithm has been derived for concurrent speakers only. To extend the algorithm to account for observations which contain only noise and no speaker signal, we can simply add an additional class for noise and cluster the observations into $K' = K + 1$ classes.

This is particulary accurate for directed noise sources but is sufficiently accurate for spherically isotropic noise models as well, as will be seen by the experimental results. For a spherically isotropic noise field one assumes uniformly distributed noise sources in the 3D space. Following [14] this does not lead to equal eigenvalues of the noise covariance matrix and, therefore, does not lead to uniformly distributed noise on the complex unit hypersphere.

5. PERMUTATION ALIGNMENT

Since the described algorithms operate independently on each frequency, a well-known permutation problem arises: Even if the source separation were perfect for each frequency bin, it is likely, that component one of one frequency bin does not correspond to the same speaker as component one of another frequency bin. In order to reconstruct the individual sources, the labels for each frequency have to be reordered to match each other along all frequencies. For the experimental results reported here we employed a clustering algorithm similar to the EM algorithm reported in [15], which is based on maximizing the intra-class similarities from vectors $\mathbf{a}_{fk} = (\gamma_{f1k}, \ldots, \gamma_{ftk}, \ldots, \gamma_{fTk})$ containing the a posteriori probabilities γ_{ftk} in each frequency bin.

6. PERFORMANCE EVALUATION

We employed room impulse response measurements from the MIRD database [16], which were resampled to $16\,\mathrm{kHz}$. They were convolved with speech segments of $5\,\mathrm{s}$ length obtained from the TIMIT database [17].

These acoustic room impulse responses correspond to a linear array consisting of $D = 6$ sensors of type AKG CK32 with sensor distances $3\,\mathrm{cm}$, $3\,\mathrm{cm}$, $8\,\mathrm{cm}$, $3\,\mathrm{cm}$, and $3\,\mathrm{cm}$. The $K = 2$ speech sources are placed on a half circle in front of the linear array with an angular difference of 60° and a distance of $1\,\mathrm{m}$ from the array center.

An artificially generated spherically isotropic noise field was generated by the algorithm presented in [14] and added to the speech images at the sensors. The STFT was then applied to the aforementioned signals with an FFT size of 1024 samples, a shift of 256 samples and using a Blackman window.

We have chosen $K' = K + 1$ normalized observations at random and used them as initial values for the mode directions. For the EM-algorithm, the initial weights and concentrations have been set to $1/K'$ and 20, respectively.

After the EM or the clustering algorithm had estimated the mode vectors, the signals are then reconstructed using a linearly constrained minimum variance (LCMV) beamformer:

$$\mathbf{H}_{\mathrm{LCMV}} = \mathbf{\Phi}_{NN}^{-1}\mathbf{H}_{\mathrm{all}}(\mathbf{H}_{\mathrm{all}}^{\mathrm{H}}\mathbf{\Phi}_{NN}^{-1}\mathbf{H}_{\mathrm{all}})^{-1}\mathbf{g}, \qquad (12)$$

where $\mathbf{\Phi}_{NN}$ is the covariance matrix of the observations $\mathbf{Y}$ masked with the noise mask, $\mathbf{H}_{\mathrm{all}}$ is the $(D \times K')$-matrix of relative acoustic transfer function estimates $\mathbf{W}_k$ obtained from either Eq. (7) or (11). Further, $\mathbf{g}$ is the response vector which is set to one for the desired source and is set to zero for all sources to be suppressed.

Fig. 1 shows the performance in terms of the SIR gain as defined in [18] for three different reverberation times T_{60} of $0.16\,\mathrm{s}$, $0.36\,\mathrm{s}$ and $0.61\,\mathrm{s}$, respectively. The following algorithms were compared:

- **k-means** with k-means++ initialization [19], Euclidean distances and prototypes calculated by a mean operation. It operates on $\widetilde{\mathbf{Y}}_t = \widetilde{\mathbf{Y}}_t \exp(-\mathrm{j} \arg Y_{1t})$.
- **k-mode** is the proposed spherical k-mode clustering algorithm, see Sec. 3.
- **EM** algorithm for cWMM as in Sec. 2.
- **IT13** is our implementation of the permutation-free BSS algorithm of [9]. It also employs a cWMM.
- **IT13+PA**: Although the algorithm of [9] does not require a permutation alignment, we found that it can still be improved by an additional permutation alignment.

We calculate the SIR gain intrusively by estimating the beamforming vectors on the observed signal and applying the beamforming vector to each signal image separately. We opted to not use BSSEval as in [18] since due to the simulation setup we had the source images available and did not need to rely on an additional estimation.

The results in Fig. 1 and Fig. 2 summarize 100 simulations for each setting and each algorithm. They are presented as box plots, where the box starts at the $25\,\%$ quantile q_1 and ends with the $75\,\%$ quantile q_3 while the line inside the box indicates the median. The small whiskers mark the values $q_3 + 1.5(q_3 - q_1)$ and $q_1 - 1.5(q_3 - q_1)$ and the outliers are presented as plus signs.

At first, it can be observed that our implementation of IT13 did not perform well on the given scenario. It achieved

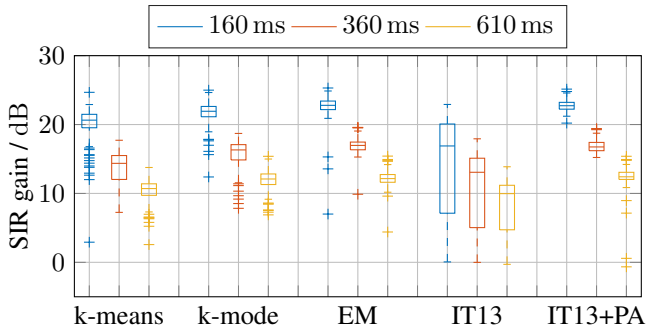

Fig. 1: SIR gain for different algorithms and different reverberation times (T_{60}) and spherically isotropic noise with an SNR of 15 dB.

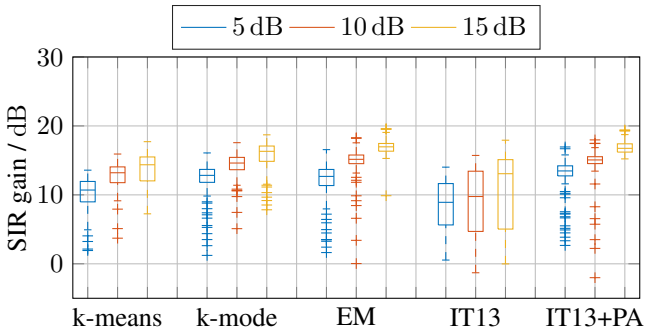

Fig. 2: SIR gain for different algorithms and fixed reverberation time of $T_{60} = 360\,\text{ms}$ and spherically isotropic noise with varying SNR.

the lowest SIR gain of the tested algorithm and the scatter of the results was large. For example, the algorithm achieved a median SIR gain of 13.1 dB in scenarios with a reverberation time of 360 ms and a noise level of 15 dB. We therefore decided to combine it with an additional permutation alignment step to provide a solid baseline for the other compared algorithms. This additional step increased its performance by 3.6 dB to a solid SIR gain of 16.7 dB.

IT13 with additional permutation alignment, the EM and the k-mode algorithm achieved comparable SIR gains, as can be seen from Fig. 1. For example, in scenarios with a reverberation time of 360 ms and a noise level of 15 dB the proposed spherical k-mode achieved a median SIR gain of 16.3 dB whereas the EM algorithm achieved 17.0 dB. The proposed algorithm therefore shows a 0.7 dB gap to the best performing algorithm.

On the contrary, the k-means with k-means++ initialization and with additionally phase normalized features achieved a median SIR gain of 14.4 dB, which is clearly inferior.

Fig. 2 now shows the same algorithms in different noise conditions. It can be observed that the relative performance of the algorithms remains the same, irrespective of the input SNR: k-mode, EM, and IT13+PA achieve similar SIR gains, while k-means is clearly less effective. Further, not surprisingly, the SIR gain consistently increases with increasing input SNR.

To get a feeling of the computational complexity, Table 1 lists the number of operations for the key algorithmic parts for each of the investigated methods. The greatest speed-up of the k-mode algorithm in comparison to the EM algorithm is achieved by the replacement of the likelihood calculations by squared cosine similarity calculations. This drastically reduces the computation time of the E-step, while the effort in the M-step is dominated by the eigenvalue decomposition and thus remains largely unaffected. The k-means algorithm further drastically reduces the computational effort of the M-step. Note that in the original EM algorithm the E-step and the M-step require approximately the same computation time. It is worth noting, that the k-means uses additional input normalization and simply replacing the principal component analysis in the k-mode algorithm by a mean operation yields much worse results. We decided against reporting actual CPU times, since we implemented the different parts in different languages.

Table 1: Number of selected operations per frequency for each algorithm: calculated likelihoods, eigenvalue decompositions, solved implicit equations, permutations. The number of iterations is I.

	k-means	k-mode	EM	IT13	IT13+PA
#likelihoods	0	0	TKI	TKI	TKI
#eigenv.	0	IK	IK	IK	IK
#implicits	0	0	IK	IK	IK
#permutes	K^2	K^2	K^2	$IK!$	$IK!$+K^2

7. CONCLUSIONS AND RELATION TO PRIOR WORK

We have presented a novel clustering algorithm for observations on a complex unit hypersphere. It has been used in a blind source separation scenario and shows comparable or even superior performance to existing algorithms, while at the same time being considerably less computationally complex.

The work presented here can be viewed as an extension of prior work in two respects. First, it is a simplification of our early proposed EM algorithm for BSS [5] having a much simpler E-step, while achieving similar source separation performance. Second, it is shown that it relates to the EM algorithm for cWMM in the same way as the spherical k-means algorithm to the EM for a mixture of von Mises-Fisher distributions [13, 12]. We therefore believe that the proposed spherical k-mode algorithm can find applications beyond speech source separation in fields, where complex-valued directional data are to be modeled and analyzed, such as in statistical shape analysis.

8. REFERENCES

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, Wiley and Sons, 2001.

[2] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

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

[4] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Audio, Speech, Language Process.*, vol. 21, no. 5, pp. 971–982, 2013.

[5] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 241–244.

[6] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 21, no. 9, pp. 1913–1928, Sept 2013.

[7] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[8] I. Jafari, R. Togneri, and S. Nordholm, "On the use of the Watson mixture model for clustering-based underdetermined blind source separation," in *Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014, pp. 988–992.

[9] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 3238–3242.

[10] K. V. Mardia and I. L. Dryden, "The complex Watson distribution and shape analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 4, pp. 913–926, 1999.

[11] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[12] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," *Journal of Machine Learning Research*, , no. 6, pp. 1345–1382, 2005.

[13] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning*, , no. 1, pp. 143–175, 2001.

[14] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.

[15] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Audio, Speech, Language Process.*, vol. 19, no. 3, pp. 516–527, 2011.

[16] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 313–317.

[17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, D. S. Pallett, and N. L. Dahlgren, *The DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CDROM*, U.S. Department of Commerce, 1993.

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

[19] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.